Computer Speech and Language (1993) 7, 283–301

Intonation component of a text-to-speech system for Hindi

A. S. Madhukumar, S. Rajendran and B. Yegnanarayana

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

Abstract

In this paper, we describe some features of the fundamental frequency (F_0) contours of speech in Hindi and propose an approach to represent and activate this intonation knowledge for an unrestricted text-to-speech system for Hindi. As in most languages, the declarative sentences in Hindi show a declining pattern of F_0 contour, whereas interrogative sentences show a rising F_0 contour. The backdrop declining or rising pattern is characterized by local falls and rises which are determined by the phonological pattern of the constituent words. In complex declarative sentences the F_0 contour resets at major syntactic boundaries. Experiments to obtain the inherent F_0 of segments are described. The intonation knowledge derived from analysis of speech is coded in a production system format. Intelligibility and naturalness of the synthesized speech improved significantly after incorporation of the intonation rules.

1. Introduction

A text-to-speech system converts an input text into a speech signal. Humans use several knowledge sources such as phonetics, phonology, morphology, syntax, semantics and pragmatics to produce speech from a text. 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 signals corresponding to the basic units of speech does not produce natural sounding speech. Rules that govern prosodic aspects of sentences and discourse have to be incorporated. While building an unrestricted text-to-speech system for Hindi, an Indian language, we have also addressed prosodic aspects (Yegnanarayana, Murthy, Sundar, Alwar, Ramachandran, Madhukumar & Rajendran, 1990). In this paper we discuss acquisition and incorporation of the intonation component of the prosodic knowledge of the system.

Intonation pattern is defined as pitch pattern over time. An utterance may convey different meanings due to changes in intonation, even if it is composed of the same segmental phonemes. In a text-to-speech system intonation refers to the periodicity of the glottal pulse source for voiced speech sounds (Klatt, 1987). One of the functions of intonation is to group words into syntactic blocks for semantic interpretation of the

Correspondence to Prof. B. Yegnanarayana.

0885–2308/93/030283 + 19 $08.00/0

utterance. Emphasis of words in an utterance is signalled by significant fall and rise patterns in the pitch contour, the perceptual correlate of the fundamental frequency (F_0) (Lehiste, 1970). In speech synthesis proper modelling of the pitch pattern leads to natural sounding speech, while in speech recognition it can assist the acoustic–phonetic module to disambiguate alternative sequences of units at various stages (e.g., Lea, 1980; Akers & Lennig, 1985; Waibel, 1988).

This paper is organized as follows: Section 2 describes important properties of intonation patterns in Hindi. Major properties of intonation patterns such as declination/rising tendency of F_0, fall–rise patterns, resetting of F_0 and inherent F_0 are discussed here. Section 3 is concerned with a text-to-speech system for Hindi. Here we discuss some of the issues involved in the design of a text-to-speech system. Finally section 4 deals with representation and activation of intonation knowledge for the text-to-speech system.

2. Properties of intonation patterns

There is no systematic study available on the properties of intonation patterns for continuous speech in Hindi, particularly in the context of the development of speech systems for Hindi. The properties discussed in the following sections are based on our observations on a speech database of *read sentences*. A corpus of 500 sentences was read out by two adult male speakers of Hindi. The speech data 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 algorithm for pitch extraction was based on group delay functions (Yegnanarayana *et al.*, 1990). In the following sections major properties of the intonation patterns, namely, declination/rising tendency of F_0, fall–rise patterns, resetting of F_0 and inherent F_0 are discussed.

2.1. Declination/rising tendency of F_0 contour

The F_0 pattern of a declarative sentence has a tendency to decline gradually during the utterance (Cohen & t'Hart, 1967). This is reported to be a common feature in languages like English (O'Shaughnessy, 1976; Pierrehumbert, 1981), French (Delgutte, 1978), Japanese (Fujisaki & Hirose, 1985), German (Ladd, Silverman, Tolkmitt, Bergmann & Scherer, 1985), Dutch (Cohen & t'Hart, 1967) and Danish (Thorsen, 1980). The attributes of declination can be divided into global and local. The global attributes are those characteristics of the fundamental frequency which can be defined over an entire clause or a phrase. Declination tendency and rising tendency are the salient global attributes in declarative sentences and interrogative sentences, respectively. Local attributes represent those characteristics which include no more than two adjacent words. Fall–rise patterns of F_0 contour and the effects of lexical stress, among other things, are local attributes (Cooper & Sorensen, 1981).

Properties of F_0 declination can be summarized as follows: (1) F_0 values fluctuate between two abstract lines—a top line and a base line, drawn near or through all maxima and minima of F_0 values in a sentence, respectively. (2) There will be a repeated succession of F_0 falls (valleys) and rises (peaks). (3) Range of F_0 (difference between valley and peak in a word) decreases with time. That is, both the top line and the base line monotonically decrease and the slope of the top line is steeper than the base line (Vaissière, 1983). (4) In a neutral declarative sentence the maximum value of F_0 will be

located on the stressed syllable of the first *content word* (semantically meaningful word) itself. (5) In connected speech the *content word* together with the preceding or following monosyllabic *function words* (words which have only grammatical meaning), if any, form a pitch accent group called *prosodic word* under certain conditions. This will be determined by rhythmic factors and other linguistic constraints. For example, the class of *function words* in Hindi includes case markers, postpositions, complementers, negative markers, conjunction, relative pronouns, emphatic markers, etc. Some of the monosyllabic *function words* (e.g., postposition */ko:/*, 'to') may get accented in continuous speech, and therefore prosodically they conjoin with the immediately preceding *content word.* In such cases the peaks of the *content word* will get shifted to the function word—this conjugation of *content word* and *function word* is treated as a *prosodic word.*

