Sādhanā, Vol. 19, Part 1, February 1994, pp. 147–169. © Printed in India.

Significance of knowledge sources for a text-to-speech system for Indian languages

B YEGNANARAYANA[1], S RAJENDRAN, V R RAMACHANDRAN and A S MADHUKUMAR

Department of Computer Science and Engineering, Indian Institute of Technology, Madras 600036, India

[1]E-mail: yegna@iitm.ernet.in

Abstract. This paper discusses the significance of segmental and prosodic knowledge sources for developing a text-to-speech system for Indian languages. Acoustic parameters such as linear prediction coefficients, formants, pitch and gain are prestored for the basic speech sound units corresponding to the orthographic characters of Hindi. The parameters are concatenated based on the input text. These parameters are modified by stored knowledge sources corresponding to coarticulation, duration and intonation. The coarticulation rules specify the pattern of joining the basic units. The duration rules modify the inherent duration of the basic units based on the linguistic context in which the units occur. The intonation rules specify the overall pitch contour for the utterance (declination or rising contour), fall-rise patterns, resetting phenomena and inherent fundamental frequency of vowels. Appropriate pauses between syntactic units are specified to enhance intelligibility and naturalness.

Keywords. Text-to-speech system; prosodic features; coarticulation; intonation; formants; content word; function word.

1. Introduction

The function of a text-to-speech system is to convert a symbolic input (text) to an output speech waveform. To produce speech from a given text, human beings use several knowledge sources such as phonetics, phonology, morphology, syntax, semantics and pragmatics. It is necessary to incorporate these knowledge sources in a suitable form for a text-to-speech system to accomplish the same task. Mere concatenation of the signals corresponding to the basic units of speech does not produce intelligible and natural-sounding speech. The rules governing various knowledge sources are essential. Other than the production of isolated utterances of basic units, most of the knowledge sources are acquired by human beings without explicit training or learning. Moreover, these knowledge sources by themselves do not make up speech. Hence these knowledge source can be viewed as metalevel

knowledge. This paper addresses some issues related to the role of various knowledge sources in the development of a text-to-speech system for Indian languages.

Speech is the primary method for communication between human beings. Of the many varieties of life sharing our world, only human beings have developed the vocal means of coding to convey information beyond a rudimentary level. Through the development of a system for speech communication between man and machine, we constitute a whole new range of communication services to extend man's capabilities, serve his social needs, and increase his productivity. As computers become increasingly popular in nearly all segments of society, it is quite natural to consider a natural mode as medium of communication. Speech is obviously the most useful medium of communication between computers and its human users. The other possible method for representing natural communication is in text mode which can be considered as a string of conventional symbols (Allen 1985). Text is often considered a more durable medium of communication and is preserved more reliably. Hence it is widely used for both input and output of computers. But text requires specialized equipment as well as typing and reading skills which many potential users may not possess. On the other hand, speech is the most widely used communication medium between humans and requires no special training. Due to these advantages, there is a growing trend towards the development of speech systems over the past three decades.

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

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

This paper is organized as follows: § 2 discusses the design issues in the development of our text-to-speech system. The role of various knowledge sources in a text-to-speech system for Hindi is discussed in § 3. Section 4 discusses the improvement in the quality of synthetic speech by the addition of these knowledge sources.

2. A text-to-speech system for Hindi

We are developing a text-to-speech system for Indian languages based on a parameter concatenation model (Yegnanarayana *et al* 1990, pp. 467–76). As the speech is modelled using parameters, the voice characteristics can be manipulated and thus prosodic features can be incorporated by changing these parameters. The parametric representation is highly flexible and needs much less storage compared to the waveform concatenation model.

The design of our text-to-speech system is modular. It enables us to make changes in any of the modules independently. These modules can be integrated to the rest of the system. Each of the modules added to the system was developed in parallel. The various modules available in our system include the knowledge sources related to coarticulation phenomena (Ramachandran & Yegnanarayana 1992), duration (Rajesh Kumar 1990) and intonation (Madhukumar *et al* 1993).

Figure 1 shows the block diagram of our text-to-speech system. The input to the system is Hindi text stored in the form of ISCII (Indian Script Code for Information

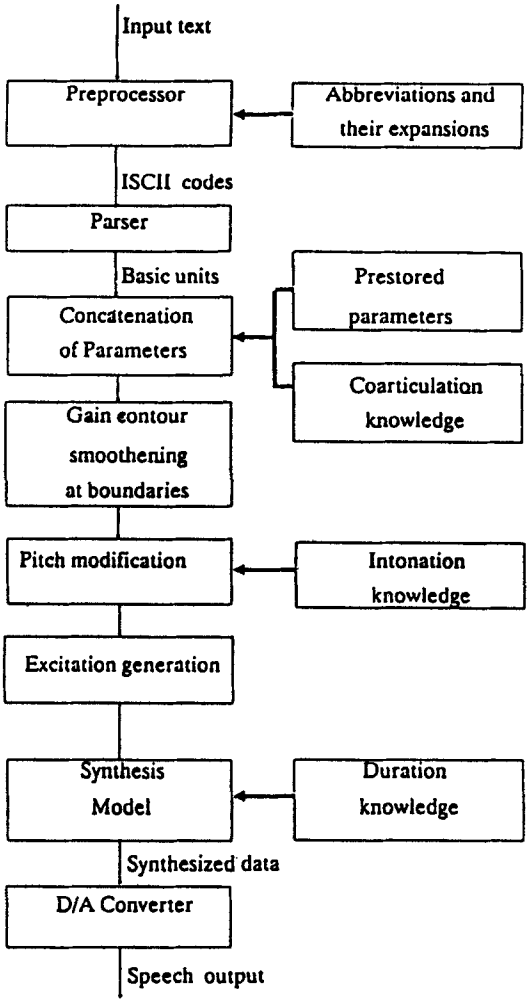


Figure 1. Block diagram of a text-to-speech system for Hindi.

Interchange) codes. The preprocessor scans the string of ISCII codes to locate abbreviations, numbers, dates and special symbols and replace them by their expansions in spoken form. Basic units are extracted from the expanded text using a simple parser. For synthesizing speech, parameters of the basic units of input text are concatenated and coarticulation rules that operate across adjacent basic units of speech are applied. The gain contour is smoothed at the boundaries between adjacent basic units. The pitch contour is modified to incorporate the intonation knowledge for the sentence being synthesized. The modified pitch and gain contours are used to generate an excitation signal. The excitation signal and the system parameters are used to generate the speech waveform.

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

2.1 *Choice of basic units*

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

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

The cluster characters can be generated from the constituent *CV* combinations and other consonants. For example, the cluster character /*kya:*/ can be generated by concatenating the consonant /*k*/ and the *CV* combination /*ya:*/. This results in the reduction of the number of basic units (from about 5000 to 400) and the storage requirement. Therefore the basic units in our text-to-speech system are: (1) isolated consonants (*C*); (2) isolated vowels (*V*) and (3) the consonant-vowel combinations (*CV*).

2.2 *Collection of basic units and extraction of parameters*

The basic units were extracted from the carrier words in isolation. The carrier words selected are meaningless words to avoid the undesirable prosodic bias introduced subconsciously by the speaker. Also it allows us to quickly form a suitable carrier word to make the extraction of the basic units easier. The required basic unit is placed in the word medial position followed by a stop consonant with some exceptions (Rajesh Kumar 1990).

To synthesize speech from a given text, our text-to-speech system uses the following speech parameters: 1) linear predictive coefficients (LPC), 2) formants, 3) pitch and 4) gain. LPC and formants represent the vocal tract system and pitch and gain parameters

correspond to the source information. In the following paragraphs, we discuss the extraction of these parameters briefly.

