DURATION AND INTONATION KNOWLEDGE FOR TEXT-TO-SPEECH CONVERSION SYSTEM FOR TELUGU AND HINDI

A THESIS

submitted by

K. KIRAN KUMAR

for the award of the degree

of

MASTER OF SCIENCE

(by Research)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, MADRAS.

MAY 2002

To My

Amma, Nānna

Annayya and Bābāi

THESIS CERTIFICATE

This is to certify that the thesis entitled **Duration and Intonation Knowledge for Text-To-Speech Conversion System for Telugu and Hindi** submitted by **K. Kiran Kumar** to the Indian Institute of Technology, Madras for the award of the degree of Master of Science (by Research) is a bonafide record of research work carried out by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Chennai-36 — Prof. B. Yegnanarayana

Date: — Dept. of Computer Science and Engg.

ACKNOWLEDGEMENTS

I am extremely thankful to my guides Prof. B. Yegnanarayana and Dr. C. Chandra Sekhar, and to all the members of Speech and Vision Lab for making my stay in the lab memorable. I have been very fortunate to have *Prof.* as my guide. I have been benefited greatly from his advice and support throughout my stay in the lab. He has been a great source of inspiration and has provided the right balance of suggestions, criticism and freedom. This thesis would not have been possible without his generous efforts. I would also like to thank him for having maintained excellent facilities and favorable working conditions in the lab.

I also would like to thank Dr. C. Chandra Sekhar for his advice and guidance during the initial phases of my work. I am very unfortunate to miss him in the later part of my work. Special thanks to Prof. S. Raman, Head of CSE Dept., for providing excellent infrastructure in the department.

I thank all my lab members Anil, Anvita, Deveraj, Gaurav [Gabbar], Gupta, Jinu, Kiran [P], KSR [Srinivasa Rao], Meera, Nayeem, Prasad [Adavi], Prasanna, Ram, Rambo [Siva], Ramesh, Sharat, Spal, SVG [Surya], Venky [Monky], Vinod and Vishwam. Special thanks to KSR for his efforts in patiently correcting my thesis in sections, chapters and drafts and supplying me with right suggestions and comments. Thanks Prasannaji for your help during the preparation of the paper. I express my sincere thanks to SVG, my hostel mate and lab mate, for supporting me personally and technically. I can not forget those wonderful *night outs* spent with Venky, Vinod, Anvita and Gabbar while writing my thesis, and the taste of bread amlets @ early mornings.

I am fortunate to have been part of IIT for nearly three years. I thank all my hostel mates and friends for making my stay at IIT wonderful. I can not forget the [B]treats at Besant Nagar beach in the evenings along with Saradhi, Reddy, Sisodia, Kamatam, Das, Chandu, Ranjith and saifulla, and the B'day bumps. I am very fortunate to have such a great bunch of friends. I also would like to mention the late night discussions with Vijay @ 367, Krishna.

I am a part of this institute writing this thesis only because of the love, encouragement and cooperation given by my parents, brother and babai.

– *Kiran Kumar*

ABSTRACT

Keywords: *speech synthesis; parameter concatenation model; suprasegmental features; parser; positional and contextual factors; duration and intonation rules.*

The task of a Text-To-Speech (TTS) conversion system is to produce intelligible and natural sounding speech output from an unrestricted text input. Intelligibility refers to the segmental quality of the speech signal. Naturalness of the synthetic speech refers to how close the speech is to human speech in terms of suprasegmental or prosodic features. The suprasegmental features duration, intonation and stress are referred to as prosodic features. Segmental quality depends on the synthesis technique and the size of the basic synthesis unit. Waveform concatenation technique with large size of the synthesis units gives high intelligibility. Naturalness of the speech depends on how well the prosodic features are modelled and incorporated into the system.

The two important tasks to improve the intelligibility and naturalness of the synthetic speech are acquisition and modeling of prosodic knowledge, and incorporating the acquired knowledge. Knowledge-based and data-based approaches are the two well known methods used to model the prosodic knowledge in a TTS system. Knowledge-based approaches have superior generalization capabilities, and can handle new contexts in an optimum manner. Hence knowledge-based approach is adapted to model the prosodic knowledge in our TTS system. In our case the basic system uses parameter concatenation for synthesizing speech. The basic units of the system are consonants, vowels and consonant-vowel units. Linear Prediction (LP) synthesis is used for synthesizing consonantal regions, and cascaded formant synthesis is used for vowel regions. To derive the sequence of basic units from an unrestricted text input, the parser is modified to accept the input text coded using ITRANS code for Indian languages.

The duration rules of the basic system are derived by analyzing the durations of the basic units from a limited database. These rules need to be modified to reflect the durational characteristics of continuous speech. New set of rules are obtained by analyzing the durations of the basic units in news reading style of continuous speech for the Indian language, Hindi. A set of new rules for modeling the duration for a TTS conversion system for Telugu is also developed by analyzing the durations of the basic units in reading style of speech. A database consisting of the labelled speech data of news in reading style for two male and one female speaker is used for studying the durational characteristics of the basic units of Telugu. A study of the intonation patterns of the speech in Telugu is carried out in the present work. The derived rules are incorporated by activating the rules depending on the context of the basic unit in the input text.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION TO TEXT-TO-SPEECH (TTS) SYSTEM

Speech is the most natural means of communication. Also, speech mode of communication does not require any specialized skills. Hence, this could also be the most desirable way of communicating with machines. To achieve this goal, machines should be made to speak and recognize human speech. Here comes the importance of speech synthesis. The function of a Text-to-Speech (TTS) conversion system is to produce natural sounding and intelligible speech from an input text. There are two types of TTS conversion systems [1]. They are: (a) Voice Response Systems and (b) Unrestricted TTS systems. Voice response systems are used for limited domain synthesis of speech. In limited domain synthesis, we know *a priori* the probable vocabulary of the output speech. As the vocabulary size is small, normal speech can be recorded and appropriately accessed for generating the desired speech. So the sizes of the synthesis units can be very large, and can vary from words to phrases or even to sentences. These systems are suitable for applications such as speaking toys, warning systems in machines, telephone directory assistance and railway information access systems.

The second type of systems generate speech from an unrestricted text, i.e, the size of the vocabulary is not limited. These systems typically use smaller size units for synthesis. Unrestricted TTS consists of two phases. In the first phase, the input text is processed and then parsed to derive a sequence of basic speech units. The second phase is to synthesize speech from the sequence of the basic units. Each phase has its own influence on the quality of the output speech.

1.2 SYNTHESIS TECHNIQUES

There are different techniques developed for synthesizing speech from a text input. The techniques are usually classified into the following three types:

- Articulatory synthesis, which attempts to model the human speech production system [2].
- Formant synthesis, which models the resonant frequencies or the transfer function of the vocal tract system based on source-filter model [3].
- Concatenative synthesis, which uses varying lengths of prerecorded speech samples directly or the coded version of the speech samples derived from natural speech [4].

1.2.1 Articulatory Synthesis

Articulatory synthesis tries to model the human vocal organs as closer as possible, so it has potent of the most satisfying technique to produce high quality synthetic speech. But it turns out to be the most difficult technique as it is impossible to model and control all the finer movements of all the vocal organs. Also, the computational complexity is considerably higher than the other techniques [4]. Thus it has received less attention than other synthesis techniques.

1.2.2 Formant Synthesis

Probably the most widely used synthesis method during last decades has been formant synthesis, which is based on the source-filter model of speech production mechanism. There are two basic structures of formant synthesizers in general. They are cascade and parallel formant synthesizers. But for better performance of the synthesizer, some combination of these two structures is generally used [3]. The excitation signal used in formant synthesis consists of a periodic impulse train for voiced speech and white noise for unvoiced speech. The formant filters represent the resonances of the vocal tract. This technique is capable of generating a variety of speech sounds, which makes it very flexible.

1.2.3 Concatenative Synthesis

Concatenating prerecorded speech utterances is probably the easiest way to produce intelligible and natural sounding synthetic speech [5]. One of the most important aspects in concatenative synthesis is to select suitable basic sound units. The selection is generally a trade-off between the size of the memory and the quality of speech. With larger size units high naturalness can be achieved due to fewer number of concatenation points. But as the number of such large size units increase, the size of the memory required to store them will also increase. Though the quality of synthetic speech decreases with small size units due to increase in the number of concatenation points, the size of the memory decreases significantly.

Concatenation of words is relatively easy to perform, and also coarticulation effects within the words are captured in the stored units. However, due to differences between the words spoken in isolation and in continuous speech, the synthetic speech will not be natural sounding. The current synthesis systems are mostly based on using smaller units like phonemes, diphones or syllable-like units, and hence are not suitable for waveform concatenation.

1.3 PROSODIC FEATURES

Segmental features are those which can be derived by processing short segments of speech of duration of about 10-30 ms each. For example, pitch period, formants, linear prediction coefficients (LPC), and short time energy of the signal are some of the segmental features. Suprasegmental features are those which can be observed over longer durations of about 1 sec or more of speech signal. Initially, segmental features were used for different speech applications like automatic speech recognition, speaker recognition and speech synthesis. Systems developed for these applications use segmental features or the features derived from them like LP cepstral coefficients, delta cepstral coefficients and mel-frequency cepstral coefficients. All the above mentioned features are spectral features obtained from short segmental analysis. Performance of these speech application systems can be improved further by incorporating suprasegmental features. Duration, intonation, and stress are some of the suprasegmental features. These features are also referred to as prosodic features. With advancements in computing facilities, many researchers developed methods for acquisition and incorporation of prosodic features in most of the today's speech applications.

Here is an illustration that explains the importance of prosodic features qualitatively. If we listen to a sentence of speech by splitting the signal into a number of small segments of duration of about 20-30 ms each, we will not be able to recognize either the speaker or the linguistic content of the signal. On the other hand, if we listen to a complete sentence, even it is degraded, we can easily understand what is being spoken and many times the speaker also can be recognized. In the later case we use prosodic knowledge, which we acquire without any explicit training. Similarly, during speech production, human beings use the so called prosodic knowledge to incorporate his/her emotional state in the speech produced. This shows the importance of prosodic features. As said earlier, human beings need not require any explicit training for acquiring suprasegmental knowledge. But for a machine to use the prosodic knowledge it needs explicit specification/training.

1.4 USE OF DURATION AND INTONATION KNOWLEDGE IN SPEECH APPLICATIONS

In recent years, prosodic knowledge is used in many of the speech systems to some extent though not to its full potential. Here, we explain the importance of the duration knowledge and intonation knowledge in various speech applications. Segmental duration is one of the cues that helps to distinguish between segments, determines features of adjacent segments, distinguishes between stressed and unstressed sound units, signals phrase and sentence boundaries, and helps to indicate the presence or absence of emphasis. Major changes like resetting in F_0 contours (discussed in chapter 3) are useful in word boundary detection. This section illustrates the usage of duration and intonation in various applications of speech like speech recognition, and speaker recognition in general, and synthesis of speech, in particular.

1.4.1 Speech Recognition

Durational patterns of phonetic segments and pauses convey information about the linguistic content of an utterance. Most of the current automatic speech recognition (ASR) systems grossly under-utilize the knowledge provided by the durational cues due to vast number of factors that influence speech timing, and the complexity with which they interact.

Chung and Seneff have done studies on duration modeling [6]. They demonstrated the benefits of duration modeling in the case of full word recognition. They used duration as a post-processor in the word spotting task with a pair of highly confusable city names, "New York" and "Newark". They passed all the misspotted tokens to the post-processor, which reconsiders those words and make a final decision with the duration component incorporated. With this post-processing technique, they found that the number of confusions were reduced from 60 to 19 out of 323 words. This improvement shows the power of duration knowledge in instances where acoustic-phonetic features are less effective. Many researchers have shown that duration model is most effective when the maximum amount of lexical knowledge is provided, wherein the model is able to take advantage of the various durational relationships among the components of the sub-lexical parse structure [6].

House, Klatt, and Lahiste have shown the usefulness of the duration of the preceding vowel in automatic speech recognition [7–9]. Apart from the usage in the post-processor, as described in the full word recognition, durational knowledge is also useful in segmenting the speech signals prior to recognition. For example, in the segmentation of a connected digit utterance, the durational statistics of each one of the digits can be used [10].

As such intonation knowledge may not be directly applicable for an ASR system because by itself intonation does not contribute to speech information. But as a meta level knowledge, it can assist the acoustic-phonetic module to disambiguate alternative sequences of units similar to the usage of duration knowledge in the post-processor [6]. Lea presented an algorithm which used major fall-rise patterns in the pitch contour along with major pauses to identify sentences, clauses and major phrase boundaries for English [11]. Kohler studied the influence of fundamental frequency in rhythmic and semantic structuring of continuous speech in German [12]. From the literature, it is obvious that the properties of the intonation knowledge are not exploited to its full potential, except in a few cases.

In speech recognition, it is observed that prosodic features have been exploited rather poorly compared to the segmental aspects of speech [13]. Integration of prosodic knowledge with segmental knowledge in an automatic speech recognition system will improve the performance of the system. But it is not known how to integrate the prosodic knowledge into an ASR systems. Also methods to combine the evidence and scores obtained from different sources are not well explored.

1.4.2 Speaker Recognition

Speaker recognition is the task of identifying people by their voice. One type of recognition system is text-dependent speaker recognition. In this, the same text is uttered during training and testing phases [14]. Here, the relative duration of each word is also used for taking final decision [15]. Properties of pitch contours can be used for recognizing speakers from their voice characteristics [16]. As the properties of pitch contours will not be affected by the spectral characteristics of the recording equipment or of the transmitting channel. Hence pitch contours can be used more effectively in real life situations like recognition of speaker from telephone speech.

1.4.3 Speech Synthesis

Prosody plays an important role in speech synthesis, both in terms of naturalness and intelligibility. Prosodic structure includes modeling of duration, intensity and fundamental frequency contours. These features have to be modelled in accordance with information from segment structure, lexicon, syntax and semantics to get high quality synthetic speech.

Speech sounds of any language do not have arbitrary duration [17]. The sound units follow certain specific duration patterns in each language. Much of the research in the last decade has been geared towards discovering these patterns. Also it is known that combination of sound units into a linguistic message is not simply a concatenation process. For example, the duration value of */la/* in the word */baalak/* can not be used for the */la/* in the word */ladaki/*, as they are at different positions in these two words. Phrases, words and syllables have an internal durational structure. Hence, this structure must also be taken into account while synthesizing speech. Bunnell, Hoskins and Yarrington have shown that 50% of the natural prosody and synthetic origin sentences were judged as natural, but only 20% of the natural origin and synthetic prosody sentences were considered to be natural [18]. This illustrates the importance of prosodic features in synthesis applications.

Duration of linguistic units is very much affected by the environment in which they occur. Generally, it is observed that stressed segments are longer in duration than unstressed segments, but not simply a proportional scaling down of each of them. So in order to account for the effects described above, we need a good duration model for a text-to-speech system to deliver good quality synthesized speech.

Several studies are reported in the literature that examine the durational behavior of speech sounds in connected and continuous speech [19, 20]. Huggins has examined the use of timing information in identifying the various speech sounds in continuous speech. These studies may give useful clues for both speech synthesis and speech recognition. A study on vowel and consonant durations in English at Bell Laboratories was done by Umeda [21, 22]. A text-to-speech system, based on articulatory model, was also built using Umeda's results.

Many durational rule systems exists for English [21–26]. The rules contained in these systems are similar, but differ in the way of describing the interacting phenomena. For example, rules in [26] and [27] rely heavily on multiple stress levels conditioned by syntactic category and word frequency. In recent years the study on intonation knowledge is mainly motivated by its application in speech synthesis. Unlike other speech systems, in synthesis systems we are able to appreciate the importance of intonation model easily through perceptual examination.

1.5 SCOPE OF THE PRESENT WORK

In the problem of speech synthesis, there are a large number of issues to be addressed at different levels. At the text analysis level, issues like how to assign the prosodic features to the basic units depending on the context of the unit, and also how to determine the various contexts. Issues in developing new models for synthesis, the sizes of the synthesis units and various signal processing intricacies are addressed at the synthesis level. At the level of prosodics, issues relating to acquisition of prosodic knowledge and incorporation of the knowledge into the synthesized speech signal are to be addressed.

In the context of prosodic modeling, the present work is aimed at the acquisition of duration and intonation knowledge from natural read style of speech read by news readers. Studying prosodic properties from normal conversational speech is difficult as these properties change significantly due to rate of speech, attitude towards what is being spoken, age and gender of speaker, etc. Also, we may not be able to make a valid generalization in such cases. Hence, we restricted our studies to read style of speech only. In the present work, issues relating to the acquisition of duration and intonation knowledge from continuous speech for two Indian languages are addressed. Duration characteristics of the basic sound units of Hindi language in continuous reading style of speech are studied and are compared with the results stated in the basic system. Also, The duration and intonation properties in continuous speech of Telugu are explored.

1.6 ORGANIZATION OF THE THESIS

The main focus of this thesis is acquisition of prosodic knowledge from continuous speech in Telugu and Hindi and to make a comparative study of knowledge in these two languages. The thesis is organized as follows:

Chapter 2 gives a review of the different approaches used for modeling duration and intonation knowledge for a TTS system. Here, only some of the most popular models used in the area of speech synthesis are described.

Chapter 3 explains the details of the basic TTS system based on which the present work is carried out. The issues addressed during the development of the basic system are explained in detail.

Chapter 4 describes the acquisition of duration knowledge from continuous speech data in read style in Hindi. Differences in durational characteristics of the basic units in continuous speech in read style and utterances in a limited database is explained in this chapter.

Chapter 5 is devoted to the analysis of the durational characteristics of the basic units of continuous speech in Telugu. Comparative study of the durational knowledge of Telugu and Hindi are studied in this chapter. The effect of positional and contextual factors among different speakers and genders are also studied.

Chapter 6 deals with the study of intonation. The intonation characteristics of continuous speech in Telugu is studied. Comparative study of intonation patterns of Hindi and Telugu is also made.

Chapter 7 concludes the thesis by summarizing the work.

CHAPTER 2

REVIEW OF APPROACHES TO DURATION AND INTONATION MODELING FOR TTS

2.1 INTRODUCTION

Prosody has an important supporting role in guiding a listener's recovery of the basic message and a starring role in signalling the speaker's attitude. Prosody is a complex weave of physical and phonetic effects that is being employed to express attitude, assumptions, and attention as a parallel channel in our daily speech communication. The semantic content of a spoken message is referred to as its *denotation*, while the emotional and attentional effects intended by speaker are part of the message's *connotation*. Prosody includes fundamental frequency, duration and stress. We discuss the fundamental frequency contour generation and duration assignment separately, because it is convenient to separate them when building systems. But we should not forget that all the prosodic features are highly correlated in natural human speech. The effect of stress is not nearly as important in synthesizing speech as the effect of the other two factors. Hence we concentrate on duration and intonation in the present study.