Speech waveform and the corresponding F_0 contour for a natural utterance of a simple declarative sentence are shown in Figure 1a. The F_0 contour that starts at the initial syllable of the first word rises towards the next target (peak), that is, the final syllable of the first content word. Further, the F_0 falls towards the initial syllable of the next content word and again rises towards the final syllable of the same word. Again it falls towards some point in the final syllable and rises towards some peak within the same syllable (word). The F_0 damps off towards the end of the utterance. It is possible to draw a line connecting all the peaks (top line) and another line connecting all the valleys (base line). Both lines decline monotonically and converge towards the end.

The intonation pattern in Hindi is not similar for all interrogative sentences. Questions expecting yes/no answers have a continuous rise in the F_0 contour. That is, both the top line and the base line diverge from each other. But in the case of question–word type questions, the top line and base line decline up to the question–word (e.g., */kya:/* 'what') and then rise. Even though the global properties of F_0 contour are different for questions, the local attributes such as the fall–rise pattern remain the same. Figure 2a shows the speech waveform and F_0 contour for a *yes/no* type question where the rising pattern of F_0 can be seen clearly. Figure 3a shows the speech waveform and F_0 contour for a *question–word* type interrogative sentence. Here the question word is */kahã;/* 'where' (second word in the sentence). The F_0 pattern declines up to the question word and then rises.

2.2. Prediction of intermediate peaks and valleys

In this section we describe our experiments to predict the intermediate peaks and valleys. We have selected 10 simple declarative sentences uttered by two adult male speakers. In all the sentences the number of content words are the same. The maximum and minimum F_0 values of each content word fluctuate between the top line and the base line. The results of our analysis show that we can approximately predict the intermediate peaks and valleys. If P_0 is the initial peak and P_n is the final peak at timings T_0 and T_n, respectively, then an intermediate peak at time T_1 can be expressed as

$$P_1 = P_0 + (T_1 - T_0)*(P_0 - P_n)/(T_0 - T_n).$$

Since the aim was to capture the properties of F_0 contour by some mathematical formula, the individual differences among the phonetic segments, speakers and sentences were normalized. For example, the inherent F_0 of the vowels changes with the phonetic properties as well as the properties of adjacent consonants. Variation in F_0 due to phonetic properties causes errors in the prediction of intermediate peaks and valleys.

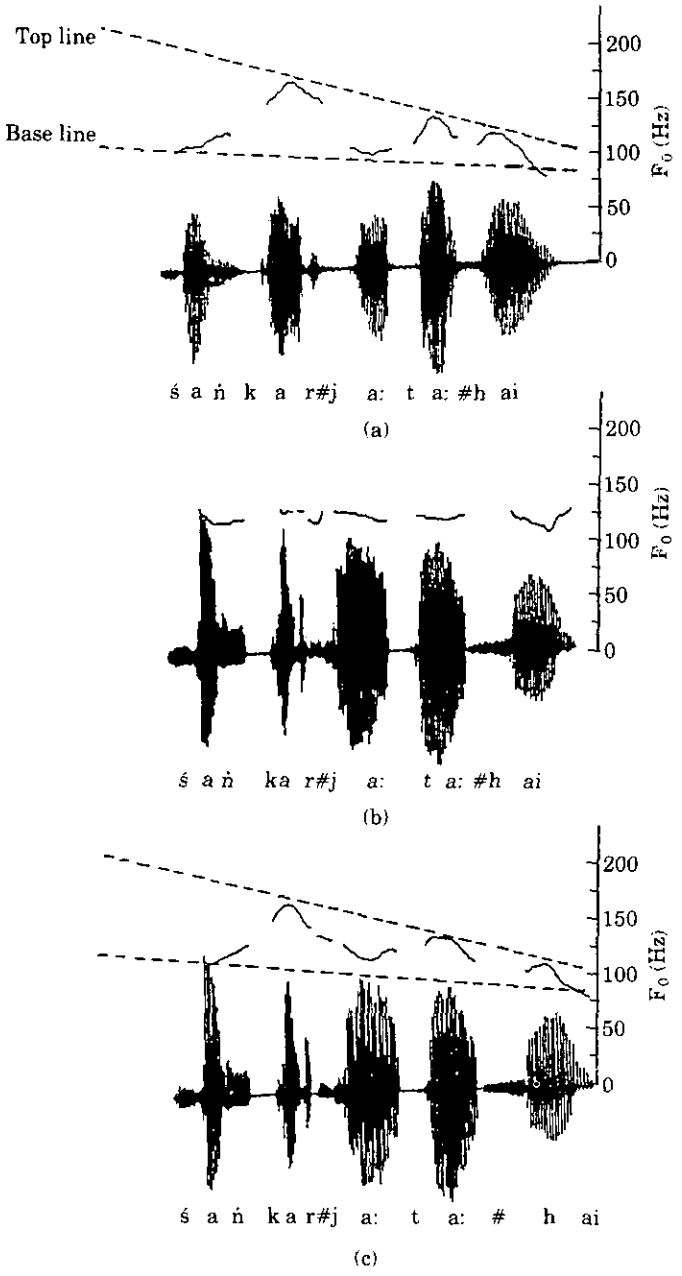


Figure 1. Speech waveform and F_0 contour for a simple declarative sentence:
/śankar#ja:ta:#hai/
Shankar go is
'Shankar goes' (# indicates word boundary).
(a) Natural speech signal; (b) Synthesized speech signal without applying intonational rules; (c) Synthesized speech signal after applying intonational rules.