Our text-to-speech system is based on the linear prediction method. A set of 14 LPC parameters are used to model the vocal tract system. These are computed using the autocorrelation method (Makhoul 1975). The coarticulation effect is manifested in the speech wave mainly as a transition pattern of formants (the resonant frequencies of vocal tract) (Ohman 1966). A difficult signal processing problem is encountered in incorporating the coarticulation rules due to the incompatibility of the parameters used for basic untis (LPC) and for specification of the rules (formants). In order to solve this problem, all basic units are converted to a representable scheme in which the vowel regions are stored using formants and the consonant regions using LPC. Formants are extracted using the properties of group delay functions (Yegnanarayana *et al* 1991).

We modify the intonation pattern by incorporating intonation knowledge obtained by analysis of continuous speech in Hindi. The voiced/unvoiced decision of the basic units is stored separately and this is sufficient enough rather than the actual values of pitch. Later, based on the intonation knowledge the pitch contour of the utterances in voiced region is modified. In the unvoiced region the pitch value is taken as zero.

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

2.3 *Preprocessor and parser*

Preprocessor and parser are two preliminary modules in our text-to-speech system. The input and output of the preprocessor module are in ISCII code itself. The text is preprocessed to locate nonphonetic strings (such as numerals and abbreviations) which are replaced by their spoken form. For example, the abbreviation */ḍā:./* (Dr.) expands to its spoken form */ḍa:kṭar/* 'Doctor' and the numeral 120·45 expands to */e:k sau bi:s dašamlav ca:r pa:nc/* 'One hundred and twenty point four five'. It also helps the intonation module to make out if a particular word is a numeral or not. The preprocessed ISCII codes are transferred to the parser.

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

2.4 *Synthesis of speech from the parameters of basic units*

To synthesize speech from the parameters, we used a linear predictive technique (Atal & Hanauer 1971) in the consonant part and a cascade formant synthesizer (Klatt

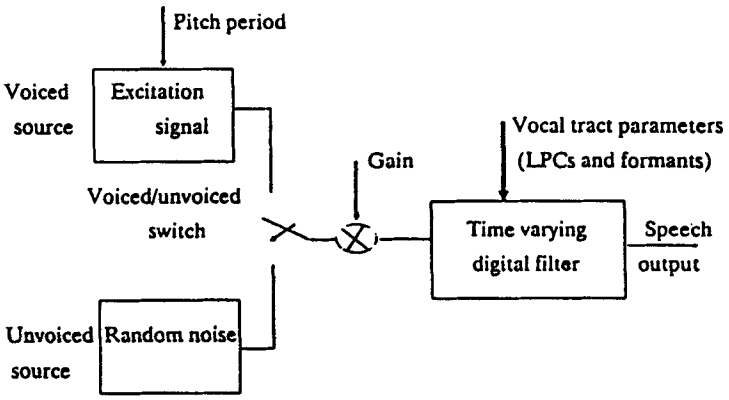


Figure 2. Block diagram for the basic speech production model. The vocal tract system is represented by LPC and formants. Pitch period and gain are the source parameters. The voiced source of the speech is represented by an excitation signal and the unvoiced source is represented by random noise.

1980) in the vowel part of the basic units. By using this hybrid method for synthesis, it is possible to capture the coarticulation behaviour of speech sounds as a set of transition patterns of formants in the vowel region.

The basic model for speech synthesis is given in figure 2. It consists of an excitation signal and a time-varying filter representing source and system parameters of speech signals, respectively. The excitation signal is periodic for voiced speech signals and a sequence of random numbers for unvoiced sounds. This excitation signal is fed to a time-varying digital filter which models the vocal tract for generating speech. The time-varying filter is represented by either LPC or formants based on the type of basic units.

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

3. The role of knowledge sources in a text-to-speech system for Hindi

For getting natural and intelligible speech output from a text-to-speech system, incorporation of various knowledge sources at segmental and suprasegmental domains is very important. Segmental-level knowledge sources are concerned with the appropriate representation of speech events by suitable parameters of basic units. Suprasegmental knowledge sources are those whose domains extend beyond a segment.

The important knowledge sources for a text-to-speech system are: (1) coarticulation rules corresponding to changes in acoustic parameters due to the influence of adjacent segments; (2) durational rules corresponding to the inherent duration of segments and their variation due to context; (3) pitch rules corresponding to inherent pitch, pitch variations across words, phrases and sentences, and (4) rules corresponding to

intensity variations across basic units, words and phrases. Rules related to duration, pitch and intensity are classified as prosodic knowledge. In the following sections, we discuss the acquisition and incorporation of knowledge corresponding to coarticulation, duration and intonation.

3.1 *Coarticulation*

The purpose of incorporating coarticulation knowledge in the text-to-speech system is to smooth the transition between adjacent basic units to improve the naturalness and fluency of synthesized speech. Figure 3 shows the waveforms and spectrograms

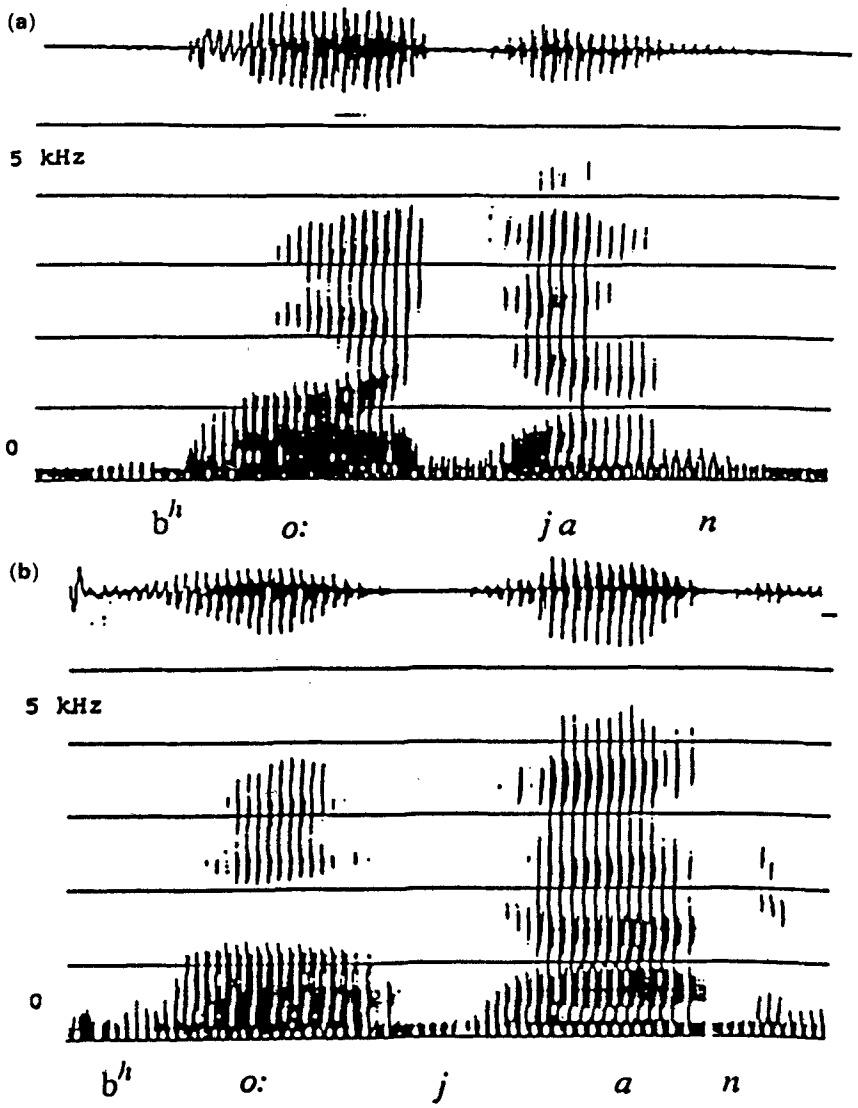


Figure 3. Illustration of the coarticulation effect between speech segments: (a) waveform and spectrogram of the word /b^ho:jan/ 'food'; (b) waveform and spectrogram of the characters /b^ho:/, /ja/ and /n/ uttered in isolation. The changes in the spectrogram of the vowels of /b^ho:/ and /ja/ in (a) are due to the context.

of the utterance /b^ho:jan/ and the isolated characters /b^ho:/, /ja/ and /n/. The change in formants towards the end of the vowels in /b^ho:/ and /ja/ are caused by the coarticulation. We study the coarticulation effect in terms of acoustic features like formant transitions, durational changes and intensity variations and formulate rules which predict how these parameters are to be manipulated in various contexts to get the desired coarticulation effect.