The models used for prosody control range from rule-based methods to trainable data-based methods. Both these extreme ends of the continuum range of methods have their own merits and demerits. The data-based methods are generally dependent on the quality and quantity of the available training data. On the other hand, the generalization is very high in rule-based models. As the present study is mainly concerned with segmental durations and intonation, we discuss some of the existing models for duration and intonation modeling in the following two sections. The models concerning loudness, pause insertion, and pause length prediction are not discussed. The models described here represent only a fraction of the prosody models in existence, and were chosen because of their influence on the development of TTS systems, in particular, and prosody research in general.

This chapter is organized as follows: In Section 2.2, we review various modeling techniques used for duration modeling. Different approaches followed for modeling the intonation knowledge for a TTS system are reviewed in Section 2.3. The last section summarizes the review.

2.2 APPROACHES TO DURATION MODELING IN A TTS SYSTEM

Different approaches have been developed for modeling duration of units in the development of TTS systems: knowledge-based models [28], data-based approaches [29], linear statistical models [30] and neural network approaches [31].

2.2.1 Knowledge-Based Models

Segmental duration modeling used by Klatt in MITalk system, where durations are predicted with rules such as *if vowel v is stressed, subtract* m_v *ms from the intrinsic duration* i_v*, multiply by 1.4, and add* m_v [32]. The mathematical structure of this rule is knowledge-based, and the values of parameters m_v and i_v are obtained via statistical estimation. The knowledge involved here is reasonable as it is derived by manually analyzing various contexts.

Klatt's model was based on information presented in phonetic literature about factors affecting segmental duration. The duration of each phoneme was estimated according to the equation

$$D = (D_i - D_m) * \frac{P}{100} + D_m$$

Where D_i and D_m are the inherent and minimum duration for the phoneme respectively. P is the shortening in percent of the duration change which is determined by the rules themselves. The minimum duration is dependent on speaking rate. The inherent duration is the base for the rules of the knowledge base. In his model, Klatt used ten rules that were based on the effects of phonetic environment, emphasis, stress level, etc. on the current phoneme. Each rule adjusts the P term in a multiplicative manner, and the final result is the sum of the effect of final rule and the product of the effects of the rest of the rules. Klatt's model is distinctly different from Umeda's model [33]. Umeda's model involves fixing a number of parameters in an equation to predict the vowel durations. This process is cumbersome.

The French text-to-speech system [34] bases its set of rules on the study performed by Shaughnessy [35]. In his rule-based model, he proposes a *base duration* for each unit. His rule-based model is based on numerous readings of a short text containing all phonemes of French. Each rule then specifies a percentage factor to modify the base duration, in case the conditions for applying the rule are satisfied. Thirty three rules accounted for the modification of a mean segment duration according to segment position, segment type, and context of the phoneme. The duration of sound classes not involving prepausal lengthening was predicted with a standard deviation of 9 ms by this model. But this model was less accurate in predicting the durations for prepausal vowels. Also, the model was not able to predict silent inter-lexical pauses. Another significant study on modeling duration for French TTS system was done by Bartkova [36]. He has developed a set of rules using data obtained from a number of speakers. This was different from earlier studies of Klatt and Shaughnessy, where the rules in the knowledge base were obtained from speech corpus pertaining to a single speaker. Bartkova's model is also based on *base duration* similar to Shaughnessy's, but the exact equations for predicting durations are different. Also, Bartkova's model takes into account the variations of stress and the variations of speech rate. A comparison of predicted and measured durations in ten sentences results in a mean difference on segmental duration of $\pm$ 15 ms.

In Chinese and Japanese text-to-speech systems, more emphasis is placed on pitch concatenation rules, due to the tonal nature of these languages. The Chinese TTS systems uses a simplified approach for duration rules developed by Lee [37]. In this model, a fixed duration is assigned for each syllable. The model has a few rules for change of this duration depending on the context. Apart from the above mentioned languages, models of this approach have been developed for many other languages including Swedish [38] and Brazilian Portuguese [39].

2.2.2 Data-Based Approaches

The availability of large speech corpora and means for analyzing them has made some researchers to propose statistical approaches. In this approach, general purpose statistical techniques are used to extract parameters from a training database. A typical example for this approach is the usage of classification and regression trees (CART) for the prediction of segmental durations proposed by Riley *et al.* [40–42]. The principle of CART is to construct a binary branching tree whose inputs are instances of phones which are fed in from the top node. The phones then pass through the arcs satisfying their constraints. Basically, the algorithm clusters durations according to their context. Generally, the factors that are used include stress level of the current phoneme, position of the unit in the word, position of the unit in the phrase, etc. Vectors for predicting the durations are formed with the factors affecting the duration of basic units as the components. The algorithm takes combinations of segmental durations and vectors for duration prediction as input, and sets up the tree by branching on vector components so as to minimize the variance at each terminal node. Increase in data storage, computational speed and new algorithms seem to make data-based approaches attractive. At the same time, it has become apparent that these approaches have fundamental shortcomings in generalizing from unit types seen in the training data to unseen unit types. Maghbouleh found that a knowledge-based approach generalized much better than CART to test material drawn from a different corpus than the training materials [43]. He showed that, even when CART was given two orders of magnitude more training data than the knowledge-based approach, the difference in performance hardly decreased.

2.2.3 Linear Statistical Models

The sum-of-products (SoP) models proposed by Jan van Santen are an example of the linear statistical models. SoP model is linear and is based on a collection of equations that are determined according to prior phonetic and phonological information as well as information collected by analyzing data. This model of segmental duration accounts for data imbalance and factor interaction problems. The model is so called because the segment duration is represented as a sum of factor's product terms.

The SoP model can be explained mathematically as follows. Suppose $\mathcal{F}$ is a set of factors $F_1, F_2, ...F_n$ that affect segment duration. Then the SoP model ϕ is a mapping from a space of all possible combinations of factor levels represented by a Cartesian product to the space of real numbers $\Re$

$$\phi \rightarrow \Re : \phi(F) = \sum_{i \in K} \prod_{i \in I_i} S_{ij}(f_j)$$

Where K is a set of integers, I_i are subsets of K, and $S_{ij}(f_j)$ is a factor scale for a factor F_j having a level f_j. The model is based on the assumption that most of the factors that affect duration have the property of *directional invariance*. This property states that the factors interact in the magnitude of their effect rather than in the direction of their effects, i.e., the direction of the effects of a factor is not affected by the other factors.

2.2.4 Neural Network Approaches

Campbell has used neural network model for predicting phone durations [31]. The intention of using neural networks is that they can learn the underlying complex interactions among the factors affecting duration of a syllable. Also if the networks are able to represent the underlying interactions, then they can also perform better with unseen data. Therefore a better generalization can be achieved.

Campbell computed a feature vector for each syllable which consisted of information about the syllable length in terms of number of phones, the nature of syllable nucleus, stress level, etc. Then the phone durations within the predicted syllable were determined by their elasticity. The elasticity is computed from a normalized duration which is calculated by *subtracting the means and expressing the residuals in terms of their standard deviations to give a zero with unit variance for each phoneme distribution [31].* He, then expressed the amount of lengthening or compression with single constant in majority of the cases. The constant was obtained from database by solving the equation for k:

$$d = \sum_{i=1}^{n} exp(\mu_i + k\sigma_i)$$

where μ_i and σ_i are the mean and standard deviation of the log-transformed duration for the realization of phoneme class i, and d is the duration of a syllable with n phones in ms, and is computed from the database.

2.3 APPROACHES TO INTONATION MODELING IN A TTS SYSTEM

Many methods have been proposed for the generation of fundamental frequency contours for TTS systems. These models can be categorized into two groups: phonological models and phonetic models. Phonological models use a phonological representation of F_0 contour. They use an inventory of abstract phonological categories, with each category having its own linguistic function. Here, the model described under rule-based category is an example of phonological models.

On the other hand, the phonetic models are developed from acoustic (F_0) data. They attempt to describe F_0 movements, and they are usually continuous in nature. Often, the description of F_0 movements is linked in some way to the linguistic level. The tilt model, Fujisaki model, and most of the neural network models come under this category.

2.3.1 Rule-Based Models

In rule-based models, the input text is first analyzed to extract linguistic cues. Phonological rules are then used to generate required information. One of the rule-based system based on an in-depth analysis of Tone and Break Indices (ToBI) labeled reference corpus has 60 rules for generating the F_0 contours [44]. This rule set can realize each of ToBI accent and boundary in different phrasal contexts. This model belongs to tone sequence model that follows a traditional phonological description of intonation. The tone sequence models interpret the F_0 contour as a linear sequence of phonologically distinctive units, that are local in nature. The advantages of rule-based systems are: they provide a good generalization in handling unseen sequences. As the rules are mostly derived by manual analysis of databases, the information is very much reliable and accurate. But the only disadvantage is that, it requires large manual effort in deriving rules.

2.3.2 Tilt Model

Tilt model is a bi-directional model proposed by Taylor that gives an abstraction for the F_0 contour directly from the data [45]. The basic unit of this model is the *intonational event*. The basic types of intonational events are pitch accents and boundary tones. Pitch accents are F_0 excursions associated with syllables, and boundary tones are rising events which occur at the edges of phrases. The amplitudes and durations of rising and falling functions combined with the position of the pitch accent/boundary tone constitute the basis for the Tilt description. These continuous valued parameters, known as *tilt parameters*, are determined from examination of the local shape of the event's F_0 contour. The importance of this model is in it's ability to capture both phonetic and phonological aspects of intonation.

2.3.3 Fujisaki's Model

The Fujisaki model was developed for generating F_0 contours of Japanese words and sentences [46]. The model is widely used in TTS systems for the languages like Japanese, German, English, Greek, Polish, Spanish and French.

This model is based on the assumption that any F_0 contour can be considered to be consisting of two kinds of elements: the slowly varying component which consists of one or more slowly varying components, and a more rapidly varying accent component. These components are said to be related to the actions of the laryngeal muscles.

The model is driven by a set of commands in the form of impulses and a set of stepwise functions. These commands are fed to a critically damped second-order linear filters, and are then superimposed to produce the final F_0 contour in the logarithmic domain, which is then transformed to absolute pitch values.

2.3.4 Neural Network Approaches

The neural network approach is proposed by Traber in [47]. Neural network models have been developed for many languages using feedforward neural network (FFNN) models and recurrent neural network (RNN) model. For the prediction of global F_0 shape in Japanese sentences, Sagisaka has used a three layered FFNN. The input to the network was a 14 to 18 variable dimension feature vector depending on factors affecting the F_0.

Chen developed another model for generating F_0 contours for Mandarin TTS system [48]. He proposed a four layered RNN model for this purpose. The input layer and the first hidden layer operate with a word-synchronized clock to represent current word phonological state within the prosodic structure of the input text. The second hidden layer and the output layer operate on a syllable-synchronized clock and use the outputs from the preceding layers, along with the additional syllable-level inputs directly fed to the second hidden layer to generate desired the desired prosodic parameters.

2.4 SUMMARY

In this chapter we discussed some of the prosodic models that are used in speech synthesis. Of the different models discussed, the rule-based models are good in generalizing the effects of various factors on the prosodic features. In the present study, rules for duration modification and generation of F_0 contour are developed.

CHAPTER 3

DESCRIPTION OF THE TTS SYSTEM FOR HINDI

A basic framework for Text-to-Speech (TTS) system is formulated, and a TTS system was built according to the framework using parameter concatenation model [49] for Indian languages. The basic units in the existing TTS system are consonants (C), vowels (V), and consonant-vowel (CV) units. Typically these are the characters of the language. Speech signal of each of the basic units is extracted from the utterances of the carrier words containing the desired basic unit in the middle position. This signal is parametrized using linear predictive coding. Formants and bandwidths, and pitch and gain contours of each basic unit is stored in the database along with the linear prediction coefficients (LPC). As the parameters can be changed easily, it is easier to incorporate prosodic features. Parametric representation is highly flexible and needs much less storage space compared to the direct waveform representation.

The design of the TTS system uses modular approach. Figure 3.1 shows the block diagram of TTS system. Modular design approach enables us to make changes in any of the modules and then can easily be integrated with the rest of the system.

The input Hindi text is given as an equivalent phonetic transcription in English script to a parser that extracts a sequence of basic units. Coarticulation rules that operate on the segments near to the boundary between two successive basic units are activated. Gain contour is smoothened at boundaries between adjacent basic units. Pitch contour is modified to incorporate the intonation knowledge. The modified gain and pitch contours are used to generate an excitation signal. The LPCs are used to model the vocal tract system and Fant's excitation model [50] is used to generate excitation source signal. Stored parameters are excited with the source signal to generate the synthetic speech waveform.

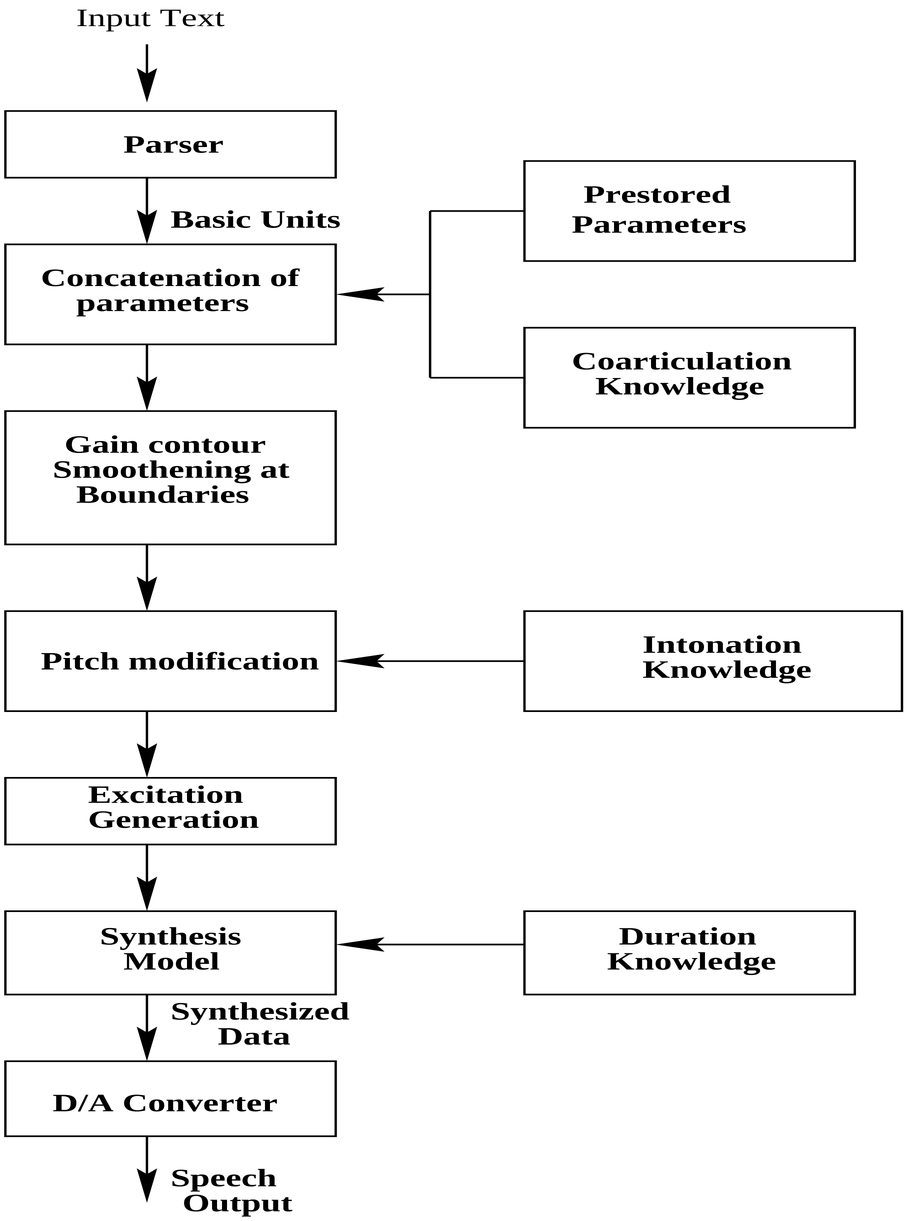

Fig. 3.1: Block diagram of the TTS for Hindi.

3.1 PREPROCESSOR AND PARSER

Preprocessor and parser along with the text analysis module constitutes the front end of the TTS system. During the preprocessing stage, the input text is analyzed to locate nonphonetic strings such as numerals, date, abbreviations, and some special symbols. Then, all these strings are replaced by their spoken form. The preprocessed text is passed on to the parser module in order to extract a sequence of the basic units from the input text. The choice of the basic units involves tradeoff between the size of memory needed to store the basic units and the computation during synthesis. In this system character is used as a basic unit, which is an equivalent to the orthographic representations of speech sounds in Indian languages. A character in an Indian language is close to a syllable. So characters are also referred to as syllable-like units. A character in Hindi language represents a speech sound in the form of a Consonant (C) or Vowel (V) or CV or cluster CCV unit. The cluster characters are generated from the constituent CV combinations and other consonants. This particular choice of syllable-like units as basic units, makes the implementation of parser easier.

3.2 PRODUCTION MODEL

To synthesize speech from the stored parameters, LP synthesis technique [51,52] is used for consonantal region, and a cascade formant synthesizer [3] for the vowel part of the basic unit. The basic production model for parametric speech synthesis is shown in Figure 3.2. It consists of a time varying digital filter excited by time varying excitation signal. The excitation signal is a sequence of periodic pulses for voiced regions of speech and white noise for unvoiced parts of speech signal. Excitation signal is fed to the time varying digital filter to synthesize the speech signal. The digital filter is modelled by LPCs during the synthesis of consonantal regions and by a set of formants and bandwidths during the synthesis of vowel regions. The choice of excitation model also affects the quality of synthetic speech [4, 53]. In the existing system, the Fant's excitation model is used.