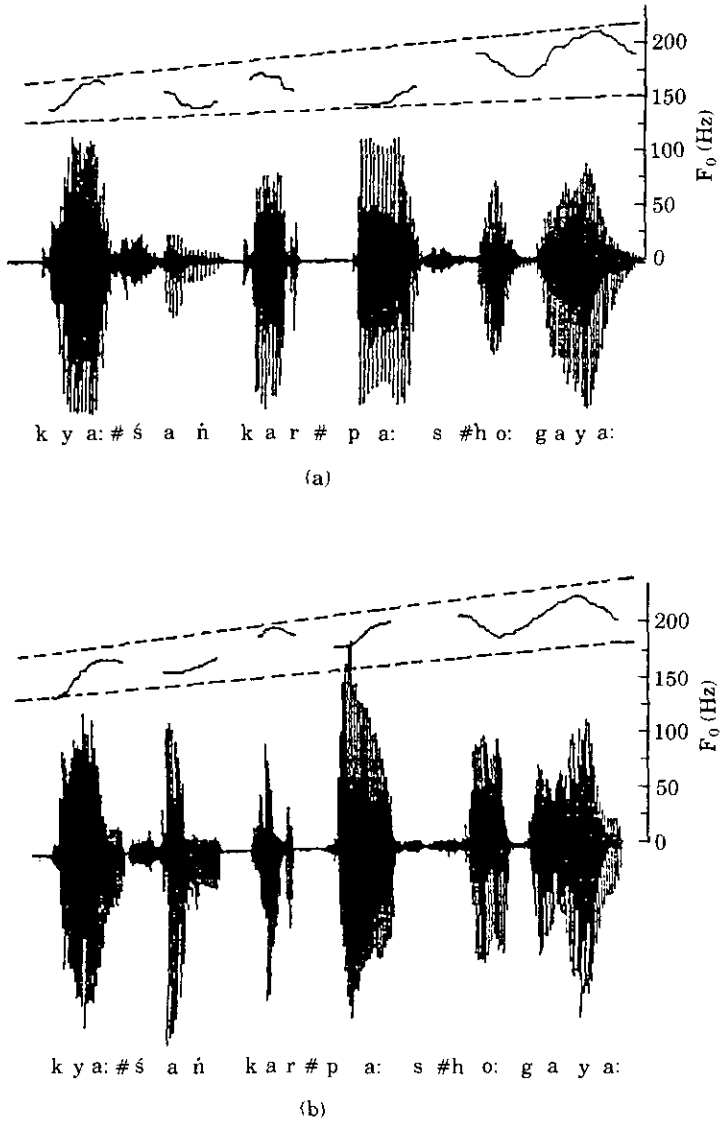
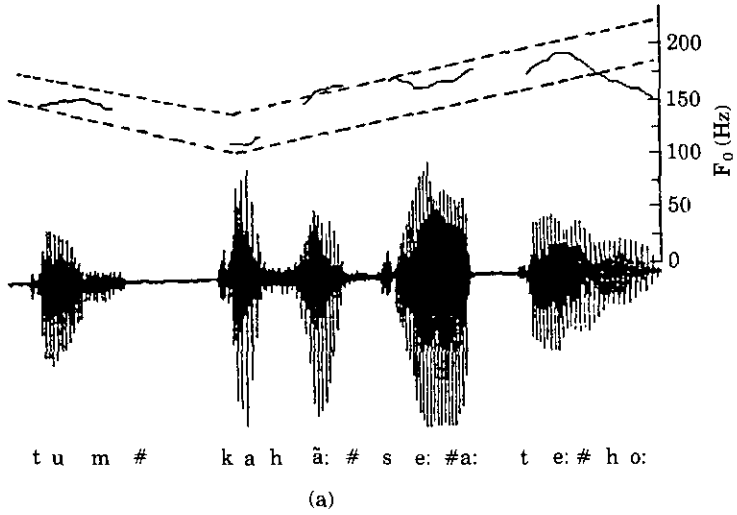


Figure 2. Speech waveform and F_0 contour for an *yes/no* type interrogative sentence:

/kya:#śankar#pa:s#ho:gaya:/
what Shankar pass become
'Has Shankar passed' (# indicates word boundary).
(a) Natural utterance; (b) Synthesized speech.

Table I shows the actual intermediate peaks, predicted peaks and the percentage errors for each sentence for a single speaker. In our experiments, we noted that the error is always less than 10% of the actual intermediate peak value, keeping the range behaviour the same.