3.1a *Nature of coarticulation in Hindi*: The phenomenon of coarticulation involves changes in the articulation and acoustics of a phoneme due to the phonetic context (O'Shaughnessy 1987). The coarticulation phenomenon has an anticipatory and a carry-over component. In the former, features of a phoneme get modified depending upon the following phoneme or phonemes. In Hindi speech, this form of coarticulation is identified to be the following: (1) Change of the features of the vowel in vowel consonant (*VC*) and vowel to vowel (*VV*) sequences; (2) the vowel gets nasalized if the preceding or following sound is a nasal, and (3) consonant spectrum is influenced by the following vowel, that is, it may be different for the same consonant before different vowels. Carry-over coarticulation is the one in which features of the phoneme are modified by the preceding phoneme or phonemes. It is observed in Hindi speech as (1) the features of the vowel in a consonant vowel (*CV*) sequence is affected by the preceding *C* and (2) the vowel is nasalized to a larger extent if the preceding consonant is nasal.

The acoustic manifestations of the various coarticulation effects are (1) the transition from *C* to *V* in a *CV* unit; (2) the transition from *V* to *C* across a *VC* sequence; (3) the transition from vowel to vowel in a *VV* sequence; (4) nasalization of the vowel; (5) variation in the duration of the same unit in different contexts, and (6) changes in the spectral features of a consonant in different contexts, namely the *CV*, *VC* and *CC* contexts. Of these, the first three transitions are characterized by formant transition patterns and gain variation across the transition whereas the nasalization is characterized by the presence of antiformants in the vowel spectrum. The *CV* transition is embedded (or preserved) in the basic unit representation and hence we do not need any contextual rules to make these changes. Our interests are primarily in *VC*, *VV* and *CC* transitions and nasalization of vowels. These are not represented or preserved in basic units since the context which decides the changes is not in the same unit. Hence it is necessary to develop contextual rules to impart coarticulation across unit boundaries and also for nasalization of vowels.

In the study of the coarticulation between two phonemes, we do not consider the influences of segments beyond the immediately neighbouring units. This greatly simplifies the study by reducing the number of cases to be considered. Coarticulation effects due to the phonemes which are not immediately adjacent are less important in fluent speech production though they reduce the speaker's effort in articulation (O'Shaughnessy 1987).

3.1b *Acquisition and formulation of coarticulation knowledge*: The issues involved in the acquisition and formulation of the coarticulation knowledge, from a point of view of incorporating into a text-to-speech system are (1) identification of the domain of coarticulation; (2) classification of the coarticulation patterns, and (3) formulation of the coarticulation patterns.

The domain of coarticulation is the transition between basic units of speech. The transitions between the basic units can be only one of the sequences of vowel-to-vowel (*VV*), vowel-to-consonant (*VC*) and consonant-to-consonant (*CC*).

Considering all combinations of two basic units, the total number of junctions possible is about a few hundred. We classify these into a small number of basic transition patterns on the basis of the similarities in the transition patterns. The nature of the transition pattern in a *VC* depends on the articulatory features of the consonant (Ohman 1966). The *VC* transitions are grouped into six distinct classes, each of which corresponds to a place of articulation of the consonant. Each *VC* class is further divided into subgroups based on the manner of articulation of the consonant. Each *VC* subgroup contains formant transition patterns for five different vowels of Hindi. The vowel to vowel formant transition patterns in Hindi are few in number compared to the *VC* transitions. Consonant to consonant transitions (*CC*) are part of cluster characters. The coarticulation across *CC* transitions is more complicated. Few common effects associated with *CC* transitions in Hindi are (1) release of word final cluster, (2) shortening of geminated clusters (geminated cluster is the one in which both consonants are the same), and (3) lengthening of the consonant before a semivowel.

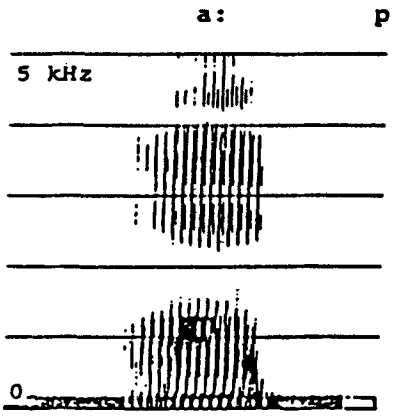
For each *VC* subgroup, rules for transition patterns are formulated for the five different vowels as the percentage deviation of formant frequencies from the steady values and the transition duration. The parameters required for the rule specification are obtained from the analysis of natural speech data. Figure 4 shows the spectrogram of a *VC* transition */a:p/* and also the rules formulated for all bilabial *VC* transitions.

Vowel-to-vowel (*VV*) transitions are characterized by gradual transition of formants from one vowel to the following one. This allows us to take care of all vowel to vowel transitions by a single rule which does the interpolation of vowel formants. Two nasalization rules – one for *CV* and other for *VC* transitions – simulate the nasalization effect. The nasals and nasalized vowels are characterized by the presence of antiresonances in the frequency spectrum. The antiresonance is simulated using a pole-zero pair below the first formant of the vowel. The consonant cluster rules manipulate the duration and the release part of the clusters but do not modify the spectral information of the constituent consonants.

3.1c *Incorporation of coarticulation knowledge*: Incorporation of the coarticulation into a text-to-speech system involves the application of the rules formulated to modify the default parameters of the basic units of the input text before synthesis, by making use of the phonetic context of the basic units in the text. The issues to be addressed in this are those of using the appropriate representation and activation methods. Since coarticulation modifies most of the parameters of the basic unit, it is very convenient to merge the knowledge activation with the synthesis process. The block diagram of the knowledge activation and synthesis scheme is shown in figure 5.

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

(a)



(b)

Vowel	a	e	i	o	u
% change					
in F1	-40	-20	0	0	0
F2	-30	-38	-32	-20	-26
F3	0	-8	-20	0	0
Trans Dur (in frames)	4	4	3	4	4

(c)

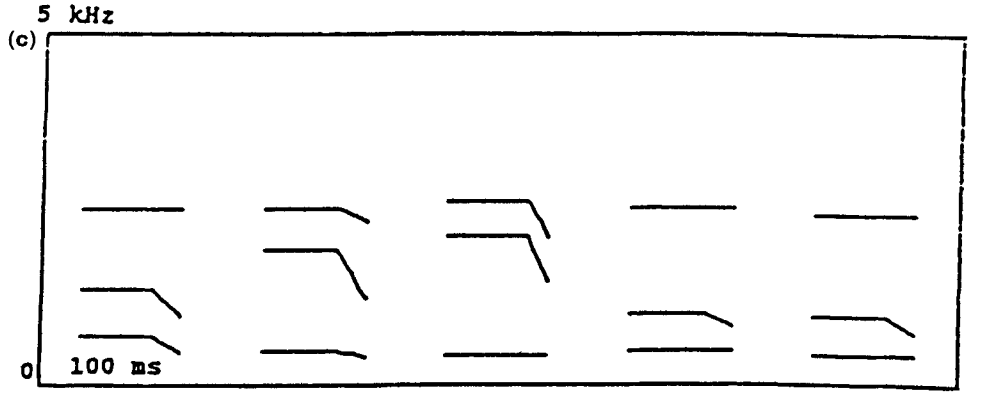


Figure 4. (a) Spectrogram of the *VC* sequence */a:p/*. (b) Table of *VC* formant transitions from different vowels to the bilabial stops. The transition is specified as the percentage changes in steady formant values of the vowel and the transition duration in frames of 6·4 ms. (c) Formant transition patterns generated from the above table. The first one of these corresponds to the spectrogram in (a).

rules are activated in this step. These rules are stored in a table and retrieved and activated as we synthesize the basic units in the TEMPCVTABLE one by one.