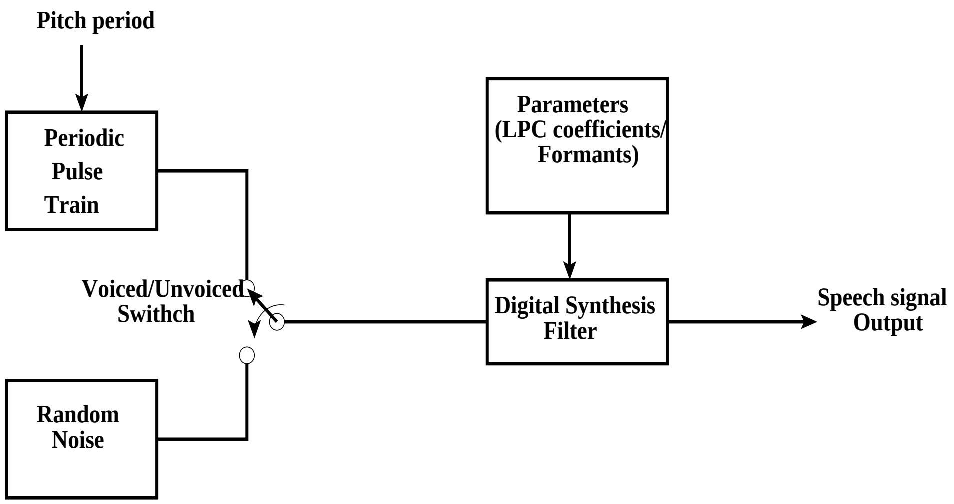

Fig. 3.2: Basic synthesizer block diagram.

Incorporation of knowledge extracted from segmental and suprasegmental domains is very important to produce intelligible, natural sounding synthetic speech. In the following sections we describe how this knowledge is acquired in the existing system.

3.3 COARTICULATION

To avoid discontinuities and to ensure a smooth transition between adjacent segments at the boundaries between adjacent basic units, coarticulation knowledge is incorporated. This improves naturalness and the fluency of the synthetic speech. The phenomenon of coarticulation involves changes in the articulation and acoustics of the basic unit due to its context. Coarticulation has its anticipatory and carry over components [54, 55].

In the TTS system, the acoustic manifestations of various coarticulation effects considered are the transitions (a) from C to V in a CV unit, (b) from V to C in a VC sequence, (c) from V to V in a VV sequence and (d) nasalization of vowels. Then nature of transition pattern in a VC sequence depends on the articulatory features of the consonant [7]. The VC transitions are grouped in to six distinct classes based on place of articulation. Each class is further divided into subgroups depending on the manner of articulation. Each subgroup contains formant transition patterns for five different vowels. The coarticulation in a CC cluster is achieved through consonant cluster rules. The CV transitions are already preserved in the basic unit itself.

Incorporation of the acquired coarticulation knowledge is done by the application of a set of rules to modify the default parameters of the basic units before synthesizing speech. Since coarticulation modifies most of the parameters, application of rules is merged with the process of synthesis. Stored representations of all the basic units in the input text are loaded into a temporary buffer, so that the parameter modification is done by the subsequent processing blocks. Figure 3.3 shows the block diagram to explain how coarticulation knowledge is activated in the existing system.

3.4 SEGMENTAL DURATIONS

Segmental duration depends on the inherent properties of the basic unit and on the phonetic and structural constraints [28, 56]. Durations of various basic sound units vary considerably due to factors such as position of the unit in the text, the nature of the preceding and the following units, speaker (male vs female), speaking style (reading vs speaking), speaking rate (fast vs slow) and the psychological state (happy, anger or fear) of speaker. Apart from the above factors, durations of the basic units are also affected by factors like native language of the speaker and accent of his/her speaking style. For the sake of simplicity, durational effects were studied individually as effects due to position (POS), syllable boundary (SYB), prepausal lengthening (PPL), postvocalic consonant effect (PVC), place of articulation (POA), changes in cluster environment (CCL), semantic novelty (NOV), and polysyllabic shortening (PSS). POS, SYB, and PPL account for the effects that modify durations of the basic units depending on their position in the text. Contextual effects were accounted by PVC, POA and CCL. In the following sections we describe how these effects were analyzed and incorporated in the existing system. These descriptions pertain to a limited number

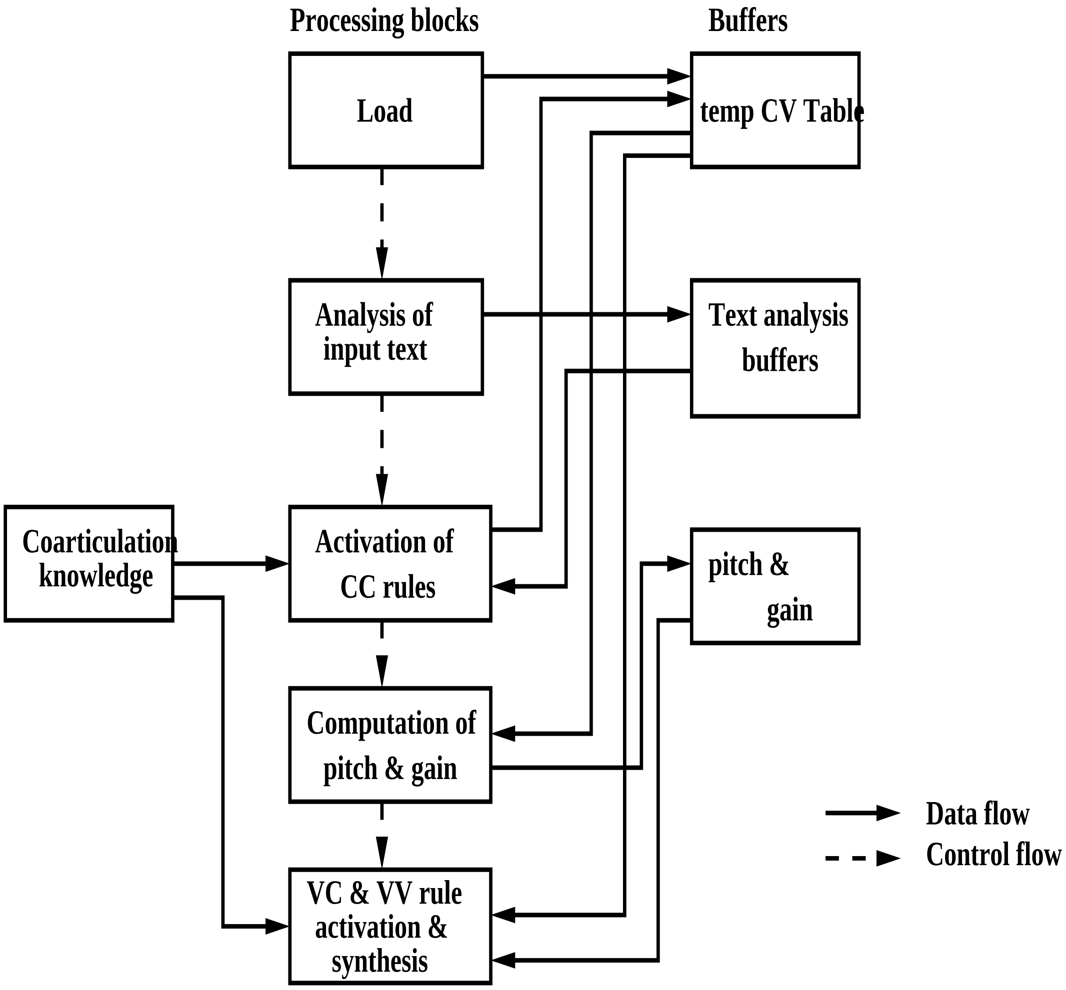

Fig. 3.3: Block diagram for the incorporation of coarticulation knowledge.

of nonsense words in controlled reading environment by a single male native Hindi speaker.

3.4.1 Position Effects (POS)

In the existing TTS system for Hindi, the POS results were obtained after analyzing durations of about 50 basic units. To arrive at these results, three nonsense words were formed for each basic unit. The first word contained the basic unit in the word medial position. The second and third words contained the basic unit in the word beginning and word ending positions, respectively. The durations of the basic units in all the three words were calculated. The percentage change in the duration of a basic unit in the word initial or the word final position over that in the word medial position was calculated. Based on the results rules were formulated for the position effects. From the results it was observed that the basic unit is more lengthened in the word final position than in the word beginning position, which in turn is longer than in the word medial position. To study the POS effect alone and to nullify the PVC effect, the basic unit is followed by the same category of unit in each of the three words.

3.4.2 Syllable Boundary Effect (SYB)

The SYB result was obtained by analyzing the durations of 24 basic units. For each basic unit two nonsense words each containing three characters were formed. The first word contained the basic unit in the word medial position. The second word contained the basic unit in the word initial position with a syllable boundary following it. The percentage change in duration of the basic unit in the second word over that in the first word is calculated. This result is the combined effect of position and syllable boundary. The POS effect is removed to get the SYB effect alone. This also was formulated in a rule.

3.4.3 Prepausal Lengthening (PPL)

The PPL results were obtained after analyzing durations of 20 basic units. This study was performed on natural speech data. The database consists of about 25 sentences spoken at natural rhythm and intonation so that PPL effect was clearly observed. The duration of a basic unit before a pause (due to either phrase boundary or breath group or sentence ending) was measured. The duration of the basic unit when it is followed by an unvoiced unaspirated stop was also measured. The percentage change in the duration of the basic unit in the former over the latter was calculated. This is the combined effect of POS and PPL. The POS effect was removed to get the PPL effect.

3.4.4 Post Vocalic Consonant Effect (PVC)

The PVC results were obtained after analyzing 20 basic units in different contexts. The basic unit could be either a vowel or a CV unit. This effect was examined for the vowels /a/, /i/ and /u/, and it was later verified for the remaining vowels. For each basic unit the effects due to stop-consonant-vowel (SCV), non stop-consonant-vowel (non SCV) classes of sounds were measured. For calculating the PVC effect due to SCV class, four nonsense words were formed with the basic unit in the word medial position. The unit is followed by four different categories of PVC. They are (a) Unvoiced Unaspirated (UVUA), (b) Voiced Unaspirated (VUA), (c) Unvoiced Aspirated (UVA) and (d) Voiced Aspirated (VA). The percentage change in the duration of the vocalic portion of the basic unit in (b), (c) and (d) over (a) was calculated. From these results, rules were formulated for each category of PVC. Similarly, for calculating the PVC effect for non SCV class of sounds, five nonsense words were formed with the basic unit in the word medial position. Four different non SCVs considered were trill, fricative, nasal and semivowel. POA and CCL effects handle coarticulation. The other effects such as PSS and NOV were not analyzed.

Production system approach was followed to represent the duration knowledge into TTS system. In this approach, each rule is an independent fragment of knowledge, and it does not rely on the correctness of other rules. This enables successive updating. Activation of the rules was done by a rule interpreter. If more than one rule is fired, then the rules combine multiplicatively to get the combined effect. The set of rules developed based on the analysis of durations of basic units is given in Appendix A.

3.5 INTONATION

Intonation pattern is defined as the variation of fundamental frequency with time [57, 58]. Like in many other languages, intonation pattern in Hindi also shows some regular features. F_0 contour of declarative sentences shows a declining tendency with time. But for interrogative sentences, fundamental frequency contour gradually rises towards the end of the sentence. F_0 contour gets modified across major syntactic boundaries which is called resetting of the F_0 contour.

The above conclusions were drawn after analyzing 500 Hindi sentences. Speech was read by two adult native male speakers of Hindi [59]. Speech was digitized to 12 bits/sample at 10 KHz sampling rate. A 256 sample analysis window with a shift of 64 samples was used for extracting pitch. The algorithms for pitch extraction were based on simple inverse filter tracking [9] and properties of group delay functions [60, 61]. The following properties of intonation were observed and incorporated in the TTS system.

3.5.1 Declining/Rising Tendency

The fundamental frequency contour consists of peaks and valleys. In the case of declining F_0, these peaks and valleys were fitted between two abstract lines - a top line and a base line. These lines were drawn nearer to all maxima or all minima of the F_0 contour. Figure 3.4 shows a typical pitch contour plot for a declarative sentence in Hindi.

In the case of interrogative sentences, two patterns were observed. One pattern for *yes-no* type questions, and the other pattern for *question word* type questions. In yes-no type questions, the F_0 contour shows a continuous rising pattern from beginning to the end of the utterance. But in the question word type, initially F_0 contour shows a declining tendency up to the question word, and then rises towards the end of the utterances. Rising or declining tendencies of F_0 gets reset across major syntactic boundaries with a pause. Also, it was observed that the global pattern of the F_0 contour for every intonational phrase is similar to that of a simple sentence. These observations in the intonation pattern were incorporated into a TTS system through a rule-based approach.

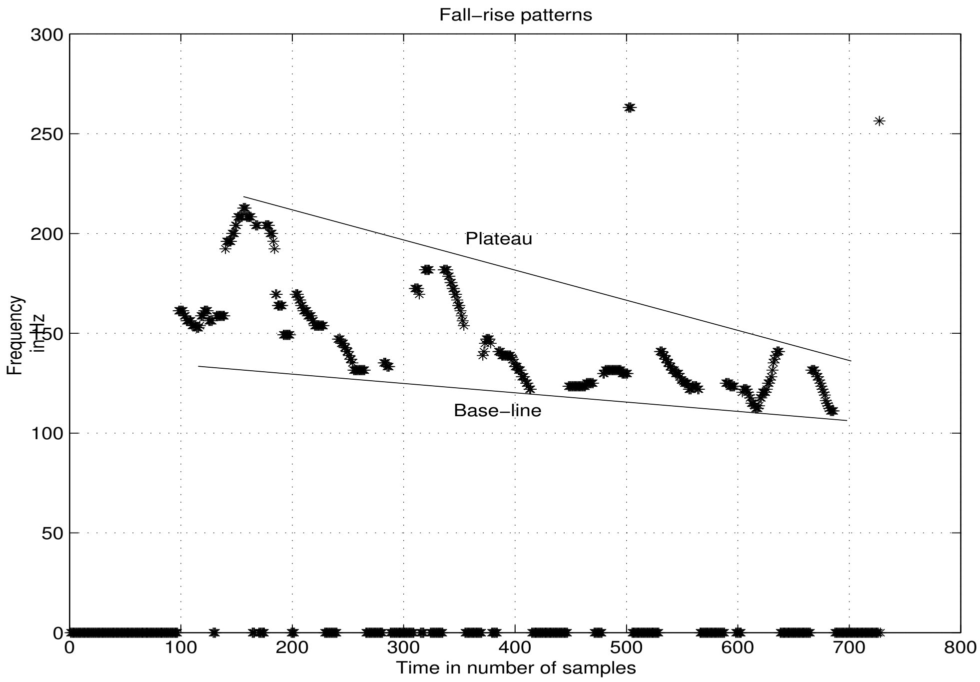

Fig. 3.4: Typical pitch contour for a phrase of a declarative sentence in Hindi.

3.5.2 Resetting of F_0 Contour

The declining characteristic of the F_0 contour exhibits some changes across major syntactic boundaries. This is imposed by both physiological factors of speech production mechanism and syntactic properties of language. The subglottal air pressure drops during phonation. Speakers are forced to give significant amount of pause between the the utterances to compensate for the drop in the subglottal air pressure. During a pause, the subglottal air pressure is built-up again, and is manifested as resetting in the F_0 contour. This tendency was seen for continuous speech in Hindi also.

3.6 SUMMARY

In this chapter, we described how various issues for the design of a TTS system were addressed in a system for Hindi. We explained how various prosodic features were studied. Also, we discussed how each of the prosodic features were analyzed and quantified in the form of rules that were incorporated in the TTS system. This chapter presents a detailed description of the basic TTS system for Hindi language. The present work is aimed at improving the quality of synthetic speech from the basic TTS system by refining the prosodic rules.

CHAPTER 4

DURATIONAL CHARACTERISTICS OF BASIC UNITS IN CONTINUOUS SPEECH FOR HINDI

The quality of synthetic speech both in terms of naturalness and intelligibility depends upon its prosody modeling [62–65]. It was shown that the information such as syntactic structure, lexical and propositional semantics, and pragmatic context is required to predict the appropriate prosodic realization of the input text [66–68]. Hence, we need a good text analysis module that can predict appropriate prosodic features from the text. The prosodic features derived from natural utterances are superior compared to those derived from utterances in constrained sequences. With this motivation, we analyzed natural utterances of read speech taken from the broadcast news data to model the durational characteristics of basic units of Hindi speech.

This chapter is organized as follows: A parser module is developed based on ITRANS code for Hindi [69] and it is explained in Section 4.1. Section 4.2 gives a description of the database used for the study. Comparison between the durational changes of the basic units in continuous Hindi speech and the utterances in constrained sequences due to positional and contextual factors is discussed in Sections 4.3 and 4.4, respectively. Sections 4.5 and 4.6 are devoted to the study of similarities and deviations in the durational properties of continuous speech for three speakers. In Section 4.7, the scheme followed to incorporate the duration knowledge into the TTS system is discussed.

4.1 PARSER

Any Text-To-Speech (TTS) conversion system has to handle the unrestricted text input at the input stage [70, 71]. For a TTS system to produce a good quality speech output, it has to extract the relevant information from the input text. If we can acquire useful linguistic as well as nonlinguistic information from the text, then we can obtain better prosodic quality in the output speech. In a TTS system, the first module that handles the unrestricted text input is the *text analysis* module. For a TTS system to produce high quality speech it needs well specified prosodic features corresponding to the input text. Also, an unambiguous parsing of text into a sequence of basic units leads to better intelligibility of the output speech [72].

In the context of Indian languages which doesn't have keyboards designed for them, the text input is generally given in a phonetic equivalent of the English alphabetics through the standard 7-bit ASCII keyboard. In phonetically equivalent representation of Indian language scripts, different conventions are followed. Hence, the sequence of basic units cannot be extracted uniquely. For example, the word *bhaarath* can be parsed into */bhaa/, /ra/, /th/* or */bha/, /a/, /ra/ /th/*. Here, only the former sequence of basic units is correct and not the latter sequence. In such ambiguous cases, the intelligibility will be lost. In order to overcome these kinds of ambiguity, we developed a parser module that can be used for all the Indian languages for extracting unambiguous sequence of basic units from the unrestricted text input, after it is preprocessed. The parser is based on ITRANS code, which is a method of coding the Indian language scripts using the English alphabets [69].

The input text is preprocessed before it is given to a parsing module. In the preprocessing stage, all the nonorthographic items such as numbers, dates, and abbreviations are converted into their equivalent orthographic form. For example, the abbreviation */DÃ./* (Dr.) is expanded to its extended form */DAkTar/* (Doctor) and the 420.13 is expanded to */chAr sau bIs dashamlav Ek tIn/* (Four hundred and twenty point one three). The schematic of the parser module is given the Figure 4.1.