An experiment similar to the above was conducted for interrogative sentences also. The intermediate peaks can be predicted as in the case of declarative sentences. Table II shows the prediction statistics for intermediate F_0 peaks in yes/no type questions. In



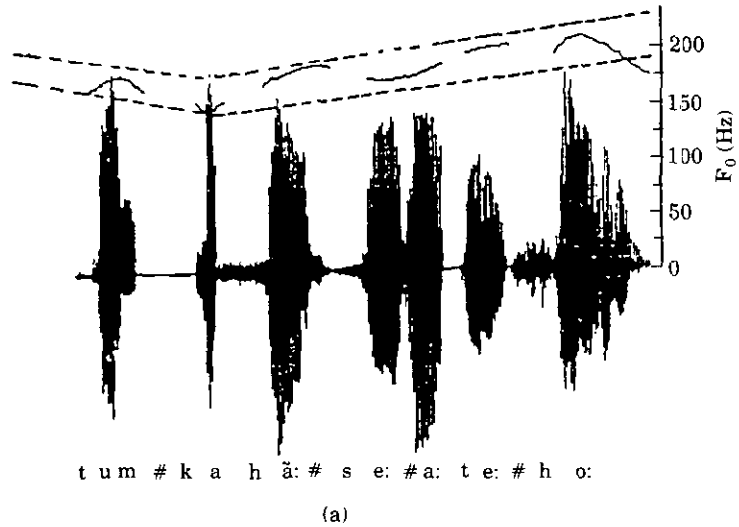


Figure 3. Speech waveform and F_0 contour for a question–word type interrogative sentence:

/tum# kahã:#se:#a:te:#ho:/
you where from come
'Where do you come from'. (# indicates word boundary).
(a) Natural utterance; (b) Synthesized speech.

interrogative sentences we have observed that the accuracy in the prediction of valleys was less compared to that of peaks.

2.3. Range of F_0 for prosodic words in sentences

In declarative sentences both the top line and the base line have negative slopes and these converge towards the end. As a result, frequency range between the valley and the

TABLE I. Prediction of intermediate peaks in declarative sentences

Sentence	Peak 1 Actual	Peak 1 Predicted	Peak 1 %Error	Peak 2 Actual	Peak 2 Predicted	Peak 2 %Error
1	127·63	129·14	1·19	123·69	123·64	0·04
2	132·88	135·56	2·02	117·13	124·88	6·62
3	147·37	141·42	4·04	125·00	132·56	6·04
4	135·44	138·32	2·13	121·07	128·41	6·06
5	134·19	137·64	2·57	127·63	127·69	0·05
6	132·88	136·84	2·98	121·07	123·78	2·24
7	132·88	135·69	2·12	121·07	125·00	3·25
8	142·06	135·04	4·94	118·44	127·89	7·98
9	144·69	152·50	5·40	135·50	140·25	3·51
10	157·81	154·29	2·23	144·69	139·33	3·70

TABLE II. Prediction of intermediate peaks in *yes/no* type questions

Sentence	Peak 1 Actual	Peak 1 Predicted	Peak 1 %Error	Peak 2 Actual	Peak 2 Predicted	Peak 2 %Error
1	176·18	175·82	0·20	226·04	216·16	4·37
2	198·49	188·37	5·10	186·68	202·97	8·73
3	199·80	205·87	3·04	237·85	228·63	3·88
4	199·80	203·34	1·77	239·42	242·38	1·24
5	202·42	198·97	1·71	237·22	217·46	8·33
6	203·73	193·68	4·94	190·17	204·83	7·71
7	201·12	216·23	7·51	253·55	252·39	0·46
8	207·67	189·32	8·84	215·54	194·13	9·93
9	220·79	213·49	3·31	250·97	237·42	5·40
10	193·24	199·05	3·01	236·54	223·84	5·37

following peak decreases with time. Table III shows the F_0 range for each prosodic word in 10 simple declarative sentences. The Table also shows the mean and standard deviation (SD). All sentences have an equal number of prosodic words. From Table III it is obvious that the range decreases with respect to the position of the prosodic word.

We have conducted similar experiments for *yes/no* type interrogative sentences. Here the top line and base line diverge and hence the F_0 range of prosodic words increases with the position. In contrast with the observations for declarative sentences, here the final prosodic word has the maximum range and the first word has the minimum. Results of these observations are shown in Table IV. When sentences end with monosyllabic words (e.g., */hai/* 'is') the range would be less due to the tapering effect, and hence the SD (27·02) is very high for the range of final word.

2.4. Fall–rise patterns

By analysing large amounts of data, we have observed some general features of local falls and rises of F_0 which are determined by the phonological constituents of the words. The following are some observations for Hindi sentences: (1) F_0 contour of prosodic words exhibits a regular pattern of a valley preceding each peak. (2) The valleys and peaks are mostly associated with the vowels which are the nuclei of the syllables.