The coarticulation rules incorporated into the system are tested to (1) verify their correct activation and (2) evaluate their perceptual significance. This was done by

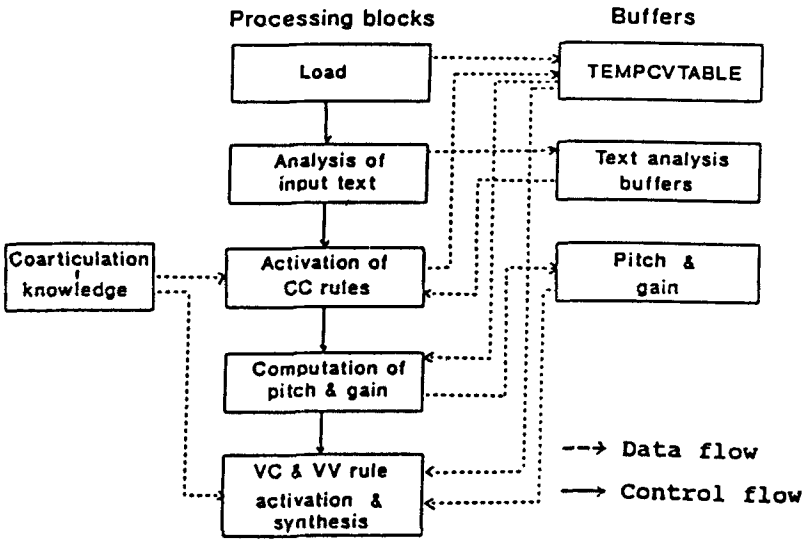


Figure 5. Incorporation of coarticulation knowledge in a text-to-speech system for Hindi.

analytical and perceptual means. Analytical testing of the rules is done by simulating the context in which the rules would be activated and then analysing the synthesized speech by signal processing means. The *VC* and *VV* transition rules and the *CC* or cluster consonant rules are tested. Figure 6 shows formant transition patterns of some

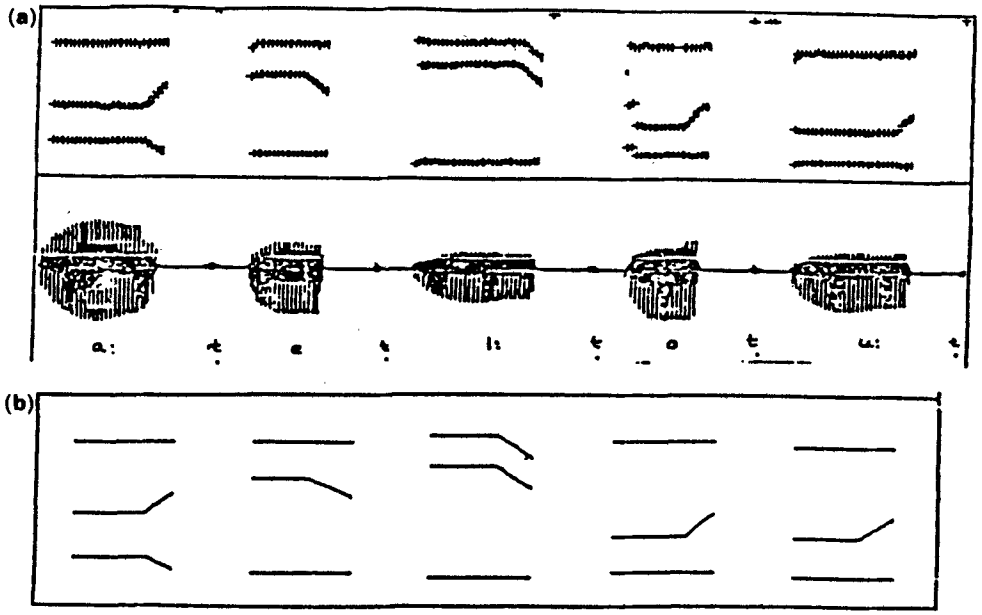


Figure 6. Formant transition patterns of *VC* sequences involving /t/. (**a**) Formant contours extracted from synthesized speech. (**b**) Formant transition patterns specified by rules.

synthesized *VC* transitions and the actual transitions specified by the rules. In perceptual testing, words synthesized with and without the transitions incorporated, were heard in pairs. The rules make significant improvement in the perceptual quality. The perceptual evaluation is done at higher levels like in sentences and paragraphs. Selected sentences and paragraphs are synthesized both with and without the coarticulation rules. The speech synthesized with the application of the rules showed significant improvements in naturalness.

3.2 *Duration*

Segmental duration is dependent both on inherent properties of the input unit concerned and on a large number of phonetic and structural constraints imposed contextually (Klatt 1976). The durations of various speech units vary considerably due to factors such as speaker (male vs female), speaking style (reading vs conversational) and speaking rate. Duration of speech units also depends upon the psychological state (fear, anger, sorrow etc.) of the speaker. When a person speaks more slowly than normal, pauses account for more durational increase than the speech units. At faster rates all normal durations of speech units shorten by a certain amount. The interplay of all these factors in natural continuous speech makes duration an extremely difficult feature to study.

3.2a *Nature of duration of speech sounds in Hindi*: For the sake of simplicity, durational effects can be roughly categorized as the effects due to position (POS), syllable boundary (SYL), prepausal lengthening (PPL), post-vocalic consonant (PVC), place of articulation (POA), change in cluster environment (CCL), semantic novelty (NOV), and polysyllabic shortening (PSS). Broadly speaking, the durational effects adjust the durations of basic units depending upon their (i) position and (ii) context in the given text. POS, SYB and PPL modify durations of basic units depending upon their position in the text. Among the durational effects that modify durations of basic units depending upon context, PVC, CCL and POA handle coarticulation at the phonetic boundaries. NOV and PSS cover other contextual phenomena, which are less frequent. In the following sections we discuss each of these effects in detail. These discussions pertain to a limited number of words or nonsense syllables in controlled reading situations by a single speaker. To the extent that the speaker is consistent, effects due to speaking rate are limited. From these studies, rules have been formulated to modify the duration of basic units in the given text.

(i) *Positional effect (POS)* – A character is more lengthened in a word final position than in a word beginning position, which in turn is longer than in a word medial position. The POS results were obtained after analysing durations of about 50 basic units. We proceeded as follows: For each basic unit we formed three nonsense words. Each word contained two or three characters. The first word contained the basic unit (under consideration) in word medial position. The second and the third words contained the basic unit in the word beginning and the word final positions respectively. In these three words, the basic unit is followed by a speech sound of the same category (such as voiced aspirated or voiced unaspirated). This was done to nullify other effects, PVC in particular, so that POS could be studied in isolation. The durations of the basic unit in all three words were measured. The percentage increase of the duration of the basic unit in the word beginning (or word final) position over that

in the word medial position is taken as the value α for that basic unit for word beginning (or word final) lengthening effect.