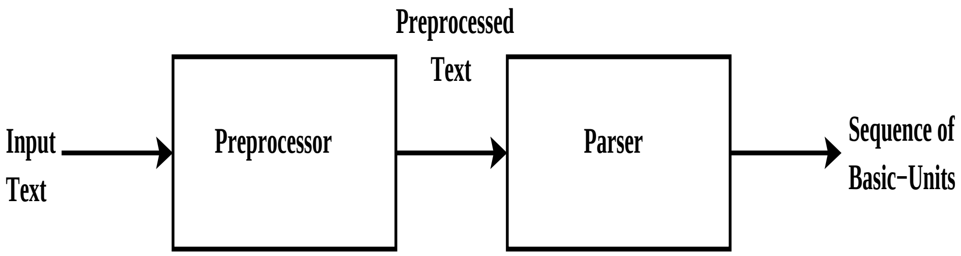

Fig. 4.1: Parser module based on ITRANS code.

For example, the sentence */ harI, sa. 1950 mE Ap kyA karte thE?/* (Hari, what were you doing in the year 1950?) have a numeral denoting the year and another abbreviation for */san/* (year). The preprocessing module converts the abbreviation */sa./* into */san/* and the numeral *1950* to */unnis sau pachAs/*. Finally, the output of the preprocessor module corresponding to the given sentence is */harI, san unnis sau pachAs mE Ap kyA karte thE?/*. The preprocessed text will have orthographic forms equivalent to all the spoken forms of the processed input. Then, the preprocessed text is given to the parser module to extract the sequence of basic units. The sequence of basic units extracted for the present sentence will be, */ha2/, /rI2/, /COMMA/, /BLANK/, /sa2/, /n2/, /BLANK/, /u2/, /n2/, /nI2/, /s2/, /BLANK/, /sau3/, /BLANK/, /pa2/, /chA2/, /s2/, /BLANK/, /mE/, /BLANK/, /A2/, /p2/, /BLANK/, /k2/, /yA2/, /BLANK/, /ka2/, /r2/, /te2/, /BLANK/, /the3/, /QMARK/*. Each parsed unit is a basic unit of the system. The indices "2" and "3" specifies whether the basic unit is a standalone basic unit, or it occurs in a word with more than one basic unit. The parsed units of words with more than one CV unit are assigned with index "2" and those of monosyllabic words are assigned with index "3". The */BLANK/* units are used to locate the word boundaries by the parser module. The special units like */COMMA/, /BLANK/, /QMARK/* etc., are used by the text analysis module while deciding the syntactic boundaries, and also while specifying the prosodic parameters. The parser module is based on ITRANS code and can be used for all the Indian languages.

The following two sections are devoted to the study of variations in duration exhibited by the basic units of continuous speech in Hindi. A study was carried out in [1] on the effects of various factors on durations of the basic units in the utterances of constrained sequences. The effect of the same factors in continuous speech is dealt with in the present chapter. The factors affecting the durations were divided into two categories: positional and contextual factors. The positional factors affect the duration due to a change in the position of the unit with respect to its position where the base duration of the unit is measured. Base duration of a basic unit is the reference duration with respect to which all the modified durations are expressed as percentage changes. The contextual factors affect the duration due to a change in the nature of the preceding and/or the following sound units.

4.2 DESCRIPTION OF THE DATABASE

Speech database consisting of *DoorDarshan* broadcast news data is analyzed for this study. Speech data is taken from three news bulletins each one of about 15 minutes duration, read by three different speakers. Two of the three bulletins were read by male speakers and the other by a female speaker. Each bulletin consists of 200 - 250 sentences, each of about 5 sec. The speech signal is sampled at 16 KHz and stored as 16 bit integers. All the bulletins are transcribed into text using ITRANS code convention for Hindi.

4.3 EFFECT OF POSITIONAL FACTORS ON DURATION

4.3.1 Word Final Position

As in many other languages [36,37,73], continuous speech in Hindi also lengthening in the durations of the basic units at the word final position is observed. The increments in the durations are different for consonant-vowel (CV) units and units in consonant-consonant-vowel (CCV) clusters. In a CCV cluster like */pra/*, the consonant */p/* is denoted as CCV_1 and the CV unit /*ra*/ is denoted as CCV_2. In the case of CV units, the variation of the duration in word final position is nearly equal for all the three speakers. On average the variations in the duration of basic units with respect to the base duration are found to be 38% and 29% for male speakers and 37% for female speaker. CV units in the word final position of continuous speech shows 5% more variation compared to utterances of constrained sequences. Also, the CCV_1 units in CCV clusters have shown similar effect. Table 4.1 lists the average percentage variations with respect to the base duration in the new rule-base and in the rule-base of the basic system. The values in the second column of the Tables 4.1 to 4.6 are the results obtained from the analysis of durations in the utterances of constrained sequences reported in [1]. All the values given in the Tables are percentage changes with respect to the base durations of the basic units.

Table 4.1: Changes in duration of basic units in the word final position.

Word final position	Percentage change with respect to base duration			
	Rules of the basic system	New rules		
		Speaker 1	Speaker 2	Speaker 3
CV unit	30	38	29	37
CCV_1	15	15	10	13
CCV_2	30	39	42	46

4.3.2 Word Initial Position

The duration characteristics of CV units and CCV cluster units in word initial position are shown in the Table 4.2. From the values in the first and last rows of the this table, it is clear that significant difference is there in the lengthening of CV units not in a cluster and those in clusters. In the case of speaker 1, the CCV_2 has more variation than CV unit. But in the case of the other two speakers, CV units are lengthened more compared to that for CCV_2. The pure consonant part of the clusters, that is CCV_1 units are observed to change by smaller amounts than CV and CCV_2 units.

Table 4.2: Changes in duration of basic units in the word initial position.

Word initial position	Percentage change with respect to base duration			
	Rules of the basic system	New rules		
		Speaker 1	Speaker 2	Speaker 3
CV unit	10	19	20	24
CCV_1	5	12	14	10
CCV_2	10	28	17	21

4.3.3 Syntactic Boundaries

The syntactic boundaries studied are: phrase boundary, sentence ending position, end of a breath group, and syllable boundaries. The percentage changes in the duration of the basic units at these syntactic boundaries is listed in the Table 4.3. The units at the sentence ending position exhibit a large change compared to the change by the units at the other boundaries in the case of speaker 2 and speaker 3. But, speaker 1 showed larger lengthening in the units at the end of the breath groups rather than in the units at the sentence ending position. The effect of syllable boundary is much less on the duration compared to the effect due to the other syntactic boundaries in the speech for all the speakers. We can also observe that, both in utterances of constrained sequences and in continuous speech, sentence ending position effected to a large extent. This result is consistent with the results obtained in [74, 75].

4.3.4 Penultimate Position to Syntactic Boundaries

The legal sound units of Hindi language are syllabic in nature, i.e, all the basic units can be represented by syllables by defining syllable as follows: typically a syllable is a

Table 4.3: Changes in duration of basic units at syntactic boundaries.

Syntactic boundary	Percentage change with respect to base duration			
	Rules of the basic system	New rules		
		Speaker 1	Speaker 2	Speaker 3
Phrase boundary	30	44	54	48
Sentence ending	35	49	61	68
Breath group	35	51	43	48
Syllable boundary	10	8	10	12

vowel with zero or more consonants preceding and/or following it. The vowel is called the nucleus of the syllable, the preceding consonant(s) if any, are called onset(s) of the syllable and the following consonant(s) if any, are called the coda(s). A syllable with coda is called a closed syllable, and that without a coda is called an open syllable.

If the last syllable before any boundary is a closed syllable, then it is observed that the effect of the boundary is felt on the onset and nucleus parts of the syllable. From the Table 4.4 it is clear that the average percentage change in the duration is 30%. It is much less compared to the lengthening in an open syllable, where the average percentage lengthening is found to be 52% (from Table 4.3). Hence, we can conclude that the effect of syntactic boundaries on the penultimate basic unit is less than that on the final basic unit. The results stated above coincides with that reported by Wightman *et al* in [75] from their study on segmental durations in the vicinity of prosodic phrase boundaries.

4.4 EFFECT OF CONTEXTUAL FACTORS ON DURATION

In this section we explore the effect of contextual factors on the duration of basic units. Here, the effect of the following basic units only is studied under this category. The reason for considering the effect of the following units is explained in Section 5.1. It is observed that the effect of voiced stops and aspirated stops on trills is different from

Table 4.4: Changes in duration of basic units at penultimate syntactic boundary position.

Penultimate Position to	Percentage change with respect to base duration			
	Rules of the basic system	New rules		
		Speaker 1	Speaker 2	Speaker 3
Phrase boundary	30	25	17	30
Sentence ending	35	30	25	45
Breath group	35	41	24	31

that of the other sounds. Hence we present the effect on trills separately. The effect of the nature of the following unit on its preceding unit other than trill is explained in the Section 4.4.2. We denote the units other than trills as *non-trill* category.

4.4.1 Effect on Trills

A trill is a speech sound where the articulator is held loosely near the roof of the mouth, so that the force of the air passing sets the articulator in vibration making contact with the roof of the mouth [76]. The alveolar trill, is a common type of trill where the tongue vibrates against the alveolar ridge. In general trills show large variation in any language [77]. In Hindi also, the trills are effected by the voiced stops and aspirated stops to different magnitudes. The voiced stops in the following position lengthen the duration of trills by 18% on average. On the other hand they affect the preceding non-trill units by 25%. Similarly the aspirated stops in the following position affect the duration of trills and non-trill units by 13% and 20%. The percentage changes are shown in Tables 4.5 and 4.6.

Table 4.5: Effect of voiced stops and aspirated stops on the duration of preceding trill.

Trills followed by	Percentage change with respect to base duration			
	Rules of the basic system	New rules		
		Speaker 1	Speaker 2	Speaker 3
Voiced stops	11	26	17	12
Aspirated stops	5	13	12	14

4.4.2 Effect on Non-Trill Category

We grouped all the other units which are not trills under the category of non-trill. Also the units in the following position are classified into six categories having different manners of articulation. The categories are voiced stops, aspirated stops, trills, fricatives, nasals, and semivowels. The semivowels */ya/* and */va/* are studied separately. It is observed that both the semivowels are affecting the duration of the preceding units in similar and nearly equal amounts. The unit */ha/* is studied under fricatives as the other fricatives were not affecting the duration of the preceding unit significantly. We can observe from the Table 4.7 that the effect of aspirated sounds on the preceding unit is large in continuous speech than that in utterances of constrained sequences. Except in the case of Speaker 2, the other two speakers show nearly equal amount of increment in the duration due to the effect of the following fricative unit */ha/*. Similarly Speaker 2 is exhibiting a significant variation in the case of following nasals.

4.5 SPEAKER INDEPENDENT VARIATIONS

Though variations are observed in the durations in continuous speech, there are some similarities among the speech units from all the speakers. Here, we say that the effect is same for all the speakers, if the difference among the percentage variations is less than 10%. Any factor resulting in percentage changes more than 10% is said to be

Table 4.6: Effect of following unit on the duration of preceding non-trill.

Non-trills followed by	Percentage change with respect to base duration			
	Rules of the basic system	New rules		
		Speaker 1	Speaker 2	Speaker 3
Voiced stops	15	30	19	26
Aspirated stops	8	18	21	20
Trills	30	14	21	27
Fricative */ha/*	25	23	5	18
Nasals	8	8	-10	4
Semivowel */ya/*	10	6	-8	-10
Semivowel */va/*	15	10	-10	-8

different. From Tables 4.1 to 4.4, we observe that all the units at word final positions and word initial CV and CCV_1 units receive same effect due to positional variations. Also, the units before a breath group and syllable boundary exhibit same durational characteristics. In Table 4.7 the average changes for these factors having similar durational properties are given. Some of the effects in the utterances of constrained sequences are repeated here in Table 4.7 for convenience.

4.6 VARIATIONS WITHIN A GENDER

In continuous speech, the factors affecting duration and their interaction cause large variation even in the speech of a single speaker. Here we examine some of the factors affecting the durations for two male speakers. Table 4.8 gives the factors having different effects on the speakers of the same gender. From the table we can observe that the units before a breath group exhibit large variation even for the speakers of the same gender. Also, the semivowels and nasals have affected the duration of the preceding units in two different directions opposing the property of *directional invariance* proposed by Jan van Santen [29].

Table 4.7: Average variations in the duration due to positional effects.

Position / Context of the unit	Percentage change with respect to base duration	
	Rules of the basic system	New rules
Word final:		
CV unit	30	35
CCV_1	15	13
CCV_2	30	42
Word initial:		
CV unit	10	21
CCV_1	5	12
Before a breath group	35	47
At syllable boundary	10	10
Trills followed by aspirated stops	5	13
Non-trill followed by aspirated stops	8	20

4.7 INCORPORATION OF DURATION KNOWLEDGE

The duration knowledge acquired from the analysis of durations of the basic units in continuous speech for Hindi is incorporated in the existing system using knowledge-based model. Production system approach is used to represent the duration knowledge in the TTS system. In this approach, knowledge is represented by a set of rules and a rule interpreter or the inference engine selects one or more number of rules for activation, depending on the position and context of the unit. When more than one rule fires, then they will be activated in a multiplicative manner. In this approach, addition of new rules to the knowledge base, or deleting some invalid rules, can be

Table 4.8: Effect of factors having large variations for the speakers of same gender.

Position / Context of the unit	Percentage duration change for	
	Speaker 1 (male)	Speaker 2 (male)
Word final CCV_2	28	17
Phrase boundary	44	54
Sentence ending	49	61
Penultimate position before breath group	41	24
Non-trill followed by:		
Voiced stop	30	19
Fricative */ha/*	23	5
Nasals	8	-10
Semivowel */ya/*	6	-8
Semivowel */va/*	10	-10

done very easily. The set of rules derived by analyzing the durations of basic units in continuous speech for Hindi is given in Appendix B. The block diagram in Figure 4.2 shows the steps involved in the incorporation of duration knowledge.

4.8 SUMMARY AND CONCLUSIONS

In this chapter we explored the effects of positional and contextual factors on the duration of basic units in continuous speech for Hindi. It was observed that the effect of semivowels and nasals in the following positions violate *directional invariance* property suggested by van Santen. Finally, the variations in the effects of some of the factors on the speakers of the same gender were explained. It is observed that the incorporation of duration knowledge acquired from the continuous speech has improved the intelligibility of the output synthetic speech.

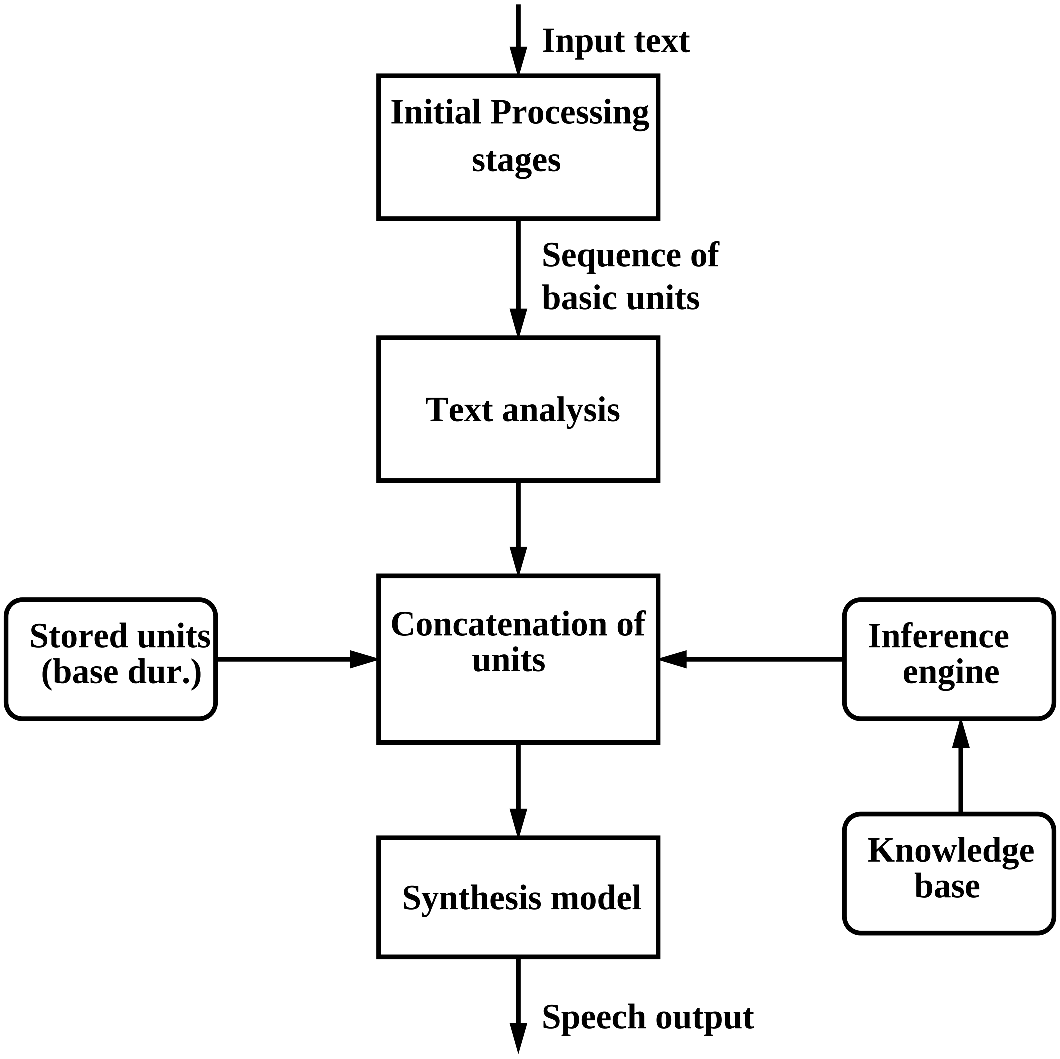

Fig. 4.2: Block diagram depicting the incorporation of duration knowledge.

CHAPTER 5

DURATIONAL ANALYSIS FOR TELUGU

5.1 INTRODUCTION

In this chapter, we describe the effect of various factors on the durations of basic units for *Telugu* language. We derive the durational information by analyzing the durations of the basic units extracted from continuous speech corresponding to *DoorDarshan* news.