TABLE III. Range of F_0 (in Hz) for fall–rise patterns in declarative sentences

	Range of F_0			
Sentence	Word 1	Word 2	Word 3	Word 4
1	38·05	22·51	13·12	11·81
2	49·86	38·12	22·31	18·37
3	45·93	26·18	16·99	14·44
4	39·36	23·62	26·24	19·68
5	40·68	18·37	18·37	10·50
6	31·49	20·99	19·68	11·81
7	35·43	22·31	7·87	3·94
8	48·55	27·56	30·18	21·03
9	61·67	57·73	44·62	26·24
10	65·61	38·05	24·93	6·56
Mean	45·66	29·54	22·43	14·44
SD	10·53	11·36	9·62	6·54

TABLE IV. Range of F_0 for fall–rise patterns in *yes/no* type interrogative sentences

	Range of F_0			
Sentence	Word 1	Word 2	Word 3	Word 4
1	13·12	35·43	89·23	57·06
2	19·68	43·30	94·48	100·27
3	30·18	41·99	74·79	51·98
4	10·50	24·93	59·97	107·85
5	10·50	77·41	80·04	86·56
6	27·56	35·43	40·24	77·42
7	18·37	28·87	85·24	93·12
8	30·18	59·11	81·36	35·43
9	38·05	64·30	56·42	21·00
10	28·87	40·68	66·92	70·87
Mean	22·70	45·15	72·87	70·16
SD	9·09	15·81	15·97	27·02

However, the exact target point (peak or valley) within the voiced regions of the syllable is determined by several other factors. For example, the peak of the nucleus gets shifted to the *coda* (the consonant that follows the vowel nucleus of the syllable) if the consonant is a nasal or lateral. (3) If the word is monosyllabic then the valley and the peak occur within the same syllable and hence F_0 will rise steadily. (4) In the cases of disyllabic and trisyllabic words the peak occurs within the region of the final syllable and the valley occurs on the initial syllable. (5) Tetrasyllabic words are characterized by two types of patterns, the F_0 rises from the initial syllable and culminates on the final syllable

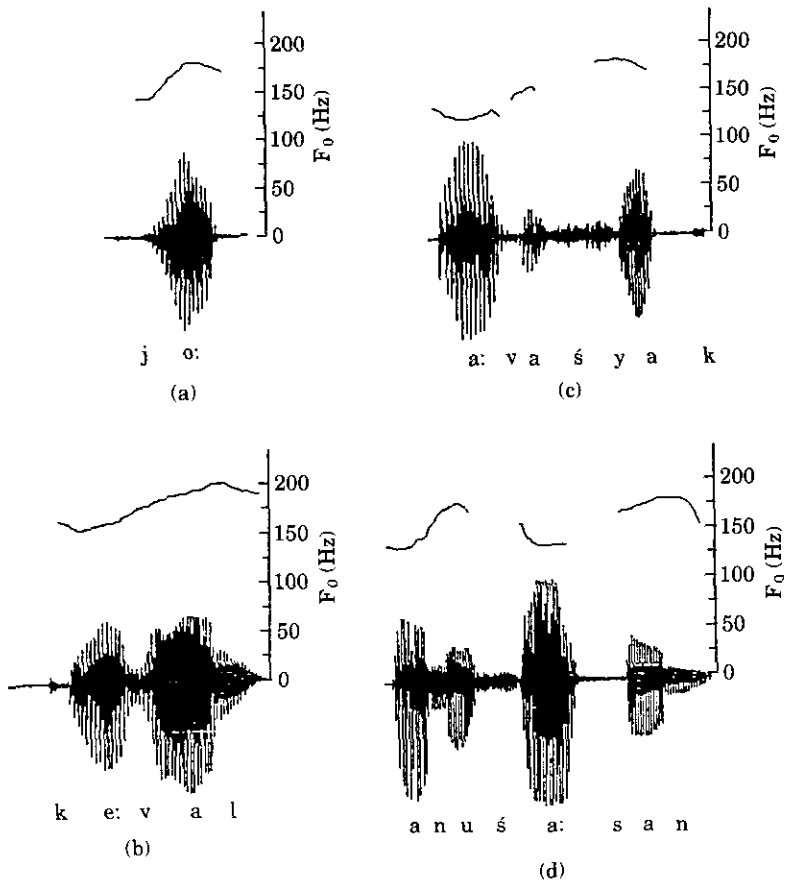


Figure 4. Local fall–rise patterns for (a) monosyllabic word */jo:/* 'which'; (b) disyllabic word */ke:val/* 'absolute'; (c) trisyllabic word */a:vaśyak/* 'necessary'; (d) tetrasyllabic word */anusa:san/* 'discipline'.

of the word, or the F_0 valleys and peaks occur at alternate syllables, that is, the peaks occur on the second and the final syllable. (6) In the case of the words with pentasyllables and beyond, the pattern will be similar to a combination of disyllabic and trisyllabic words. These fall–rise patterns are useful in synthesizing intonation (Madhukumar, Rajendran & Yegnanarayana, 1991). Figure 4 shows the local fall–rise patterns for (a) monosyllabic word, (b) disyllabic word, (c) trisyllabic word and (d) tetrasyllabic word.

Fall–rise patterns are superimposed on each prosodic word, whereas declination and rising manifest throughout the utterance. Hence operationally they are distinguishable by virtue of their different domains. F_0 contours in Figs 1a, 2a and 3a show the local fall–rise pattern at prosodic word level in addition to their global attributes.

The order of the noun phrase arguments that precede the verb in Hindi can be varied quite freely. However, the fall–rise patterns of words do not undergo any change when the order of words is modified (Madhukumar, Rajendran, Chandra Sekhar & Yegnanarayana, 1991). For example, consider the following sentence:

/usta:d ne: mi:na: ko: sita:r sikha:ya:/
Master NOM. Mina to sitar taught

'The master taught sitar to Mina'

where NOM. refers to the nominative case marker */ne:/*.