(ii) *Syllable boundary effect* (*SYB*) – The duration of the basic unit appearing just before a syllable boundary is increased. The SYB result was obtained after analysing durations of 24 basic units. We proceeded as follows: For each basic unit, we formed two nonsense words. Each word contained three characters. The first word contained the basic unit in word medial position. The second word contained the basic unit in word initial position with a syllable boundary following it. In these two words, the basic unit is followed by a speech sound of the same category (such as voiced aspirated or voiced unaspirated). The durations of the basic unit in the two words are measured. The percentage increase of the duration of the basic unit in the second word over the first is the combined effect of POS and SYB effects. We remove the POS effect from this in order to obtain the value of α (α indicates the percentage by which the 'normal' duration of the unit concerned is modified) for that unit for the SYB effect (it is in this way that we figured out that our rules ought to combine multiplicatively rather than additively).

(iii) *Prepausal lengthening effect* (*PPL*) – The duration of the character appearing before a pause is increased. PPL effect can be attributed to slowing down of speech in anticipation of a pause, aiding perceptual cues to syntactic boundaries. There is an increase in the duration of either the final or penultimate character of a word just before a pause. If the final character has a vowel, then the increase is only in the final character. Otherwise the increase is in the penultimate character.

The PPL results were obtained after analysing durations of about 20 basic units. This study was performed on continuous speech data and hence the basic units were embedded in meaningful words. The test set consisted of about 25 continuous sentences spoken with natural intonation and rhythm, so that the PPL effects are clearly observed. The duration of each basic unit before a pause (due to either a phrase boundary, or a breath group, or a sentence ending) was measured. The duration of the basic unit, when followed by an unaspirated and an unvoiced stop, is also measured. The percentage increase in the duration of the basic unit in the former over the latter case is the combination of POS and PPL effects. We remove the effect of POS to obtain the value for that basic unit for the PPL effect.

(iv) *Post-vocalic consonant effect* (*PVC*) – This effect states that the duration of a vowel changes depending upon the type of consonant following it. Voicing, aspiration, sonority and nasality of the PVC, all these affect the duration of the preceding vowel. The PVC results were obtained after analysing the durations of 20 basic units in different contexts. The basic unit could be either a *CV* combination or an isolated vowel. This effect was examined for the vowels /*a:*/, /*i*/ and /*u*/. It was later verified for the remaining vowels. For each basic unit we performed two experiments: (i) PVC effect due to a stop consonant, and (ii) PVC effect due to a nonstop consonant. These two cases are explained.

(a) PVC effect due to a stop consonant: Four nonsense words are formed with the basic unit in the word medial position. The basic unit is followed by four different cases of PVC. They are (1) unvoiced and unaspirated stops; (2) voiced and unaspirated stops; (3) unvoiced and aspirated stops; (4) voiced and aspirated stops.

The percentage increase of the duration of the vocalic portion of the basic unit in (2), (3) and (4) over (1) gives the value for each unit for each category of the PVC.

(b) PVC effect due to a nonstop consonant: Five nonsense words are formed with the basic unit in the word medial position. The basic unit is followed by five different cases of PVC. They are (1) unvoiced and unaspirated stops; (2) trill; (3) fricatives; (4) nasals; (5) semivowels. The percentage increase of the duration of the vocalic portion of the basic unit in (2), (3), (4) and (5) over (1) gives the value of α for each unit for each category of the PVC. The PVC results were verified in the case of continuous speech for some basic units.

(v) *Place of articulation effect* (*POA*) – If two adjacent characters (within as well as across word boundaries) have the same place of articulation, then one or both of the characters are shortened. This is due to relative ease of pronouncing sequences of speech sounds with the same place of articulation.

(vi) *Changes in cluster environments* (*CCL*) – In the case of cluster characters (*CCV* or *CCCV*), the durations of the various constituent basic units change due to the presence of adjacent consonants. They often shorten due to proximity of the POS, and sometimes lengthen due to relative difficulty of pronouncing certain sequences of consonants with conflicting articulatory requirements.

(vii) *Polysyllabic shortening effect* (*PSS*) – If the number of characters in a word is greater than three, then the vocalic durations of the various characters are reduced. This effect may relate to communication efficiency: words with many units are easier to identify than short words, which could allow spending less time per unit without risking perceptual mistakes.

3.2b *Incorporation of durational knowledge*: It is possible to vary the duration of a basic unit by varying the number of samples to be synthesized per frame (by default 64 samples are synthesized per frame). For instance, if the number of samples for all frames in a basic unit is doubled, the duration of the basic unit is doubled. The issues involved in incorporation of durational knowledge are (1) analysis of input text; (2) deciding the base duration for each unit; (3) representation and activation of knowledge. Figure 7 places these issues in the overall scheme of our text-of-speech system. In the following sections we discuss each of these issues in detail.

(i) *Analysis of input text* – The input text is analysed to obtain necessary information to enable the activation of durational knowledge such as (1) type of the basic unit (*C* or *CV*), (2) type of consonant in the basic unit, (3) position of the character in the word, (4) number of characters in the word, (5) markers for phrase boundaries and breath groups in the input text, and (6) syllabification within a word.

(ii) *Deciding the base duration of each unit* – The base duration of each basic unit is its duration in the carrier word where it occurs in the word medial position. Since the same basic unit will be used in all types of context and position wherever it occurs in the input text, it is imperative that the stored basic unit be devoid of the influence of any of the durational effects mentioned earlier. From this point of view some guidelines observed in forming the carrier word for a basic unit, can be explained as follows: (1) The basic unit must be followed by an unaspirated and an unvoiced stop

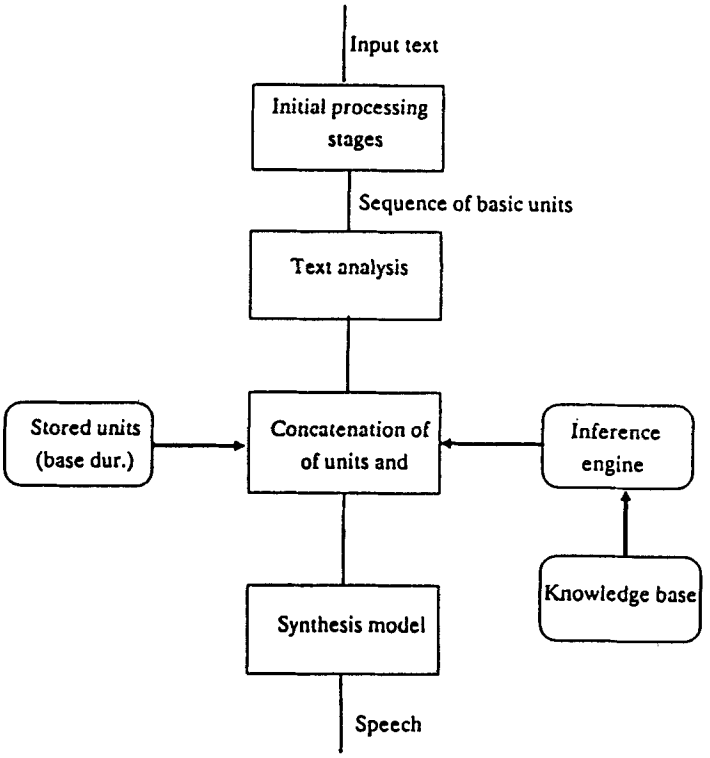


Figure 7. Incorporation of durational knowledge in a text-to-speech system for Hindi.

in order to nullify the PVC effect. (2) Each carrier word must have three characters. This is to nullify the PSS effect. (3) The basic unit and its adjacent characters must not have the same place of articulation. This is to nullify the POA effect. In other words, the base duration of a basic unit is its length in a neutral phonetic context. Depending upon the context in the given text, various durational deviations (using durational rules) are effected. This, in essence, summarizes the durational model used in our present system.