5.2 PREPARATION OF DATABASE

Database is prepared using the speech data from television news bulletins. The database consists of three news bulletins each of about 15 minutes duration. Two of these three bulletins were read by male speakers and the other by a female speaker. In each bulletin, there are about 250-300 sentences each of 2-3 sec duration. Speech signal is sampled at 16 kHz sampling frequency and stored as 16 bit integers. Speech is transcribed into its corresponding text and stored in a text file using the ITRANS code [78]. As syllable-like units are used as basic units, speech signal is manually labeled to syllable-like units. The duration of each of the labeled syllable-like units is noted separately in the form of table. A typical structure of the table containing duration information of various syllable-like units is shown in the Table 5.1. The values in the first column are the durations (in ms) of the corresponding basic units given in the second column. In the database there are a total number of 14,349 syllable-like basic units. Out of them 4,458 are uttered by the female speaker and the remaining 9,891 are uttered by the two male speakers.

Table 5.1: Typical structure of the duration information table.

Duration (ms)	Basic unit
21	SIL
83	cho
96	ra
128	bA
53	Tu
140	dA
68	ru
52	la
150	up
131	sam
67	ha
146	ra
116	Na
36	SIL

5.3 MEASUREMENT OF DURATIONS

5.3.1 Base Duration

In order to represent the changes occurred in the durations due to any of the factors, we need a reference duration which we call *base duration*. This base duration should not include the effect of any of the factors. Hence, we measured base duration of syllable-like units when they are in certain context specified by the following conditions, where the effect of any of the factors is minimum. If the unit is not a vowel, then it should be followed by unvoiced unaspirated stop, should be present in the word medial position, and should not be present in a cluster. If the unit is vowel, then it should be in the word initial position, and should be followed by unaspirated unvoiced stop. For each of the basic units, we calculate base duration. The values in the fourth column of the Table 5.2 give the base durations of some of the basic units.

5.3.2 Duration Affected by the Positional Factors

From the manually transcribed text, we locate various positions, mentioned under the positional factors in the Section 5.1, at which durations of units get affected. We find that the effect of positional factors on the basic units for different manners of articulation and voicing is not the same. Hence, we grouped the basic units into six classes: voiced, unvoiced, liquids, nasals, trills and semivowels. For each class of basic units, duration is measured when they are affected by any one of the positional factors. Basic units of all the above classes were showing a similar trend when they are in the word initial position. Hence, the basic units in the word final position only are considered separately. Also, when a basic unit is in the middle of a word the change in duration is accounted by contextual factors. The change in the duration value for each basic unit is expressed in percentage with respect to the base duration of the basic unit.

5.3.3 Duration Affected by the Contextual Factors

We have considered the effects of different manners of articulation and voicing of the following unit on the duration of present unit as contextual effects in Section 5.1. We have studied the effects of nasals, liquids, trills, semivowels, voiced and aspirated basic units in the following position.

5.4 ANALYSIS OF THE EFFECTS OF POSITIONAL FACTORS

In this section we will consider the effects of positional factors on the duration of basic units in continuous speech for Telugu. In the next few sections we present the analysis of durations of basic units which come under the category of prepausal lengthening.

Table 5.2: Illustration of the effect of word final position on obstruent category of units.

Speaker 1 (male)				
Basic Unit	**Word**	**Duration of Basic unit**	**Base Duration**	**% Change**
Word final voiced unit				
/Du/	/naTuDu/	78	67	16
/gu/	/nAlugu/	95	81	18
/dhi/	/pratinidhi/	133	116	15
Word final unvoiced unit				
/ka/	/sainika/	107	86	25
/Tu/	/pATu/	98	77	27
/ka/	/vyatirEka/	124	96	29
/ki/	/vAriki/	98	83	18

5.4.1 Word Final Voiced Vs. Unvoiced Units

In the word final position, like in many other languages, we observe lengthening of basic units in duration. The lengthening is less for voiced stop consonant-vowel (CV) as compared to that for the unvoiced stop consonant-vowel units. We have considered three speakers for this study. Tables 5.2 and 5.3 give illustrations of some of the basic units studied under this category.

Table 5.3: Illustration of the effect of word final position on obstruent category of units (Contd.).

Speaker 2 (male)				
Basic Unit	**Word**	**Duration of Basic unit**	**Base Duration**	**% Change**
Word final voiced unit				
/Du/	/mUDu/	75	60	25
/da/	/dOhada/	114	94	21
/da/	/varada/	117	94	24
/Du/	/nAyuDu/	78	60	30
Word final unvoiced unit				
/ka/	/naitika/	114	90	27
/ti/	/parimiti/	104	81	29
/ka/	/sainika/	118	90	31
Speaker 3 (female)				
Word final voiced unit				
/gA/	/badulugA/	128	112	14
/du/	/aidu/	105	91	15
/Du/	/adyakshuDu/	103	88	17
Word final unvoiced unit				
/ki/	/kshEmAniki/	120	96	25
/ku/	/mEraku/	112	90	24
/ka/	/chErika/	101	80	27

Table 5.4: Illustration of the effect of pause on CCV clusters.

Basic Unit	**Word**	**Duration of Basic unit**	**Base Duration**	**% Change**
Speaker 3 (female)				
Word initial CCV				
/p/	/prabhutva/	61	51	20
/d/	/dyEyamgA/	74	61	21
/s/	/svarNAlu/	62	61	21
Word final CCV				
/g/	/sariggA/	57	41	40
/r/	/pArTI/	44	30	46
/t/	/pottu/	38	29	32
/n/	/chEstunnA/	67	48	39
/r/	/kAryadarsi/	42	30	41

5.4.2 Word Final Liquids, Nasals, Trills and Semivowels

The variation in the durations of units which belong to the sonarant category is different from the variation in the obstruents category of units which we mentioned in the previous section. The four types of sonarant category units: liquids, nasals, trills, and semivowels show different variations in the word final position. For each one of the four types, we analyzed a number of CV units for the three speakers and some of them are given the Tables 5.5 to 5.7 as an illustration.

5.4.3 Effect of Major Syntactic Boundaries

The boundaries at the end of a phrase, before a breath group and at the sentence ending position can be considered as major syntactic boundaries, as there will be a

Table 5.5: Illustration of the effect of word final position on sonarant category of units for speaker 1 (male).

Speaker 1 (male)				
Basic Unit	**Word**	**Duration of Basic unit**	**Base Duration**	**% Change**
Word final liquids				
/la/	/digumatula/	103	70	47
/li/	/ivatali/	89	60	48
/lu/	/dArulu/	97	70	38
Word final nasals				
/ma/	/tama/	99	76	30
/na/	/chEsina/	91	68	33
Word final trills				
/rA/	/dvArA/	126	101	24
/ra/	/bhIkara/	79	65	20
/ra/	/itara/	83	65	26
Word final semivowel				
/yi/	/jarugutAyi/	74	60	23
/vu/	/vunDavu/	66	56	18
/vai/	/Ayuvai/	112	96	17
/yi/	/nilichAyi/	74	60	24
/vi/	/lAnTivi/	85	70	20

Table 5.6: Illustration of the effect of word final position on sonarant category of units for speaker 2 (male).

Speaker 2 (male)				
Basic Unit	**Word**	**Duration of Basic unit**	**Base Duration**	**% Change**
Word final liquids				
/lA/	/idilA/	152	115	32
/li/	/Avali/	127	98	29
/lA/	/vElA/	169	115	47
Word final nasals				
/ni/	/chEstAmani/	115	91	26
/ma/	/tama/	124	98	27
/na/	/jarigina/	103	85	21
Word final trills				
/ra/	/amara/	87	74	18
/ra/	/prasAra/	86	74	17
/ri/	/tivAri/	109	88	23
Word final semivowel				
/ya/	/jAtIya/	85	64	33
/yi/	/vAjpEyi/	88	69	28
/ya/	/vyaya/	82	64	28
/vi/	/avi/	100	76	31

Table 5.7: Illustration of the effect of word final position on sonarant category of units for speaker 3 (female).

Speaker 3 (female)				
Basic Unit	**Word**	**Duration of Basic unit**	**Base Duration**	**% Change**
Word final liquids				
/lu/	/dAkalu/	113	91	24
/li/	/Akali/	125	98	27
/lu/	/mukhyAmshAlu/	112	91	23
Word final nasals				
/mu/	/tAmu/	145	104	39
/ni/	/chEyAlani/	121	89	36
/na/	/chEsina/	139	98	42
Word final trills				
/ru/	/chEsAru/	96	81	19
/ra/	/bAlura/	79	70	13
Word final semivowels				
/ya/	/bhAratIya/	144	116	24
/yA/	/AsiyA/	188	147	28
/yA/	/inDiyA/	187	147	21
/va/	/mUDava/	105	86	22

resetting of F_0 contour at these instants. Of all the variations in durations of basic units due to the effect of prepausal lengthening, the variations at these major syntactic boundaries are significant. The units at a phrase boundary are incremented more in duration compared to the units at other two major syntactic boundaries. For all the three speakers, this was observed. Tables 5.8 to 5.10 show some examples of the units considered for the study of the three major syntactic boundaries.

Table 5.8: Illustration of the effect of major syntactic boundaries on the duration of basic unit in the speech of speaker 1 (male).

Speaker 1 (male)				
Basic Unit	**Word**	**Duration of Basic unit**	**Base Duration**	**% Change**
Phrase boundary				
/ni/	/chEstAmani/	152	86	76
/lO/	/amalulO/	178	102	74
/pai/	/vAripai/	221	126	75
/ru/	/okaru/	122	71	72
/DA/	/kUDA/	142	81	76
Breath groups				
/ki/	/chEyaDAniki/	159	105	51
/tE/	/ayitE/	191	123	55
/ru/	/kOrAru/	106	71	49
/ni/	/kAdani/	136	86	59
Sentence ending position				
/di/	/chEsinadi/	96	75	28
/ru/	/kOrAru/	92	71	29
/ru/	/cheppAru/	89	71	25

5.4.4 Effect of Word Initial Position

Similar to prepausal lengthening effect we have observed an increment in the duration of the basic units after a pause i.e, either word initial position or the sentence initial position. Unlike the prepausal lengthening effect, the postpausal lengthening effect show nearly equal amount of lengthening after all types of pauses. Table 5.11 list the units studied under this category.

5.4.5 Effect of Word Boundary Positions on the Duration of Consonants in a CCV Cluster

For studying the effect of pause on the duration of the consonant part of a CCV cluster, we computed the duration of the consonant part of CCV clusters in word initial and final positions. Out of three speakers considered, both the male speakers tend to increase the duration of the consonant by nearly equal amount in both word initial and word final positions. The female speaker showed a larger increment in a word final cluster compared to word initial clusters. Tables 5.4 and 5.12 show the effect on the duration of consonants in a CCV cluster environment.

5.5 ANALYSIS OF THE EFFECT OF CONTEXTUAL FACTORS

As stated in the Section 5.1, factors like nature of the preceding and the following units also affect the duration of the current unit. Different manners of articulation and places of articulation of the basic units in the preceding and following position affect the duration of the present unit to different extents. The effect of preceding sound unit can be modelled through a set of coarticulation rules [79]. Here, we study the effect of the following unit in this section.

5.5.1 Effect of Sonarants in the Following Position

Among the sonarants liquids, nasals, trills and semivowels show a major effect on the duration of the preceding unit. All the four types exhibit different magnitudes of changes. Also, the direction of the change is varying from speaker to speaker. Liquids in the following position has made the duration of the current unit to decrease by nearly equal amount for the male speakers. But the presence of a liquid sound has caused an increment in the duration of the current unit by a large amount in the case of a female speaker. The presence of nasals and trills in the following position has triggered an increment in the duration of the current unit in the speech of all the three speakers. Similarly, the occurrence of a semivowel in the following position forced the current unit to reduce its duration. Tables 5.13 to 5.15 show some example of the units considered under this category.

5.5.2 Effect of Unaspirated Voiced Stop/Aspirated Stop Units in the Following Position

In the obstruents category unaspirated voiced stops and aspirated stop consonants have exhibited considerable effect on the duration of the current unit due to their presence in the following position. All the three speakers tend to increase the duration of the basic unit if the following unit is either unaspirated voiced stop or aspirated stop. Tables 5.16 to 5.19 show some of the units considered to study the effect of the following obstruent on the current basic unit duration.

5.6 COMPARISONS WITH RESPECT TO SPEAKER, GENDER AND LANGUAGE

The effect of various factors on the duration is explained in the previous sections. There exist some similarities and deviations in the effects of positional and contextual factors among the three speakers. Some types of sound units exhibited similar effects among different speakers. The factors studied were found to be showing similar effects on the speakers of both the genders, and also on the speakers of the two Indian languages considered. In this section, a summary of the similarities and the dissimilarities in the effects of durational factors among the speakers, genders, and languages are explained. Tables 5.17 and 5.18 give the summary of the effects of positional and contextual factors for all the three speakers. The values under the column Syllable Frequency (SF) indicates the number of basic units observed for the effect studied. The values under the column Percentage Change (PC) give the average percentage change over the base duration.

Table 5.9: Illustration of the effect of major syntactic boundaries on the duration of basic unit in the speech of speaker 2 (male).

Speaker 2 (male)				
Basic Unit	**Word**	**Duration of Basic unit**	**Base Duration**	**% Change**
Phrase boundary				
/ni/	/lEvani/	141	91	56
/ti/	/udhruti/	123	81	58
/ku/	/varaku/	137	84	63
/DA/	/kUDA/	170	112	52
/nu/	/hAmIlanu/	146	93	57
/Du/	/sOdaruDu/	94	61	58
Breath groups				
/tU/	/mATlADutU/	153	108	42
/kE/	/ippaTikE/	190	129	48
/ni/	/tIsukunnAmani/	133	91	46
/ki/	/cheyyaTAniki/	110	76	46
Sentence ending				
/di/	/chEsinadi/	130	106	22
/ru/	/annAru/	108	89	21
/ru/	/cheppAru/	112	89	26
/na/	/prakaTana/	108	85	27

Table 5.10: Illustration of the effect of major syntactic boundaries on the duration of basic unit in the speech of speaker 3 (female).

Speaker 3 (female)				
Basic Unit	**Word**	**Duration of Basic unit**	**Base Duration**	**% Change**
Phrase boundary				
/ni/	/rAvAlani/	138	89	55
/tU/	mATlADutU/	305	193	58
/gA/	/unDagA/	173	112	54
/ni/	/undani/	135	89	52
Breath groups				
/ju/	/rOju/	188	129	46
/ki/	/chEyaTAniki/	139	96	45
/nE/	/taruvAtanE/	189	133	42
/ni/	/tappadani/	124	89	39
/ka/	/idikAka/	110	80	38
/lu/	/stAnAlu/	129	91	42
Sentence ending				
/ru/	/cheppAru/	113	81	40
/yi/	/rAsAyi/	126	89	41
/yi/	/chEsAyi/	128	89	44
/ru/	/annAru/	110	81	36

Table 5.11: Illustration of the effect of word initial position.

Basic Unit	Word	Duration of Basic unit	Base Duration	% Change
Speaker 1 (male)				
/ka/	/kanIsam/	112	86	30
/dA/	/dADulu/	178	132	35
/cho/	/chorabATuku/	122	88	36
/bo/	/boggu/	177	125	37
Speaker 2 (male)				
/bo/	/boggu/	164	120	36
/A/	/Ayana/	123	90	37
/i/	/idilA/	112	81	38
Speaker 3 (female)				
/vE/	/vEdikanu/	129	90	43
/bA/	/bAlikala/	185	135	37
/vi/	/vijayam/	134	99	35
/rA/	/rAjakIya/	193	140	38
/chE/	/chEsAru/	163	117	39

Table 5.12: Illustration of the effect of pause on CCV clusters.

Basic Unit	Word	Duration of Basic unit	Base Duration	% Change
Speaker 1 (male)				
Word initial CCV				
/p/	/prajalu/	64	58	10
/v/	/vyavasAya/	68	62	10
/s/	/svAgatam/	52	47	10
Word final CCV				
/k/	/vimukti/	57	50	14
/r/	/chorabATudArlu/	24	22	9
/n/	/kOrindi/	42	36	16
/s/	/bhAskar/	52	47	11
Speaker 2 (male)				
Word initial CCV				
/p/	/pratyEka/	61	56	10
/t/	/tvaralO/	32	29	9
/k/	/kramangA/	75	68	10
/p/	/prayOjanAlaku/	62	56	11
Word final CCV				
/t/	/mantri/	34	29	17
/n/	/vishayAnni/	52	48	8
/r/	/pArTi/	40	36	11
/dh/	/madhya/	57	52	10
/r/	/tIrpu/	40	36	12

Table 5.13: Illustration of the effect of sonarants in the following position for speaker 1 (male).

Basic Unit	**Word**	**Duration of Basic unit**	**Base Duration**	**% Change**
Following unit is a liquid				
/ta/	/ivatali/	96	103	-7
/ka/	/avakatavakala/	81	86	-6
/ja/	/prajalaku/	85	90	-5
Following unit is a nasal				
/si/	/chEsina/	119	99	20
/Ta/	/samghaTanaku/	106	87	22
/pi/	/kalipinadi/	95	80	19
Following unit is a trill				
/pA/	/apAramaina/	130	113	15
/vA/	/nivArana/	141	126	12
/sA/	/chEsAru/	122	107	14
Following unit is a semivowel				
/ta/	/avakatavakala/	73	84	-13
/Tu/	/eTuvanTi/	64	77	-18
/ni/	/jenivA/	69	86	-19
/kO/	/tIsukOvAlani/	89	107	-17
/bhi/	/abhiyOgAlapai/	106	121	-12

Table 5.14: Illustration of the effect of sonarants in the following position for speaker 2 (male).

Basic Unit	**Word**	**Duration of Basic unit**	**Base Duration**	**% Change**
Following unit is a liquid				
/di/	/idilA/	101	106	-4
/vA/	/havAlA/	104	112	-7
/ja/	/nadIjalAlu/	91	97	-6
/di/	/vadili/	104	106	-2
Following unit is a nasal				
/ja/	/prayOjanAlaku/	106	92	16
/la/	/pAlana/	98	83	18
/Ta/	/prakaTana/	115	102	13
/ya/	Ayana/	77	64	21
/sA/	/chEsAmani/	185	149	24
Following unit is a trill				
/ka/	/spasTIkaraNa/	104	90	16
/sA/	/chEsArani/	165	149	11
/ka/	/ekarAla/	103	90	15
/va/	/vivarAlu/	83	68	22
Following unit is a semivowel				
/tI/	/jAtIya/	137	154	-10
/ra/	/aravai/	70	74	-6
/sA/	/chEsAyi/	131	149	-12
/pA/	/rUpAyalu/	135	157	-14

Table 5.15: Illustration of the effect of sonarants in the following position for speaker 3 (female).