The sentence can be written in at least six different ways by changing the order of the subject noun phrase (NP) (*/usta:d ne:/*), direct object NP (*/mi:na: ko:/*), indirect object NP (*/sita:r/*) and the verb phrase (*/sikha:ya:/*) in the sentence. The fall–rise pattern of the words remains constant irrespective of the change in the word order. However, the range between the valley and peak will vary, and this is determined by the position of a word in the sentence. When a word occurs in the initial position of a declarative sentence, the range between the valley and the peak is around 30 to 40 Hz, whereas when the same word occurs in the final position the range is not more than 10 Hz. Figure 5 shows the constant pitch accent pattern of the same prosodic word (*/mi:na: ko:/*) with differing pitch ranges when the word occurs in different positions in a sentence.

2.5. Resetting of F_0 contour

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

Resetting of F_0 will take place both in valleys and in peaks. But the magnitude of resetting differs in both cases. From our experiments on *read sentences* of complex declarative sentences, we have observed certain features related to the resetting across syntactic boundaries. For these studies we have used a corpus of 50 sentences, each with two syntactic clauses. The general properties of F_0 resetting obtained from this analysis are summarized below.

The initial peak (the first peak of the first syntactic clause) F_0 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. The effect of resetting is directly proportional to the strength of the syntactic boundary (Cooper & Paccia-Cooper, 1980). There is no significant resetting between short phrases. Within a syntactic clause, the F_0 contour is similar to the F_0 pattern of a simple sentence. That is, the declination in F_0 contour is accompanied by local falls and rises.

There is a correlation between pause between syntactic clauses and resetting value. The duration of the pause can be considered as a cue to measure the strength of the boundary. That is, the stronger the boundary, the longer the pause. When the pause is long, subglottal air pressure is built up again and hence the value of F_0 for resetting will be high. The resetting value of F_0 increases with the pause up to a limit. Beyond the limit the pause does not show any correlation with the resetting value. In such cases each syntactic clause can be considered as an independent sentence. From our experiments, we have observed that the pause between syntactic clauses varies from 3% to 18% of the total duration of the utterance.

Figure 6a shows the effect of resetting the F_0 contour across syntactic boundaries. The sentence has two clauses, and hence one major syntactic boundary. The resetting of F_0 coincides with this syntactic boundary.

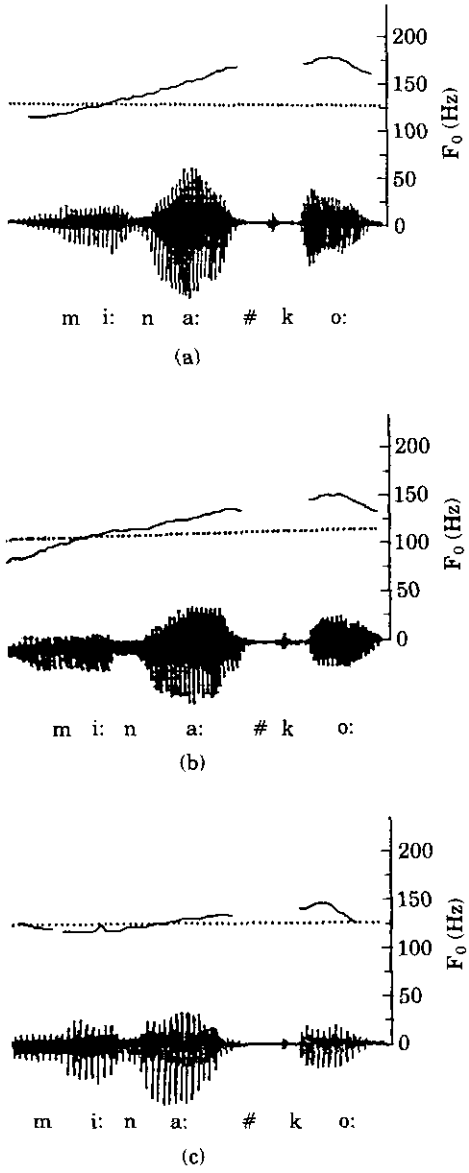


Figure 5. Fall–rise patterns of the prosodic word (*/mi:na: ko:/*) when it occurs in the following three different positions in a sentence:

(a) */mi:na: ko: sita:r usta:d ne: sikha:ya:/*
Mina to sitar master NOM. taught
'The master taught sitar to Mina'.
(b) */sita:r mi:na: ko: usta:d ne: sikha:ya:/*
(c) */usta:d ne: sita:r mi:na: ko: sikha:ya:/*

For studying the behaviour of the intonation patterns in complex declarative sentences we have selected 15 complex declarative sentences in Hindi with two syntactic clauses. Syntactic clauses can be identified from a text by the presence of function words like subordinate and coordinate conjunctions in Hindi (Kachru, 1980).