(iii) *Representation and activation of knowledge* – We used the production system approach (Rich 1983) to represent the durational knowledge in our text-to-speech system. Each rule in the knowledge base is an independent fragment of knowledge and does not rely on the correctness of other rules. This facilitates successive updating since the rules are independent of each other, and the order of declaration of rules is not important. For most artificial intelligence application domains, where the knowledge is not systematically formulated (as in our problem), the production system formalism offers a natural way of encoding the knowledge. Besides the production system rules provide an easy way of explaining the intermediate decisions taken.

Since we have represented the durational knowledge as rules, the activation of knowledge is achieved by means of a rule-based inference engine (or a rule interpreter). Depending upon the context of each basic unit in the given text, various rules may be applied. In the modelling duration, it is to be decided whether rules for lengthening

or shortening should be expressed absolutely or as percentages and whether the rules should combine by addition or multiplication. The rules combine multiplicatively if more than one rule fires for the same unit. Thus the order in which the rules combine does not matter. After application of all durational rules, the base duration of each basic unit is modified to obtain its duration to be used during synthesis. The inference engine is of the forward chaining type or is data driven. It is applied for each basic unit in the given text. After all the rules are applied, speech is synthesized using the modified durations of the basic units. The quality of the synthetic speech is improved significantly.

3.3 *Intonation*

Intonation pattern is defined as the variation of the fundamental frequency (F_0) with time. An utterance may convey a different meaning due to the changes in intonation even if it is composed of the same segmental phonemes. Intonation helps to group words into syntactic blocks for semantic interpretation of utterances. As in many other languages, intonation patterns in Hindi show some regular features. F_0 contour of declarative sentences decline gradually with time and for interrogative sentences F_0 contour rises towards the end. This backdrop declination or rising is characterized by local falls and rises. F_0 contour gets modified across major syntactic boundaries which is called resetting of F_0 contour. The resetting is used as a marker for phrase boundaries and it is accompanied by a pause. The intonation pattern of an utterance is also affected by segmental factors of constituent units. In the following sections, we discuss the issues in acquisition of intonation knowledge from continuous speech and incorporation of this knowledge in a text-to-speech system for Hindi. It includes a discussion on the properties of intonation patterns in Hindi such as declination/rising tendency, local fall-rise patterns in Hindi, resetting of F_0 contour, significance of pause and on the effects of phonetic factors on intonation knowledge, and the issues in the incorporation of intonation knowledge in a text-to-speech system for Hindi.

3.3a *Properties of intonation knowledge in Hindi*: For the present analysis, we have used the reading style of speech. A corpus of 500 sentences was read out by two adult male native speakers of Hindi. Speech was digitized to 12 bits/sample at a sampling rate of 10 kHz. A 256-sample analysis frame with a shift of 64 samples was used for extracting pitch. The algorithms for pitch extraction are based on simplified inverse filter tracking (Markel 1972) and properties of group delay functions (Yegnanarayana *et al* 1991). In the following sections we discuss the properties of intonation patterns for continuous speech in Hindi in detail.

(i) *Declination/rising tendency* – Properties of F_0 declination in Hindi can be summarized as follows: (1) Declination of F_0 contour in Hindi is characterized by falls (valleys) and rises (peaks). (2) These falls and rises fluctuate between two abstract lines – a top line and a base line, drawn near or through all maxima and minima F_0 values in a sentence, respectively. (3) The difference between valley and next peak (range of F_0 contour) decreases with time. (4) In a neutral declarative sentence the maximum value of F_0 will be located in the first *content word* (semantically meaningful word) itself. (5) In connected speech, the monosyllabic function words (words which have only grammatical value) conjoin with the preceding or following content words.

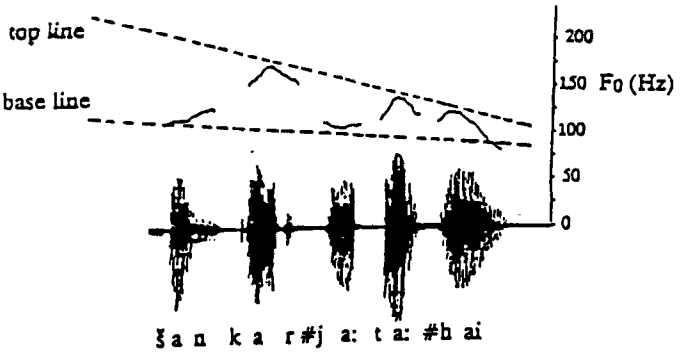


Figure 8. Speech waveform and F_0 contour for simple declarative sentence */šankar ja:ta: hai/* 'Shankar goes'. # indicates word boundary.

Speech waveform and F_0 contour for a natural utterance of a simple declarative sentence are shown in figure 8. The F_0 contour starts at the initial syllable of the first word and rises towards the next target, that is, the final syllable of the first content word. The F_0 rises and falls damp off at the end of the utterance. It is possible to draw a line connecting all the peaks (top line) and another connecting all the valleys (base line). Both lines decline monotonically and converge towards the end.

Interrogative sentences in Hindi can be broadly classified into two. They are (1) yes-no type questions and (2) question-word type questions. Yes-no type interrogative sentences in Hindi have the same grammatical structure as declarative sentences, except that optionally the question may include the question word */kya:/* usually at the beginning of the sentence. Question-word type interrogative sentences expect detailed answers and are marked in Hindi with any one of a set of interrogative words in Hindi. Intonation patterns for both types of interrogative sentences are different.

Questions expecting yes-no answers have a continuous rise in F_0 contour and hence the top line and the base line rise towards the end. The intonation pattern for question-word type interrogative sentences exhibit a dual nature. The top line and the base line decline gradually up to the question word and then rise towards the end. The fall-rise pattern does not change with respect to the type of the sentence. Figure 9 shows the speech waveform and the F_0 contour for a yes-no type interrogative

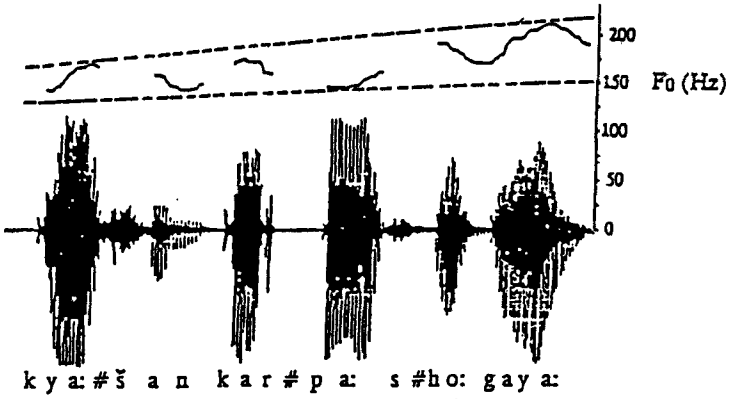


Figure 9. Speech waveform and F_0 contour for a yes/no type interrogative sentence */kya: šankar pa:s ho:gaya:/* 'Has Shankar passed?'.