Basic Unit	Word	Duration of Basic unit	Base Duration	% Change
Following unit is liquid				
/do/	/vaidolagina/	182	147	24
/rA/	/stAvarAlanu/	171	140	22
/Du/	/dADulanu/	103	88	17
/Da/	/vijayavADalO/	114	94	21
Following unit is nasal				
/va/	/lEvani/	109	86	27
/lE/	/maruvalEni/	126	98	29
/la/	/stAvarAlanu/	103	79	30
/la/	/nirmUlanaku/	101	79	28
/rA/	/samvatsarAniki/	135	110	23
Following unit is trill				
/la/	/chEyagalarani/	92	79	17
/ka/	/vargIkaraNa/	91	80	14
/da/	/kurani/	117	98	19
/ta/	/itara/	106	89	19
Following unit is semivowel				
/ja/	/vijayavADa/	77	87	-12
/pA/	/rUpAyalu/	113	121	-8
/tA/	/penchutAyani/	122	135	-10

Table 5.16: Illustration of the effect of following obstruent unit.

Basic Unit	Word	Duration of Basic unit	Base Duration	% Change
Speaker 1 (male)				
Following unit is a voiced stop				
/ka/	/nilakaDagA/	101	86	17
/Tu/	/naTuDu/	93	77	21
/Tu/	/bATugA/	91	77	18
Following unit is an aspirated stop				
/ja/	/rAjadhAni/	91	81	12
/lE/	/vilEkhari/	109	98	11
/li/	/alikhita/	94	70	13
Speaker 2 (male)				
Following unit is an voiced				
/la/	/velupalagala/	87	70	24
/bA/	/haidarAbAdulO/	177	145	22
/ta/	/jataga/	110	90	23
Following unit is an aspirated				
/ra/	/sarapharA/	82	74	10
/lE/	/vilEkharula/	129	119	9
/sha/	/vishAkha/	92	83	11

Table 5.17: Durational changes due to positional factors. The abbreviations PC and SF represents percentage change and syllable frequency respectively.

Position of the unit	Speaker 1		Speaker 2		Speaker 3	
	PC	SF	PC	SF	PC	SF
Word final position						
(i) Voiced	16	13	26	18	15	34
(ii) Unvoiced	26	31	29	26	25	26
(iii) Liquid	44	19	32	24	25	24
(iv) Nasal	32	22	26	19	40	28
(v) Trill	24	28	19	27	17	20
(vi) Semivowel	20	24	30	12	24	25
Phrase boundary	76	20	58	19	55	31
Breath group position	54	23	46	20	42	25
Word initial position	36	34	37	33	39	35
Sentence ending position	27	24	25	32	40	34
Word initial CCV	10	17	10	18	21	20
Word final CCV	13	16	10	12	40	28

Table 5.18: Durational changes due to contextual factors. The abbreviations PC and SF represents percentage change and syllable frequency respectively.

Context of the unit	Speaker 1		Speaker 2		Speaker 3	
	PC	SF	PC	SF	PC	SF
Unit followed by						
(i) Liquid	-5	25	-4	22	21	29
(ii) Nasal	20	31	17	20	27	21
(iii) Trill	14	19	16	18	17	24
(iv) Semivowel	-16	24	-10	21	-10	19
(v) Voiced stop	19	18	22	16	13	28
(vi) Aspirated stop	11	14	10	14	17	25

5.6.1 Similarities across the Speakers

A large number of factors affect the duration of the basic units as mentioned in section 5.1. Generally any speaker cannot repeat a word or basic sound unit with same duration. Also, there is large variation in the duration of the basic units occurring at different contexts uttered by the same speaker. For example, the duration of the basic unit */ka/* in the case of speaker 1, has varied from 52 ms to 160 ms in different contexts. Even though there is a large variation in the absolute durations, from the Tables 5.17 and 5.18 we observe that the percentage change in many of the cases is approximately equal. As an illustration, all the three speakers have shown nearly equal amount of change for sound units in the word initial position. The similarity in lengthening of basic units at word initial position may vary for speech in different contexts like casual conversational speech, political speech, or speech while addressing a group of delegates. But for the given case of news reading style of speech, the variation is approximately

Table 5.19: Illustration of the effect of following obstruent unit.

Basic Unit	Word	Duration of Basic unit	Base Duration	% Change
Speaker 3 (female)				
Following unit is voiced				
/la/	/vaidolagina/	89	79	13
/ya/	/chEyaDAniki/	131	116	13
/ru/	/porugu/	91	81	13
Following unit is aspirated				
/mu/	/pramukha/	122	104	17
/lE/	/vilEkharulatO/	114	98	16
/gA/	/pragADha/	129	112	15
/ja/	/rAjadhAni/	104	87	19

equal irrespective of the sound unit, speaker and gender. Hence it can be characterized as a language specific-feature. Unvoiced units in the word final position have undergone similar variations for all the speakers. Sounds under trills category, have lengthened by nearly the same amount in the word final position for all the speakers. Also, the trills category of units have affected the preceding units to the same extent.

5.6.2 Variations Between the Genders

Apart from some similarities as discussed above, there are some deviations between genders. For example, female speaker exhibited a large change in the sentence ending position compared to the change exhibited by male speakers. Also, nasal sounds at the word final position have got more lengthening in the speech of the female speaker, compared with the speech of the male speakers speech. The presence of liquid sounds lengthened the preceding unit in the female speech, whereas it has reduced in the case of male speaker. There is a significant difference in the duration of the preceding unit

Table 5.20: Average duration changes among the three speakers.

Position/Contextual factor	Average % change
Word initial position	34.67
Word final Unvoiced	26
Word final trills	23.67
Followed by trills	15.67

Table 5.21: Variations in the duration changes between the two genders.

Position/ Context	Average percentage changes	
	Male speaker	Female speaker
Sentence ending position	26	40
Word final (nasals)	29	40
Liquids in following position	-4.5	21
Nasals in following position	18.5	27
Aspirated stops in following position	10.5	17

due to the effect of nasals and aspirated stops in the following position. The above mentioned differences in the durational behavior can be used to differentiate the gender of a given speech wave automatically. Table 5.21 gives summary of the average effects in the speech of male and female speakers.

5.6.3 Variations within a Gender

Along with similarities, some deviations are also observed among the speakers of the same gender. In our analysis from the tables it is clear that the voiced units, liquids and semivowels in the word final position has differed much for the two male speakers. Similarly, the effect of phrase boundary in the speech of speaker 1 is 78% whereas for

Table 5.22: Variations in duration effects for speakers of same gender.

Position/context factor	Percentage change	
	Speaker 1	Speaker 2
Word final voiced units	16	26
Word final liquids	44	32
Word final semivowel	20	30
Following unit is semivowel	-16	-10

speaker 2 it is 58%. Semivowels in the following position has effected the duration of the preceding units to different extents for the two male speakers. Table 5.22 summarizes the above stated deviations in the durational effects.

5.6.4 Comparison of Durational Effects across Hindi and Telugu Languages

There is very little quantitative data available on the changes in segmental durations of speech with regard to Indian languages. It is often assumed that many Indian languages have common phonetic and/or phonological features [80]. Also, some differences in segments have been shown to have cross linguistic validity [81]. Here the similarities and differences in the durational patterns of the speech in Telugu and Hindi are studied.

The lengthening in duration of units in penultimate final positions of a word is observed in Hindi. This effect is observed when the final character is not having vowel in it. It means that if the final unit is a coda, then the CV unit in the final syllable excluding the coda gets lengthened. These effects also hold good for units at the penultimate phrase boundary positions, breath group positions, and sentence ending positions. But in the case of Telugu, in which most of the syllables are in the final position, coda does not exist. Hence the effect is always on the final unit of the word.

Voiced and unvoiced units in word final position show nearly equal amount of lengthening in the case of Hindi, whereas they exhibit different amounts of lengthening in the case of Telugu. In Telugu, unvoiced sounds at the word final positions receive more lengthening than voiced units. Similarly lengthening at syllable boundary was not observed in the case of Telugu. But syllable boundaries show considerable lengthening in the speech of Hindi language. The effect of voiced stops in the following position on the preceding unit is almost equal in both languages. On the other hand, the effect of aspirated units in the following position is more for speech in Hindi than in Telugu. The effect of trills and semivowels in the following positions is nearly equal for speech in both the languages. The fricative sound */ha/* in the following position has considerable effect on its preceding unit in Hindi speech.

5.7 SUMMARY

In this chapter, a detailed study on the durations of the basic units in continuous speech of Telugu language was carried out. From the analysis it is clear that different sound units receives the effect of the same factor to different extents. For example, nasals and trills in the word final position are modified to different extents for the same speaker. Also, there are some cross linguistic similarities and differences in the durational patterns of the two languages.

The acquired duration knowledge by the analysis of durations of the basic units in the continuous speech of Telugu is incorporated in the existing system with the same set of basic units that are used for the Hindi TTS system. Knowledge-based approach which is explained in Section 4.7 is used to incorporate the duration knowledge into the TTS system. Listening tests show that the quality of the output synthetic speech from the TTS system with the duration rules derived from continuous speech has improved significantly. The duration rules for Telugu are given in Appendix C.

CHAPTER 6

INTONATION KNOWLEDGE FOR TELUGU

6.1 INTRODUCTION

Intonation pattern is defined as the variation of fundamental frequency, the acoustic parameter of pitch over time. Properties of intonation patterns change with respect to the type of sentence. F_0 contour of a declarative sentence declines gradually with time, whereas for interrogative sentences F_0 contour rises towards the end of the sentence. The characteristics of these patterns are speaker dependent. Though F_0 contours are speaker dependent, there are some global properties of these contours which are language specific. Apart from the language dependent characteristics, some common intonation patterns were observed across languages in the previous studies [82]. For example, the declining tendency in the case of declarative sentences is a common feature studied for many languages. The local fall-rise patterns that are included in the F_0 contour can change from language to language, as they are determined by the phonological patterns of the constituent words. The objective of this chapter is to study the intonation patterns of the speech for Telugu.

This chapter is organized as follows: Section 6.2 describes the technique used to compute the pitch contours. The features of the fall-rise patterns in the pitch contours for the utterances of Telugu language are studied in the Section 6.3.1. It also includes the comparison of the pitch contours of Telugu and Hindi languages. A study on the resetting of the F_0 contour at syntactic boundaries is made in Section 6.3.2. The last Section summarizes the properties of intonation patterns Hindi and Telugu .

6.2 SIMPLIFIED INVERSE FILTER TRACKING (SIFT) ALGORITHM FOR PITCH EXTRACTION

There exist several algorithms to determine the pitch period of a given speech signal [83–85]. They are based on autocorrelation functions, cepstrum, simplified inverse filter tracking, linear predictive coding, average magnitude difference functions, group delay functions, etc. We have used the pitch extraction method based on simplified inverse filter tracking (SIFT), as it is computationally not very expensive compared to other methods of extraction. Also, for the study of pitch contours of clean speech like news reading, SIFT algorithm gives accurate F_0 values.

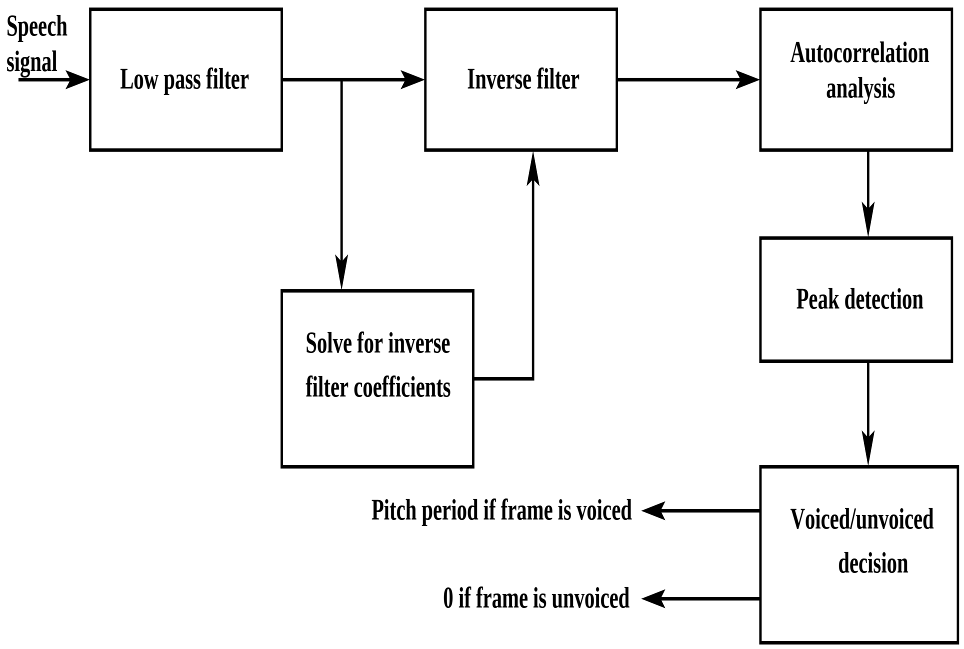

Fig. 6.1: Block diagram for pitch estimation using SIFT algorithm.

The speech signal is passed through a low pass filter with cut off frequency at 4 KHz. The low pass filtered signal is then sampled at 10 KHz. Linear prediction analysis is done on the sampled signal with a frame size of 256 samples and a frame shift of 64 samples between the successive frames. Linear prediction analysis is performed on each frame of the speech signal to obtain a set of p equations in p unknowns. The inverse filter coefficients $\{a_k\}$ can be obtained by solving the set of linear equations given by

$$\sum_{k=1}^{p} r_n(|i-k|)a_k = r_n(i), \quad 1 \leq i \leq p.$$

Where $r_n(i)$ is the set of autocorrelation coefficients for the n^{th} frame. Durbin's recursive algorithm is used for solving the set of equations. Here p denotes the order of the linear prediction analysis. Once the inverse filter coefficients are computed, the transfer function of the inverse filter is given by

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$$

Speech signal is passed through the inverse filter to get the residual signal. F_0 is estimated from the autocorrelation of the residual signal. The largest peak corresponds to the pitch period of the speech signal. Simple interpolation is applied in the region of the peak, and the decision of voiced/unvoiced is taken based on the interpolated peak. If the frame is voiced, then the inverse of the pitch period gives the F_0 for that frame.

We have considered the news data for studying the intonation properties of Telugu. The properties of declarative sentences are compared with those for Hindi studied in [86].

6.3 INTONATION PATTERNS FOR TELUGU

We study the similarities and differences in the characteristics of intonation patterns of Hindi and Telugu for a read style of speech. We have observed significant differences and quite some similarities in the global characteristics of the intonation patterns between the two languages. We consider the characteristics of the local fall-rise patterns and the resetting of the F_0 contour in the following two subsections.

6.3.1 Local Fall-Rise Patterns

Peaks and valleys of F_0 contour for an utterance form local fall-rise patterns. The fundamental frequency contour is accompanied by the local fall-rise patterns along with some global characteristics depending on the type of sentence. These local fall-rise patterns were observed for words with varying number of syllables from the database of the three news bulletins in Telugu language. There are many trisyllabic and tetra-syllabic words compared monosyllabic and disyllabic words. Most of the monosyllabic words are functional words, and thus combine with a disyllabic or trisyllabic content words either at the beginning or at the end. Hence the number of trisyllabic and tetra-syllabic words are higher in number. In disyllabic words a valley on the first syllable and a peak on the last syllable are generally observed as shown in the Figure 6.2. In the case of trisyllabic words either a valley-peak-valley or a valley-peak sequence is observed. If two valleys are present then the peak is located on the middle syllable, otherwise on the last syllable of the word. Figures 6.3 and 6.4 show these two types of patterns for the trisyllabic words */AsiyA/* and */sainika/*.

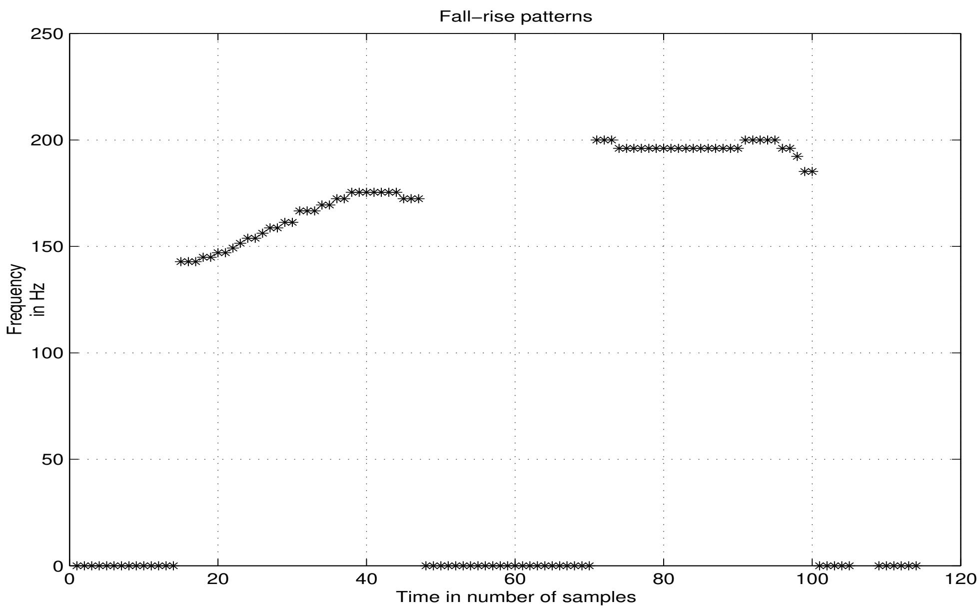

Fig. 6.2: Pattern of the F_0 contour for the disyllabic word */inkA/*.