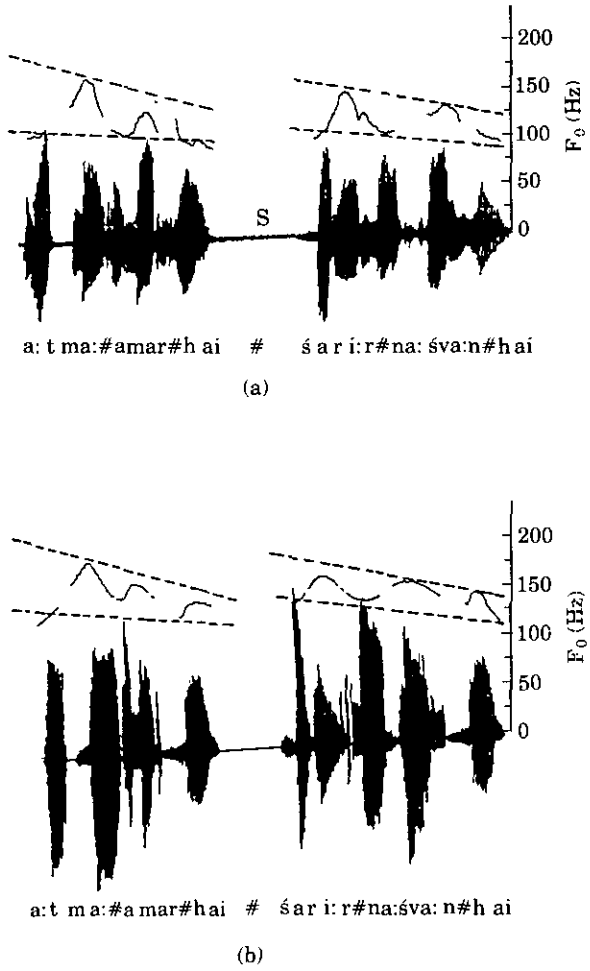


Figure 6. Speech waveform and resetting of F_0 contour at a clause boundary (S) in a complex declarative sentence:
/a:tma: #amar #hai #śari:r #na:śva:n# hai/
Soul immortal is body mortal is
'Soul is immortal, body is mortal'. (# indicates word boundary).
(a) Natural speech signal; (b) Synthetic speech signal.

The average initial peak F_0 (the first peak of the first syntactic clause) is around 130 Hz for an adult male speaker and is speaker dependent. All other significant peaks such as the intermediate and the final peaks of syntactic clauses, resetting value of F_0 (the first peak of the second syntactic clause) and tapering frequency at the end of the utterance can be related to the initial peak frequency. The final peak of the first syntactic clause is about 81% (ranging from 66% to 90% with a SD of 6·49) of the initial frequency. Resetting frequency is around 92% (ranging from 88% to 97% with a SD of 2·60) and the final peak of the second syntactic clause is around 72% (ranging from 61% to 84% with a SD of 6·88) of the initial peak frequency. End tapering frequency is always constant, that is, around 57% of the initial peak frequency. Table V shows the experimental results for the resetting of peaks.

TABLE V. Resetting values of F_0 peaks at syntactic boundaries in complex declarative sentences

Sentence	%dur 1	%pause	$\%F_{01}$	$\%F_{0r}$	$\%F_{0f}$	%ta
1	43·73	17·22	82·66	96·82	65·32	53·76
2	45·37	11·65	88·62	89·85	84·31	57·85
3	44·69	11·66	76·45	89·24	62·50	55·23
4	49·54	14·86	70·29	90·88	61·18	54·41
5	44·66	14·52	83·28	90·62	79·47	55·13
6	36·18	15·33	66·18	94·75	68·80	54·23
7	45·51	11·58	80·16	95·38	74·46	58·97
8	42·82	11·15	90·33	90·33	80·66	58·29
9	55·84	12·78	80·89	87·81	62·88	56·23
10	45·35	12·00	83·84	90·96	77·26	60·00
11	42·46	11·08	90·26	95·42	69·63	57·59
12	34·78	11·75	86·14	90·22	72·01	57·34
13	56·67	11·94	78·90	90·41	76·71	59·18
14	49·23	10·87	82·32	92·88	75·99	58·05
15	46·91	13·52	82·07	89·65	74·75	56·31
Mean	45·58	12·79	81·49	91·68	72·40	56·84
SD	5·70	1·82	6·49	2·60	6·88	1·89

%dur1 and %pause are the ratio of duration of the first syntactic clause and duration of the pause with respect to the total duration of the utterance, respectively. $\%F_{01}$ is the ratio of the final peak frequency of the first syntactic clause to the initial peak frequency of the first syntactic clause. $\%F_{0r}$ and $\%F_{0f}$ are the ratios of the first and final peak frequencies of the second syntactic clause, respectively, with respect to the initial peak frequency of the first syntactic clause. Similarly, %ta is the ratio of the end tapering frequency with respect to the initial peak frequency.