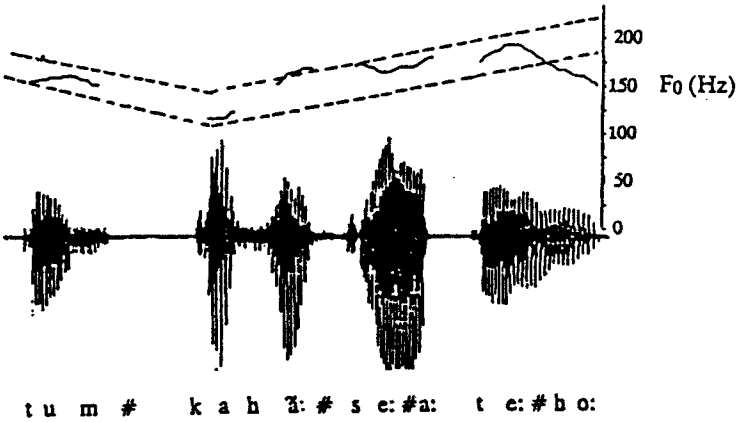


Figure 10. Speech waveform and F_0 contour for a question-word type interrogative sentence */tum kahã: se: a:te: ho:/* 'Where do you come from?'.

sentence. The F_0 falls and rises repeatedly till the end of the utterance. Figure 10 shows the speech waveform F_0 contour for a question word type interrogative sentence. Here the question word is */kahã:/* (second word in the sentence). The F_0 contour declines up to the question word and then rises.

(ii) *Local fall-rise patterns in Hindi* – F_0 contour of content words in Hindi exhibits a regular pattern of valleys and peaks, corresponding to the prominence of a particular syllable in a word or phrase. By analysing large amounts of data, we have observed some general features of valleys and peaks of F_0 contour which are determined by the phonological pattern of the words. The following are some observations for Hindi sentences: (1) The valleys and peaks are mostly associated with the vowels which are the nuclei of the syllables. However, the exact target point (valley or peak) within the voiced region of the syllable is determined by several factors. For example, the peak of the nucleus get shifted to the coda (the consonant that follows the vowel nucleus of syllable) if the consonant is either nasal or lateral. (2) If the word is monosyllabic then the valley and the peak occur within the same syllable and hence F_0 rises steadily. (3) In the case of disyllabic and trisyllabic words the peak occurs on the final syllable and the valley occurs on the initial syllable. (4) Tetrasyllabic words show two patterns: (a) a valley on the initial syllable and peak on the final syllable; (b) the valley and peak occur on alternate syllables and hence are characterized by two valleys and two peaks. The latter type is more likely in compound words. (5) The pattern for pentasyllabic words is similar to a combination of disyllabic and trisyllabic words.

(iii) *Resetting of F_0 contour* – F_0 resetting occurs across major syntactic boundaries and is accompanied by a pause. The part of utterance delimited by such a pause is called *intonational phrase.* The general properties of F_0 resetting obtained from the analysis are summarized in the following sections.

The *initial peak* F_0 (F_0 value of the first peak of the first intonational phrase) is constant for a particular speaker. All other significant peaks and valleys in the subsequent clauses can be related to the initial peak F_0. Within a syntactic clause, the F_0 contour is similar to the F_0 contour of a simple sentence. That is, the declination

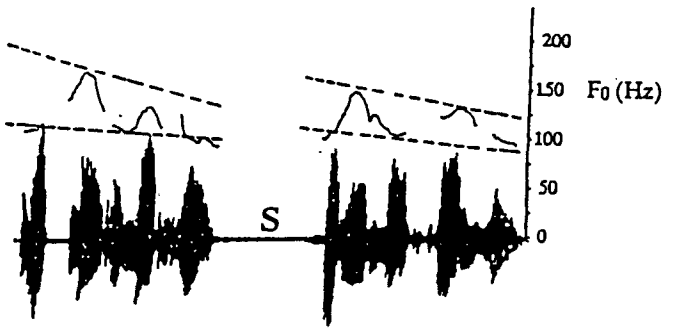


Figure 11. Speech waveform and resetting of F_0 contour at clause boundary (*S*) in a complex declarative sentence */a:tma: amar hai šari:r na:šva:n hai/* 'Soul is immortal, body is mortal'.

of F_0 contour is accompanied by local falls and rises. Figure 11 shows the effect of F_0 resetting across syntactic boundaries. The sentence */a:tma: amar hai, šari:r na:šva:n hai/* (Soul is immortal, body is mortal) has two syntactic clauses and the F_0 contour drifts down as a function of time till the occurrence of major syntactic break (at the end of */a:tma: amar hai/*), which is also marked by a significant pause of duration of about 300 ms.

The major factors which can affect F_0 resetting are physiological constraints, syntactic constraints and semantic constraints. Physiological constraints are the limitations imposed by the human speech production mechanism. Syntactic constraints include the changes in F_0 resetting with respect to the changes in the type of the sentence. Semantic constraints are the semantic aspects which control the properties of F_0 resetting.

(iv) *Significance of pause* – Pauses have been assigned to two main functions: (1) They separate large grammatical units, such as syntactic clauses. (2) They serve to clarify subgrouping of smaller units. Pauses can occur between words, intonational phrases and sentences. The characteristics of pauses for continuous speech in Hindi are summarized as below.

Speakers give different durations of pauses between words in continuous speech. The durations of pauses between words are controlled by several factors like lexical content of the words, position of the word in an intonational phrase and the phonetic factors of post-pause and pre-pause syllables. The amount of pause between words is much less than the amount of pause between intonational phrases. Between intonational phrases the amount of pause is determined by the type of constituent boundary. The amount of pause between sentences will be greater than pause between words and pause between intonational phrases.

(v) *Effect of segmental factors on F_0 contour* – Acoustic phonetic behaviour of vowels and consonants of surrounding speech units alter the F_0 values of vowels. F_0 values of vowels were studied by embedding the test words in a carrier sentence. For this study we have selected several possible combinations of disyllabic words. The test words are mostly nonsense words where the vowel characteristics are studied for both

initial and final syllables separately. The results from this analysis are summarized below.

There is a correlation between height of the vowel and its inherent F_0. If other factors remain constant, high vowels (/*i*/ and /*u*/) exhibit higher F_0 than low vowels (/*a*/). If the quantity of the vowel increases without changing any other factor, then inherent F_0 also increases. For example, the long vowel /*a:*/ has higher F_0 than the shorter counterpart /*a*/ at the same position. In all the cases, final vowels have greater F_0 than initial vowels. Within each test word further regular variations in F_0 were observed by changing the preceding or following syllables. However these changes are very small when compared with the changes due to other properties of F_0 contour.

3.3b *Incorporation of intonation knowledge*: In a text-to-speech system, intonation refers to the periodicity of glottal pulse source in voiced speech segments. There are different stages required for the incorporation of intonation knowledge in a text-to-speech system. They are summarized as follows: (1) Input text has to be parsed to find out the type of sentence and the corresponding intonation behaviour. An intonation

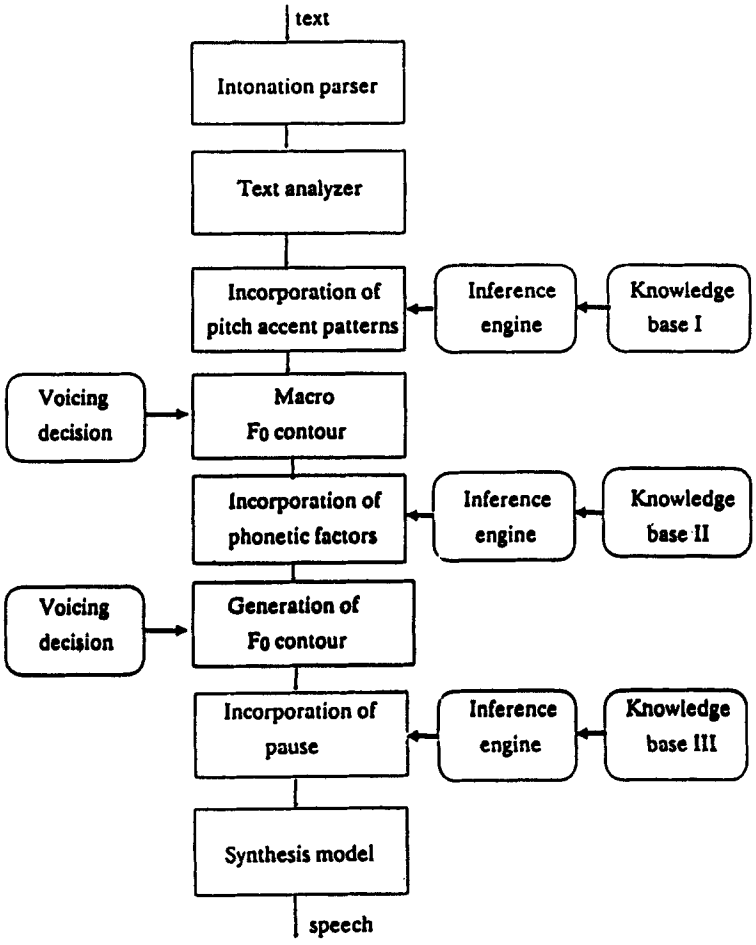


Figure 12. Incorporation of intonation knowledge in a text-to-speech system for Hindi.