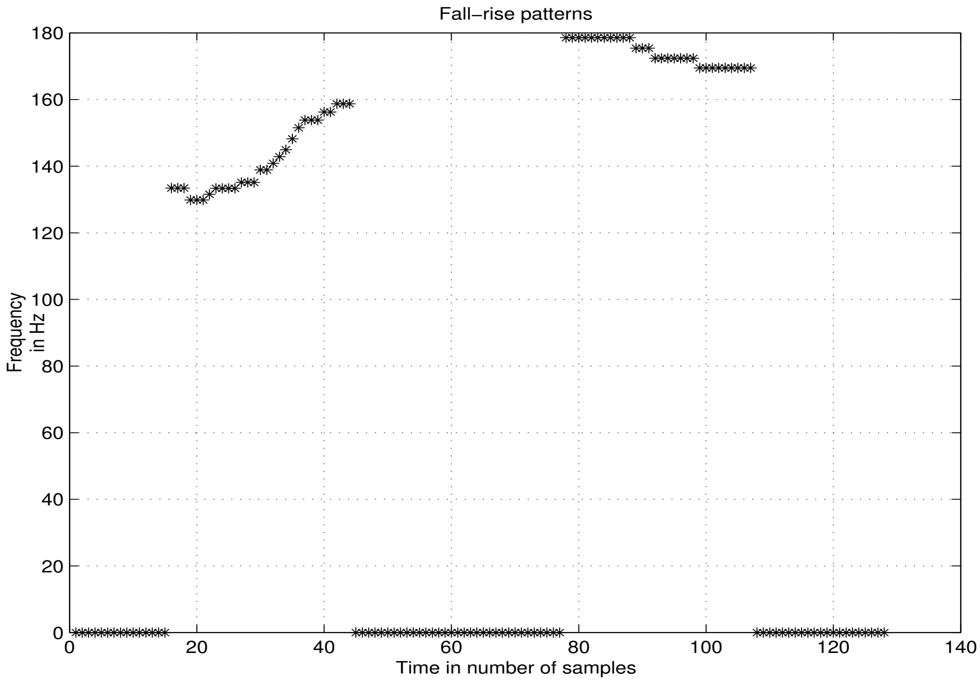

Fig. 6.3: Pattern of the F_0 contour for the trisyllabic word */AsiyA/*.

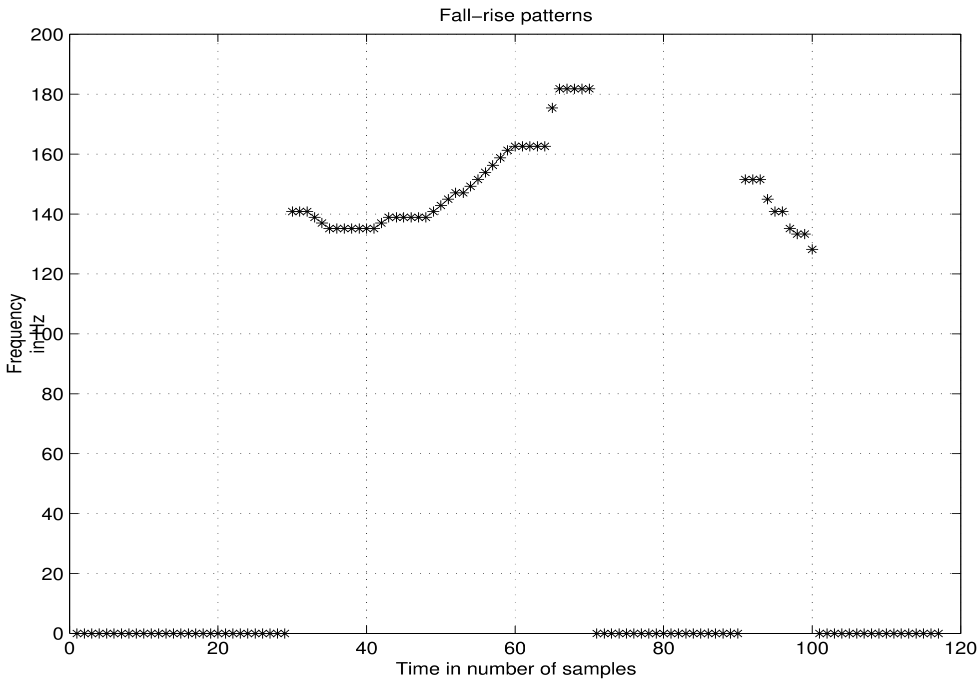

Fig. 6.4: Pattern of the F_0 contour for the trisyllabic word */sainika/*.

Tetrasyllabic words show a consistent pattern of a valley on the first syllable and a peak on either second or third syllable followed by again a valley on the final syllable as shown in the Figure 6.5.

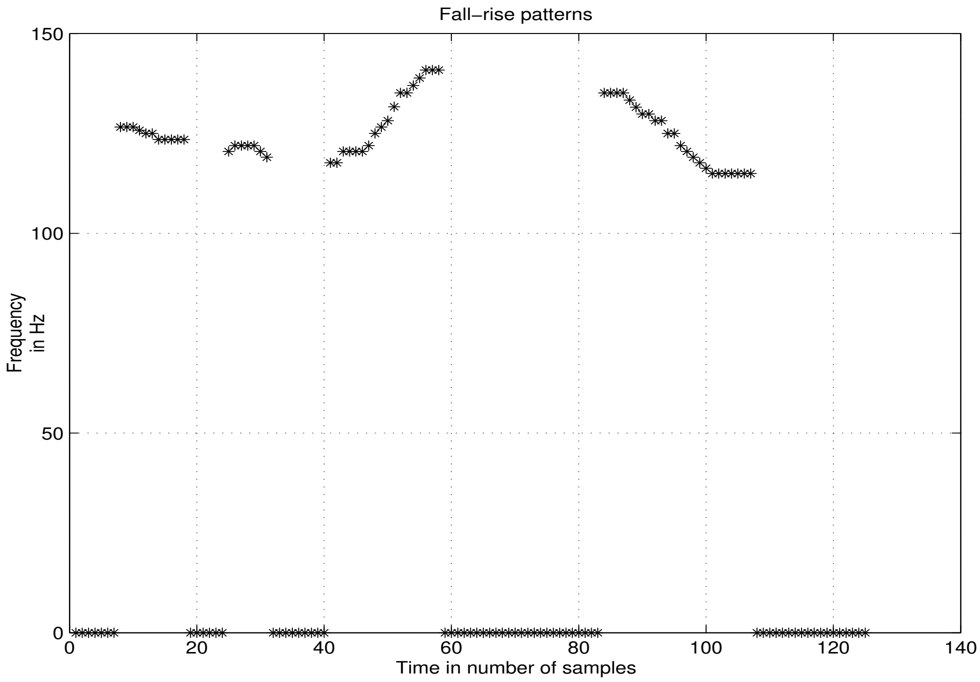

Fig. 6.5: Pattern of the F_0 contour for the tetrasyllabic word */taruvAta/*.

6.3.2 Resetting of F_0 Contour

Due to the constraint of the physical size of the lungs, the air pressure drops after certain time period. This period is called one breath group. At the end of the a breath group the fundamental frequency also drops to its minimum. Again after a pause the air pressure builds up at the beginning of the next breath group. This is called resetting of F_0 contour.

In Hindi it has been shown that a reset in fundamental frequency takes place at the beginning of a breath group or a phrase. But in our study on resetting of F_0 contour in Telugu, we observed that the resetting doesn't happen immediately but there will be a rise in the fundamental frequency of the last syllable before the preceding breath group or a fall in the fundamental frequency of the first syllable after the breath group. The two types of resetting are observed and are shown in the Figures 6.6 and 6.7.

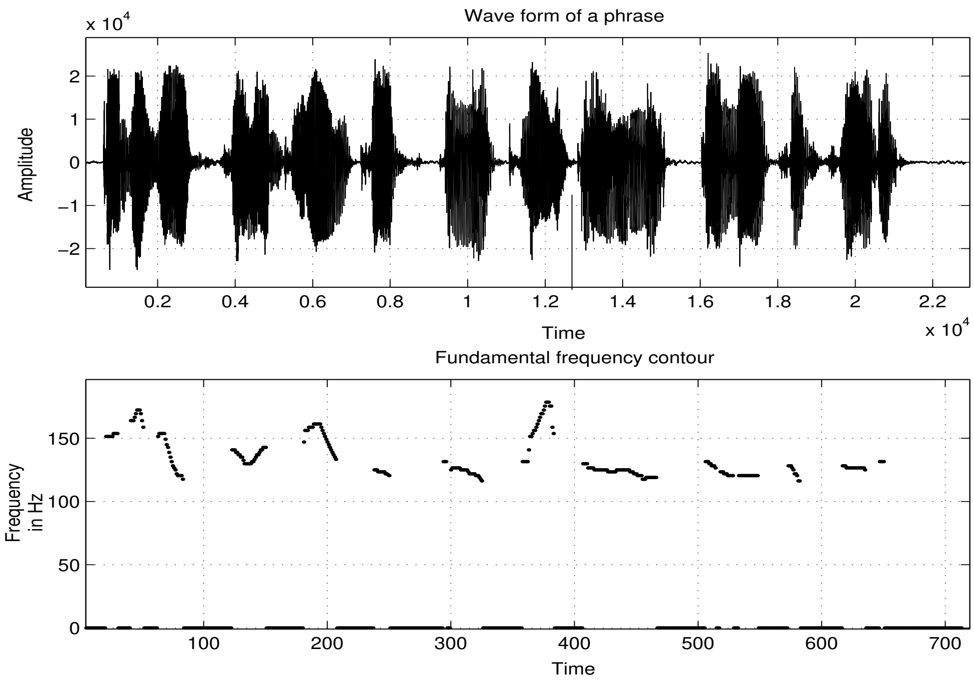

Fig. 6.6: Resetting in F_0 contour at a phrase boundary where the F_0 on the last syllable of the preceding phrase shows an inclining tendency. The vertical line indicates the phrase boundary of the compound phrase */tirigi svAdhInam chEsukonnaTlu Ayana teliyachEsAru/*.

6.4 CONCLUSIONS

In this chapter we presented a gross study on the behavior of the intonation characteristics of Telugu sentences. We have observed the fundamental frequency contours of simple and compound Telugu sentences and compared the results with those of Hindi obtained from the previous studies in [86]. We have also studied the pitch patterns of words with different number of syllables.

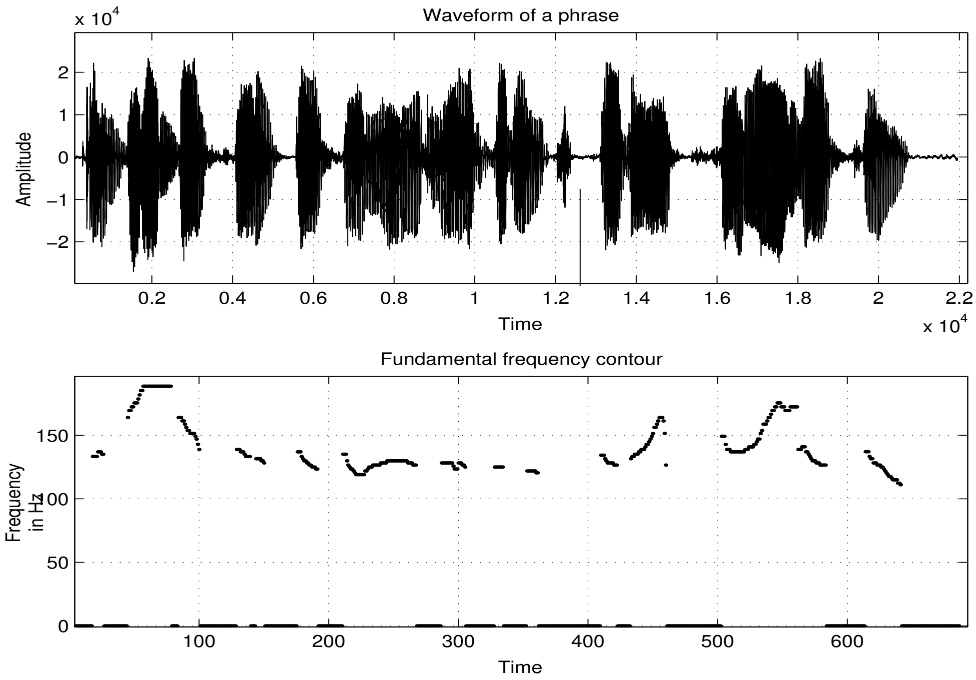

Fig. 6.7: Resetting in F_0 contour at a phrase boundary where the F_0 on the first syllable of the following phrase shows an declining tendency. The vertical line indicates the phrase boundary of the compound phrase */agra dEshAlto sambandhAla gurinchi kUDa samAvEsham/*.

CHAPTER 7

SUMMARY AND CONCLUSIONS

The main objective of this work was to achieve improvements in the quality of the synthesized speech from the existing text-to-speech (TTS) system for Hindi. In order to improve the naturalness and intelligibility of synthesized speech a set of precise rules from continuous speech of Hindi and Telugu were extracted.

Mainly three studies were carried out in the present work. The first was, the extraction of duration rules from continuous speech of Hindi. The database used for this purpose was developed from the read speech of news bulletins by three speakers: two of them were male speakers and the other was female. The rules extracted from the continuous speech are compared with those derived from the utterances of constrained sequences in Hindi. In our observation, it was found that there were major differences in the durational effects of continuous speech and the utterances of constrained sequences for Hindi.

The second study was formulation of duration rules for TTS system for Telugu. For this study also the database was prepared using the speech collected from news bulletins. As both Hindi and Telugu languages are phonetic in nature we did not find any major differences in the durational effects in these two languages.

The third study was on the acquisition of intonation rules for Telugu by analyzing the local fall-rise patterns and the resetting of the F_0 contour.

7.1 MAJOR CONTRIBUTIONS OF THE THESIS

The following are the major contributions of the present work on prosodic knowledge for TTS systems:

- Formulation of duration rules for a TTS system for Telugu
- Extraction of duration rules from continuous speech of Hindi
- Study of intonation patterns for TTS system for Telugu

7.2 SCOPE FOR FUTURE WORK

- The rules are extracted from a read style of speech which suits for a TTS system intended to perform the job of a news reader. But for a TTS system that can be used for general purpose conversation, the rules can be extracted by analysis of recording natural conversation.
- The rules interact in a multiplicative manner, whenever more than one rule is fired. A study can be carried out to devise some method dictating the interaction of the rules.
- In the present study the comparisons between the genders are studied with only one female speaker's data. A similar study can be conducted with more number of speakers of the same gender to have better conclusions of the effect of gender on duration and intonation variations.
- Similar studies can be carried out for other Indian languages, and by identifying the similarities and deviations, a common database of language-independent prosodic rules can be developed.

APPENDIX A

DURATION RULES OF THE BASIC TTS SYSTEM FOR HINDI

IF 2

character is in word final position.

character type is not cluster CCV.

THEN 1

increase duration by 30 percent.

IF 2

character is in word final position.

character type is cluster CCV_1.

THEN 1

increase duration by 15 percent.

IF 2

character is in word final position.

character type is cluster CCV_2.

THEN 1

increase duration by 30 percent.

IF 2

character is in word beginning position.

character type is not cluster CCV.

THEN 1

increase duration by 10 percent.

IF 2

character is in word beginning position.

character type is cluster CCV_1.

THEN 1

increase duration by 5 percent.

IF 2

character is in word beginning position.

character type is cluster CCV_2.

THEN 1

increase duration by 10 percent.

IF 2

character at phrase boundary position.

character has vowel region.

THEN 1

increase duration by 30 percent.

IF 2

character at penultimate phrase boundary position.

next character does not have vowel region.

THEN 1

increase duration by 30 percent.

IF 2

character at sentence ending position.

character has vowel region.

THEN 1

increase duration by 35 percent.

IF 2

character at penultimate sentence ending position.

next character does not have vowel region.

THEN 1

increase duration by 35 percent.

IF 2

character at breath group position.

character has vowel region.

THEN 1

increase duration by 35 percent.

IF 2

character at penultimate breath group position.

next character does not have vowel region.

THEN 1

increase duration by 35 percent.

IF 1

unit is at syllable boundary.

THEN 1

increase duration by 10 percent.

IF 2

unit belongs to CV category.

next character is voiced stop.

THEN 1

increase duration by 15 percent.

IF 3

unit belongs to CV category.

next character is voiced stop.

unit is trill.

THEN 1

increase duration by 11 percent.

IF 2

unit belongs to CV category.

next character is aspirated stop.

THEN 1

increase duration by 8 percent.

IF 3

unit belongs to CV category.

next character is aspirated stop.

unit is trill.

THEN 1

increase duration by 5 percent.

IF 2

unit belongs to CV category.

next character is trill.

THEN 1

increase duration by 30 percent.

IF 2

unit belongs to CV category.

next character is fricative ha.

THEN 1

decrease duration by 25 percent.

IF 2

unit belongs to CV category.

next character is nasal.

THEN 1

decrease duration by 8 percent.
IF 2
unit belongs to CV category.
next character is semivowel ya.
THEN 1
increase duration by 10 percent.
IF 2
unit belongs to CV category.
next character is semivowel va.
THEN 1
increase duration by 15 percent.
IF 2
unit belongs to V category.
next character is voiced stop.
THEN 1
increase duration by 21 percent.
IF 3
unit belongs to V category.
next character is voiced stop.
unit is trill.
THEN 1
increase duration by 15 percent.
IF 2
unit belongs to V category.
next character is aspirated stop.
THEN 1
increase duration by 11 percent.
IF 3
unit belongs to V category.
next character is aspirated stop.
unit is trill.
THEN 1
increase duration by 7 percent.
IF 2
unit belongs to V category.
next character is trill.
THEN 1
increase duration by 42 percent.
IF 2
unit belongs to V category.
next character is fricative ha.
THEN 1
decrease duration by 35 percent.
IF 2
unit belongs to V category.
next character is nasal.
THEN 1
decrease duration by 11 percent.
IF 2
unit belongs to V category.
next character is semivowel ya.
THEN 1
increase duration by 14 percent.

APPENDIX B

DURATION RULES DERIVED FROM FROM CONTINUOUS SPEECH OF HINDI

IF 2

character is in word final position.

character type is not cluster CCV.

THEN 1

increase duration by 29 percent.

IF 2

character is in word final position.

character type is cluster CCV_1.

THEN 1

increase duration by 10 percent.

IF 2

character is in word final position.

character type is cluster CCV_2.

THEN 1

increase duration by 42 percent.

IF 2

character is in word beginning position.

character type is not cluster CCV.

THEN 1

increase duration by 20 percent.

IF 2

character is in word beginning position.

character type is cluster CCV_1.

THEN 1

increase duration by 14 percent.

IF 2

character is in word beginning position.

character type is cluster CCV_2.

THEN 1

increase duration by 17 percent.

IF 2

character at phrase boundary position.

character has vowel region.

THEN 1

increase duration by 54 percent.