parser is used for the classification of sentences. The F_0 contour changes with respect to the type of the sentence. (2) Text analysis has to be performed both at the word level and at the character level. Word-level analysis decides the importance of each word in the sentence. Character analyser determines the number of syllables in each word and classifies the syllables based on their acoustic phonetic behaviour. (3) Fall–rise patterns which include the decision of valleys and peaks have to be incorporated. After the incorporation of valleys and peaks a macro F_0 contour is generated by joining successive valleys and peaks using a straight line based on the voiced/unvoiced classification of the corresponding basic units. (4) Segmental factors on the F_0 contour have to be incorporated. The final F_0 contour is generated at this stage using spline curves (Rogers & Adams 1989). (5) We have to incorporate the proper amounts of pauses between words, intonational phrases and sentences. During a pause, all source and system parameters are set to zero. (6) Intonation knowledge has to be represented using a suitable knowledge representation scheme in order to incorporate in a text-to-speech system. Our system is based on production system approach (Rich 1983). (7) Activation of intonation knowledge is achieved by means of a rule based inference engine with forward chain control strategy. Figure 12 places these issues in the overall scheme of our text-to-speech system.

The quality of synthetic speech obtained after the incorporation of intonation knowledge is tested for several sentences and compared with synthetic speech produced from the waveform concatenation model and the parameter concatenation model without intonation knowledge. The intelligibility and naturalness of synthetic speech increased significantly with the addition of intonation knowledge.

4. Evaluation of the quality of synthetic speech

As the quality of the text-to-speech system improves, it is necessary to evaluate and compare the performance of each rule added to the system. Quality of synthetic speech is usually referred to the total auditory impression the listener experiences upon hearing the speech from the system. The listener's impression is influenced by the various constraints such as familiarity with the language, the inherent limitations of the human information system, the experience and training of the listener, the linguistic structure of the message set and the structure and the quality of the speech signal (Pisoni *et al* 1985; Childers & Ke Wu 1990).

There is no well-formed method for assessing the performance of synthetic speech quality. Assessment of the perceptual response by the human listener investigates the transmission of linguistic information from the speech signal and addresses specific questions such as how accurately synthetic characters and words are recognized, how well the meaning of synthetic utterance is understood and how easy it is to perceive and understand synthetic speech. Given below are discussions of some perceptual experiments done by us.

In order to test the improvement in quality, several sentences in Hindi were synthesized using (1) waveform concatenation model, (2) parameter concatenation model and (3) parameter concatenation model with the addition of different knowledge sources and demonstrated before several native and non-native speakers of Hindi. The results of these experiments are summarized in the following paragraphs.

From our experiments, we found that the speech obtained from parameter concatenation model (type 2) is better than the waveform concatenation model (type 1).

There are abrupt discontinuities in the synthetic speech obtained from type 1 at the boundaries of the basic units. This is removed by type 2 as concatenation is now done at the parameter level resulting in a reasonably smooth transition between boundaries. But there are various distortions in type 2 due to the lack of proper prosodic knowledge in synthetic speech.

The listeners were able to perceive the improvement in the quality of the synthetic speech in type 3, that is, in the parameter concatenation model after incorporating different knowledge sources. Depending on the outcome of the performance of the system, we can modify the system further, to attain the ultimate goal – to develop a text-to-speech system which is as good as natural.

5. Summary

In this paper, we have discussed the importance of knowledge sources in text-to-speech systems. Even though the emphasis was on Hindi, it can be extended to other Indian languages as well since the special features of Indian languages are considered in the design of the system. The main issues in the design and the development of such a system are the acquisition, representation and activation of knowledge at various levels, especially knowledge related to coarticulation, duration and intonation. All these knowledge sources are coded into a suitable form to incorporate into the system. The quality of speech from the text-to-speech system can be improved significantly with the addition of rules in the knowledge base. It is also essential to acquire knowledge related to other sources such as intensity variations and pauses between words and phrases.

References

Allen J 1985 A perspective of man-machine communication by speech. *Proc. IEEE* 73: 1541–1551

Allen J, Hunnicutt M S, Klatt D H 1987 *From text-to-speech: the MITalk system* (Cambridge: University Press)

Atal B S, Hanauer S L 1971 Speech analysis and synthesis by linear prediction of speech wave. *J. Acoust. Soc. Am.* 50: 637–655

Childers D G, Ke Wu 1990 Quality of speech produced by analysis-synthesis. *Speech Commun.* 9: 97–117

Fant G 1982 The voice source – acoustic modeling, Technical report, STL-QPSR 4/1982: 28–48

Klatt D H 1976 Linguistic uses of segmental duration in English: acoustic and perceptual evidences. *J. Acoust. Soc. Am.* 60: 1208–1221

Klatt D H 1980 Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.* 67: 971–995

Madhukumar A S, Rajendran S, Yegnanarayana B 1993 Intonation component of a text-to-speech system for Hindi. *Computer Speech and Language* 7: 283–301

Makhoul J 1975 Linear prediction: a tutorial review. *Proc. IEEE* 63: 561–580

Markel J D 1972 The SIFT algorithm for fundamental frequency estimation. *IEEE Trans. Acoust. Speech Signal Process.* 24: 399–418

Ohman S E G 1966 Coarticulation in *VCV* utterances: spectrographic measurements. *J. Acoust. Soc. Am.* 39: 151–168

O'Shaughnessy D 1984 Design of a real-time French text-to-speech system. *Speech Commun.* 3: 233–243

O'Shaughnessy D 1987 *Speech communication – Human and machine* (Reading, MA: Addison Wesley)

Papamichalis P E 1987 *Practical approaches to speech coding* (Englewood Cliffs, NJ: Prentice Hall)

Pisoni D B, Nusbaum H C, Green B G 1985 Perception of synthetic speech generated by rule. *Proc. IEEE* 73: 1665–1676

Rajesh Kumar S R 1990 *Significance of durational knowledge in a text-to-speech system for Hindi.* M S Dissertation, Indian Institute of Technology, Madras

Ramachandran V R, Yegnanarayana B 1992 Coarticulation rules for a text-to-speech system for Hindi. In *Proceedings of the Speech Technolgy Workshop*, Indian Institute of Technology, Madras, pp. 211–219

Rich E 1983 *Artificial intelligence* (New York: McGraw Hill)

Rogers D F, Adams J A 1989 *Mathematical elements for computer graphics* (New York: McGraw Hill)

Yegnanarayana B, Murthy H A, Ramachandran V R 1991 Speech processing using modified group delay functions. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Toronto, 2: 945–948

Yegnanarayana B, Murthy H A, Sundar R, Alwar N, Ramachandran V R, Madhukumar A S, Rajendran S 1990 Development of a text-to-speech system for Indian languages. In *Frontiers of knowledge based computing systems* (eds) K M Rege, V P Bhatkar (Bombay: Narosa)