IF 2

character at penultimate phrase boundary position.

next character does not have vowel region.

THEN 1

increase duration by 17 percent.

IF 2

character at sentence ending position.

character has vowel region.

THEN 1

increase duration by 61 percent.

IF 2

character at penultimate sentence ending position.
next character does not have vowel region.
THEN 1
increase duration by 25 percent.
IF 2
character at breath group position.
character has vowel region.
THEN 1
increase duration by 43 percent.
IF 2
character at penultimate breath group position.
next character does not have vowel region.
THEN 1
increase duration by 24 percent.
IF 1
unit is at syllable boundary.
THEN 1
increase duration by 10 percent.
IF 2
unit belongs to CV category.
next character is voiced stop.
THEN 1
increase duration by 19 percent.
IF 3
unit belongs to CV category.
next character is voiced stop.
unit is trill.
THEN 1
increase duration by 17 percent.
IF 2
unit belongs to CV category.
next character is aspirated stop.
THEN 1
increase duration by 21 percent.
IF 3
unit belongs to CV category.
next character is aspirated stop.
unit is trill.
THEN 1
increase duration by 12 percent.
IF 2
unit belongs to CV category.
next character is trill.
THEN 1
increase duration by 21 percent.
IF 2
unit belongs to CV category.
next character is fricative ha.
THEN 1
decrease duration by 5 percent.
IF 2
unit belongs to CV category.
next character is nasal.
THEN 1
decrease duration by -10 percent.
IF 2
unit belongs to CV category.
next character is semivowel ya.

THEN 1

increase duration by -8 percent.

IF 2

unit belongs to CV category.

next character is semivowel va.

THEN 1

increase duration by -10 percent.

APPENDIX C

DURATION RULES DERIVED FROM CONTINUOUS SPEECH OF TELUGU

IF 3

character is in word final position.

character type is not cluster CCV.

character is voiced.

THEN 1

increase duration by 16 percent.

IF 2

character is in word final position.

character type is not cluster CCV.

character is unvoiced.

THEN 1

increase duration by 25 percent.

IF 2

character is in word final position.

character type is cluster CCV_2.

THEN 1

increase duration by 40 percent.

IF 2

character is in word final position.

unit is liquid.

THEN 1

increase duration by 25 percent.

IF 2

character is in word final position.

unit is nasal.

THEN 1

increase duration by 40 percent.

IF 2

character is in word final position.

unit is trill.

THEN 1

ncrease duration by 17 percent.

IF 2

character is in word beginning position.

character type is not cluster CCV.

THEN 1

increase duration by 39 percent.

IF 2

character at sentence ending position.

character type is not cluster CCV.

THEN 1

increase duration by 40 percent.

IF 2

character at phrase boundary position.

character type is not cluster CCV.
THEN 1
increase duration by 55 percent.
IF 1
character at breath group position.
THEN 1
increase duration by 42 percent.
IF 2
character is in word final position.
unit is semivowel.
THEN 1
increase duration by 24 percent.
IF 2
character type is not cluster CCV.
next character is voiced stop.
THEN 1
increase duration by 13 percent.
IF 2
character type is not cluster CCV.
next character is aspirated stop.
THEN 1
increase duration by 17 percent.
IF 2
character type is not cluster CCV.
next character is liquid.
THEN 1
decrease duration by 5 percent.
IF 2
character type is not cluster CCV.
increase duration by 27 percent.
IF 2
character type is not cluster CCV.
next character is trill.
THEN 1
increase duration by 17 percent.
IF 2
character type is not cluster CCV.
next character is semivowel.
THEN 1
decrease duration by 10 percent.
IF 3
character is not in word final position.
character is not in word beginning position.
character type is cluster CCV_2.
THEN 1
increase duration by 21 percent.

BIBLIOGRAPHY

[1] Rajesh Kumar S. R., "Significance of durational knowledge for a text-to-speech system in an Indian language," *M.S. Thesis, Dept. of Computer Science and Engineering, IIT Madras, Chennai - 36*, March 1990.

[2] Gupta S. and Schroeter J., "Pitch-synchronous frame-by-frame and segment-based articulatory analysis by synthesis," *J. Acoust. Soc. Amer.*, vol. 94, pp. 2517–2530, 1993.

[3] Dennis H. Klatt, "Software for a cascade / parallel formant synthesizer," *J. Acoust. Soc. Amer.*, vol. 67, pp. 971–995, 1980.

[4] Dennis H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Amer.*, vol. 82, pp. 737–793, 1987.

[5] Sharman R. A., "Concatenative speech synthesis using subphoneme segments," in *Proceedings of the Institute of Acoustics*, pp. 367–374, 1994.

[6] G. Chung and S. Seneff, "Hierarchical duration modeling for speech recognition using the ANGIE framework," in *Proceedings of the European Conference on Speech Processing and Technology*, vol. 1, pp. 22–25, 1997.

[7] Crystal, Thomas H. and Arthur S. House, "Segmental durations in connected speech signals: Syllabic stress," *J. Acoust. Soc. Amer.*, vol. 83, pp. 1574–1585, 1988.

[8] Crystal, Thomas H. and Arthur S. House, "The duration of American English vowels: an overview," *J. Acoust. Soc. Amer.*, vol. 16, pp. 263–284, 1988.

[9] Dennis H. Klatt, "Interaction between two factors that influence vowel duration," *J. Acoust. Soc. Amer.*, vol. 54, pp. 1102–1104, 1973.

[10] K. Power, "Durational modelling for improved connected digit recognition," in *Proc. Int. Conf. Spoken Language Processing*, vol. 2, pp. 885–888, 1996.

[11] Lea W. A., "Prosodic aids to speech recognition," in *Trends in Speech Recognition* (Lea W. A., ed.), pp. 166–205, New Jersey: Prentice-Hall, 1980.

[12] Kohler K. J., "Prosodic boundary signal in German," *Phonetica*, vol. 40, pp. 89–134, 1983.

[13] Werner, S. and Keller, E., *Prosodic Aspects of Speech.* John Wiley, 1994.

[14] J. P. Campbell, "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437–1462, September 1997.

[15] M. Mathew, "Combining evidence from different classifiers for text-dependent speaker verification," *M.S. Thesis, Dept. of Computer Science and Engineering, IIT Madras, Chennai - 36*, December 1999.

[16] B. S. Atal, "Automatic speaker recognition based on pitch contours," *J. Acoust. Soc. Amer.*, vol. 52, pp. 1687–1697, 1972.

[17] K. N. Reddy, "The influence of plosive consonant voicing and aspiration on duration of vowels in Telugu and Hindi," in *Perspectives on Language in Society* (S. K. Verma and D. singh, eds.), pp. 183–202, Delhi: Kalinga Publications, 1996.

[18] Bunnell, H.T., S. Hoskins and D. Yarrington, "Prosodic vs. segmental contributions to naturalness in a diphone synthesizer," in *Proc. Int. Conf. Spoken Language Processing*, vol. 5, pp. 1723–1726, Nov.30-Dec.4 1998.

[19] Crystal, Thomas H. and Arthur S. House, "Segmental durations in connected speech signals: Current results," *J. Acoust. Soc. Amer.*, vol. 83, pp. 1553–1573, 1988.

[20] Crystal, Thomas H. and Arthur S. House, "The duration of American English stop consonants: an overview," *J. Acoust. Soc. Amer.*, vol. 16, pp. 285–294, 1988.

[21] Umeda N., "Vowel duration in American English," *J. Acoust. Soc. Amer.*, vol. 58, pp. 434–445, 1975.

[22] Umeda N., "Consonant duration in American English," *J. Acoust. Soc. Amer.*, vol. 61, pp. 846–858, 1977.

[23] Dennis H. Klatt, "Structure of a phonological rule component for a speech synthesis by rule program," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 391–398, 1976.

[24] Mattingly, Ignatius G., "Synthesis by rule of prosodic features," *Language and Speech*, vol. 9, pp. 1–13, 1966.

[25] Mattingly, Ignatius G., "Experimental methods for speech synthesis by rule," *IEEE Trans. on Audio*, vol. 16, pp. 198–202, 1968.

[26] Coker, C. H., Umeda, N. and Browman, C. P., "Automatic synthesis from ordinary English text," *IEEE Trans. on Audio Electro acoustics*, vol. 21, pp. 293–297, 1973.

[27] Gaitenby, J. H., Sholes, G. N. and Kuhn, G. M., "Word and phrase stress by rule for a reading machine," *Status Report on Speech Research submitted to Haskins Laboratories, New Haven*, 1972.

[28] Dennis H. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *J. Acoust. Soc. Amer.*, vol. 59, pp. 1208–1221, 1976.

[29] van Santen J., "Combinatorial issues in text-to-speech synthesis," in *Proceedings of the European Conference on Speech Processing and Technology*, vol. 1, 1997.

[30] J. P. H. van Santen, "Assignment of segmental duration in text-to-speech synthesis," *Computer Speech and Language*, vol. 8, pp. 95–128, April 1994.

[31] W. N. Campbell, "Syllable-based segmental duration," in *Talking Machines: Theories, Models, and Designs* (G. Bailly, C. Benoit, and T. R. Sawallis, ed.), pp. 211–224, Elsevier, 1992.

[32] J. Allen, M. Hunnicut and Dennis H. Klatt, *From Text-to-speech: The MITalk system.* Cambridge University Press, 1987.

[33] N. Umeda, "Linguistic rules for text-to-speech synthesis," *Proceedings of the IEEE*, vol. 64, pp. 443–451, April 1976.

[34] D. O'Shaughnessy, "Design of a real-time French text-to-speech system," *Speech Communication*, vol. 12, pp. 233–243, March 1984.

[35] D. O'Shaughnessy, "A study of French vowel and consonant durations," *Journal of Phonetics*, pp. 385–406, 1991.

[36] K. Bartkova and C. Sorin, "A model of segmental duration for speech synthesis in French," *Speech Communication*, vol. 6, pp. 245–260, 1987.

[37] L. Lee, C. Tseng and M. Ouh-Young, "The synthesis rules in a Chinese text-to-speech system," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 9, pp. 1309–1320, 1989.

[38] R. Carlson and B. Granstrom, "A search for durational rules in a real-speech database," *Phonetica*, vol. 43, pp. 140–154, 1986.

[39] A. R. M. Simoes, "Predicting sound segment duration in connected speech: An acoustical study of Brazilian Portuguese," in *Proc. of the ESCA Workshop on Speech Synthesis*, pp. 173–176, 1990.

[40] M. D. Riley, "An overview of speaker recognition technology," in *Talking Machines: Theories, Models, and Designs* (G. Bailly, C. Benoit, and T. R. Sawallis, ed.), pp. 265–273, Boston: Elsevier, 1992.

[41] S.H. Lee and Y.H. Oh, "Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems," *Speech Communication.*

[42] L. Breiman, J. Friedman, and R. Olshen, *Classification and Regression Trees.* Wadsworth International, California, 1984.

[43] A. Maghbouleh, "An empirical comparison of automatic decision tree and hand-configured linear models for vowel durations," in *Proc. of the Second Meeting ACL Special Interest Group in Computational Phonology*, 1996.

[44] K. Silverman, M. E. Beckman, J. Pitrelli, M. Ostendorf C. Wightman, P. Price, J. Pierrehumbert and J. Hirschberg, "ToBI: a standard for labeling English prosody," in *Proc. Int. Conf. Spoken Language Processing*, vol. 2, pp. 867–870, 1992.

[45] Taylor, P. A., "The rise/fall/connection model of intonation," *Speech Communication*, vol. 15, no. 1, pp. 169–186, 1995.

[46] H. Fujisaki and S. Ohno, "Prosodic parameterization of spoken Japanese based on a model of the generation process of f_0 contours," in *Proc. Int. Conf. Spoken Language Processing*, vol. 4, pp. 2439–2442, 1996.

[47] Chris Traber, "F_0 generation with a database of natural F_0 patterns and with a neural network," in *Talking Machines: Theories, Models, and Designs* (G. Bailly, C. Benoit, and T. R. Sawallis, ed.), pp. 287–304, Elsevier, 1992.

[48] Sin-Horng Chen, S. H. Hwang and Y. Wang, "An RNN-based prosodic information synthesizer for Mandarin text-to-speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 6, pp. 226–239, May 1998.

[49] B. Yegnanarayana, S. Rajendran, V. R. Ramachandran and A. S. Madhukumar, "Significance of knowledge sources in text-to-speech system for Indian languages," *Sādhanā*, vol. 19, pp. 147–169, 1994.

[50] G. Fant, "The voice source-acoustic modeling," *Technical Report STL-QPSR*, vol. 4, pp. 28–48, 1982.

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

[52] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, 1975.

[53] Stephen D. Pearson, Hector R. Javkin, Kenji Matsui and Takahiro Kamai, "Text-to-speech synthesis using a natural voice source," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 193–196, 1994.

[54] D. O'Shaughnessy, *Speech Communication: Human and Machine*. Addison-Wesley, 1987.

[55] S. E. G. Ohman, "Coarticulation in VCV utterances: Spectrographic measurements," *J. Acoust. Soc. Amer.*, vol. 31, pp. 151–168, 1966.

[56] W. N. Campbell and S. D. Isard, "Segment durations in a syllable frame," *Journal of Phonetics*, vol. 19, pp. 37–47, 1991.

[57] K. J. Kohler, "Macro and micro f_0 in the synthesis of intonation," in *Papers in Laboratory Phonology I: Between Grammer and Physics of Speech* (K. J. and B. M. E., eds.), Cambridge University Press, 1990.

[58] Lehiste I., *Suprasegmentals*. MIT Press, 1970.

[59] A. S. Madhukumar, S. Rajendran, and B. Yegnanarayana, "Intonation component of text-to-speech system for Hindi," *Computer, Speech, and Language*, vol. 7, pp. 283–301, 1993.

[60] B. Yegnanarayana and D. K. Saikia and T. R. Krishnan, "Significance of group delay functions in signal reconstruction from spectral magnitude or phase," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 3, pp. 610–623, 1984.

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

[62] Bruce, G., Granstrom, B. and House, D., "Prosodic phrasing in Swedish speech synthesis," in *Proc. of the ESCA Workshop on Speech Synthesis*, pp. 125–128, 1990.

[63] Silverman, K., "The structure and processing of f_0 contours," *Ph.D. Thesis, University of Cambridge*, 1987.

[64] Terken, J. M. B. and Lemeer, G., "Effects of segmental quality and intonation on quality judgments for texts and utterances," *Journal of Phonetics*, vol. 16, pp. 453–457, 1988.

[65] A. Monaghan, "Heuristic strategies for higher-level analysis of unrestricted text," in *Talking Machines* (G. Bailly and C. Benoit, ed.), pp. 143–161, Amsterdam: Elsevier Science, 1995.

[66] Bolinger, D., "Accent is predictable (if you're a mind-reader)," *Language*, vol. 48, pp. 633–644, 1972.

[67] Hirschberg, J. and Pierrehumbert, J., "The intonational structuring of discourse," in *Proc. of the 24th Annual Meeting of the Association of Computational Linguistics*, pp. 136–144, 1986.

[68] Nooteboom, S. G. and Kruyt, J. G., "Accent, focus distribution, and the perceived distribution of given and new information: An experiment," *J. Acoust. Soc. Amer.*, vol. 82, pp. 1512–1524, 1987.

[69] *http://www.aczone.com/itrans/tblall/tblall.html.*

[70] Xuedong Huang, Alex Acero and Hsiao-Wuen Hon, *Spoken Language Processing: A guide to theory, algorithm, and system design.* Prentice Hall PTR, New Jersey, 2000.

[71] Fitzpatric, F. and Bachenko, J., "Parsing for prosody: what a text-to-speech system needs from syntax," in *Proc. of the Annual Artificial Intelligence Systems in Government Conference*, pp. 188–194, 1989.

[72] A. Monaghan, "Parsing unrestricted text: A multi-phase approach," *Proc. of the XIIth ICPhS*, pp. 526–529, August 1991.

[73] D. K. Oller, "The effect of position on in utterance on speech segment duration in English," *J. Acoust. Soc. Amer.*, pp. 1512–1524, 1987.

[74] Li-Chiung Yang, "Contextual effects on syllable duration," *Third ESCA/COCOSDA Int. Workshop on Speech Synthesis*, pp. 131–136, 1998.

[75] Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M. and Price, P. J., "Segmental durations in the vicinity of prosodic phrase boundaries," *J. Acoust. Soc. Amer.*, vol. 91, pp. 1707–1717, 1992.

[76] Richard S. McGowan, "Tongue-tip trills and vocal tract wall compliance," *J. Acoust. Soc. Amer.*, vol. 91, pp. 2903–2910, 1992.

[77] T. Balasubramanian, "The two r's and the two n's in Tamil," *Journal of Phonetics*, vol. 10, pp. 89–97, 1982.

[78] *http://www.aczone.com/itrans/tlgutx/node3.html.*

[79] Ostendorf M., "Linking speech recognition and language processing through prosody," *Journal for Integrated study of Artificial Intelligence, Cognitive Science and Applied Epistemology*, vol. 13, no. 5, pp. 79–103, 1997.

[80] Ramanujan A. K. and C. Masica, "Toward a phonological topology of the Indian linguistic area," *Current Trends in Linguistics*, vol. 5, pp. 543–576, 1969.

[81] Delattre P., "Some factors of vowel duration and their crosslinguistic validity," *J. Acoust. Soc. Amer.*, vol. 34, pp. 1141–1143, 1965.

[82] Vaissire J., "Language independent prosodic features," in *Prosody: Models and Measurements* (A. Cutler and D. R. Ladd, eds.), pp. 53–66, Springer Verlag, 1983.

[83] Yegnanarayana B., Hema. A. Murthy and Ramachandran V. R., "Speech processing using modified group delay functions," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 945–948, 1991.

[84] Rabiner L. R., Cheng M. J., Rosenberg A. E. and McGonegal C. A., "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 4, pp. 399–418, 1976.

[85] Hess W., *Pitch Determination of Speech Signals: Algorithms and Devices.* Springer-Verlag, Berlin, 1983.

[86] Madhukumar A. S., "Intonation knowledge for speech systems for an Indian language," *Ph.D. Thesis, Dept. of Computer Science and Engineering, IIT Madras, Chennai - 36*, January 1993.

LIST OF PUBLICATIONS

1. K. Kiran Kumar, K. Srinivasa Rao and B. Yegnanarayana, "Duration knowledge for a Text-to-Speech system for Telugu", *International Conference on Knowledge Based Computer Systems - 2002 (communicated).*