Like peaks, the valleys in a complex declarative sentence also show a systematic behaviour. We have analysed all significant valley points with respect to the initial peak frequency. After resetting, the values of the valleys also change similar to the values of the peaks. From the analysis we have observed that generally the initial valley (the valley of the first prosodic word in the first syntactic clause) stands out separately from the base line. Due to this and the errors in the prediction of valleys and peaks, the proposed model can be regarded as an approximation of the properties of intonation patterns in Hindi. The value of the initial valley is around 73% (ranging from 80% to 67% with a SD of 3·63) of the initial peak frequency. The base line of the first syntactic clause is the line joining the second valley (valley of the second prosodic word in the first syntactic clause) and the final valley of the first syntactic clause. In the first syntactic clause, the second valley is around 75% (ranging from 85% to 68% with a SD of 5·40) and the final valley is 64% (ranging from 71% to 59% with a SD of 3·72) of the initial peak frequency. The base line of the second syntactic clause is the line joining the resetting valley and the final valley of the second syntactic clause. The resetting valley is around 71% (ranging from 79% to 67% with a SD of 3·23) and the final valley of the second syntactic clause is around 65% (ranging from 75% to 60% with a SD of 4·34) of the initial peak frequency. Table VI gives the results obtained from the analysis of valleys in complex declarative sentences.

In all these cases we have considered the initial peak frequency as the reference point.

TABLE VI. Resetting values of F_0 valleys at syntactic boundaries in complex declarative sentences

Sentence	%iv	$\%V_{0s}$	$\%V_{0i}$	$-V_{0r}$	$\%V_{0f}$
1	76·59	69·36	65·32	67·05	60·40
2	70·75	73·73	69·85	70·45	63·88
3	72·29	80·42	70·78	70·18	64·16
4	71·43	80·47	65·01	69·68	62·39
5	69·84	69·29	59·24	67·39	62·50
6	66·85	71·47	63·04	72·55	61·96
7	70·40	72·27	58·67	69·07	65·07
8	67·60	80·45	64·53	73·46	73·74
9	71·92	84·81	68·48	76·50	74·79
10	74·24	68·42	60·11	71·47	68·98
11	72·73	82·39	65·34	74·43	64·77
12	77·94	73·07	69·34	79·08	68·19
13	80·43	73·37	61·68	69·84	61·41
14	74·25	68·77	63·01	69·59	62·85
15	75·99	69·92	62·27	69·13	60·42
Mean	72·88	74·55	64·44	71·32	65·03
SD	3·63	5·40	3·72	3·23	4·34

%iv, $\%V_{0s}$ and $\%V_{0i}$ are the ratios of the initial, second and the final valleys of the first syntactic clause, respectively, with respect to the initial peak frequency. $\%V_{0r}$ and $\%V_{0f}$ are the ratios of the initial and the final valley of the second syntactic clause, respectively, with respect to the initial peak frequency.

Hence the accuracy of modelling F_0 resetting for a text-to-speech system completely depends on the choice of the initial peak frequency.

2.6. Inherent F_0

Studies on English and other languages suggests that there is a definite correlation between inherent F_0 and the height of the vowel and other segmental factors (Lehiste & Peterson, 1959). One of the aims of the present study was to make a systematic study of the properties of inherent F_0 and contextual effects in Hindi. This was studied by embedding the test words in a carrier sentence */me:ra: na:m* _____ *hai/* 'My name is _____'. For this study, we have selected several combinations of disyllabic words. The test words are mostly nonsense words wherein the vowel characteristics are studied both in the initial and final syllables separately. The observations from this study are as follows:

There is a correlation between the height of the vowel and its inherent F_0. If other factors are constant, high vowels */i,u/* exhibit higher F_0 than that of low vowel */a/*. Our study shows that in Hindi the difference between high vowels and low vowels is 15 to 20 Hz. In our experiments we have noticed that the inherent F_0 of mid vowels */e:,o:/* were closer to the F_0 of high vowels than of low vowels. The difference is 2–7 Hz from high vowels. The length of the vowel has a definite correlation with F_0 contours: the longer the vowel the higher the F_0. The prevocalic consonant has an impact on the vowel. When a voiceless consonant occurs in an accented syllable, the peak of the F_0 contour is slightly shifted towards the vowel onset position. The peak of the F_0 contour occurs at the onset

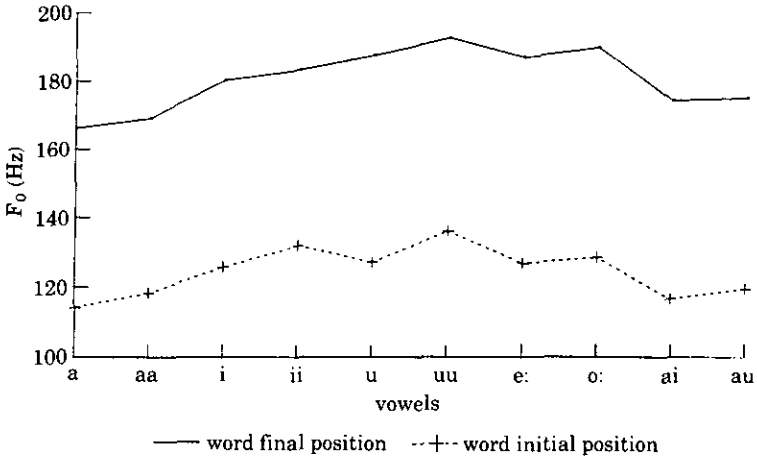


Figure 7. Inherent F_0 of vowels.

of the vowel and this is about 4–10 Hz higher than the F_0 in the middle of the vowel. However, in the case of voiced consonants the F_0 contour rises gradually from the onset of the vowel and the peak occurs in the middle of the vowel nucleus.

Figure 7 shows the inherent F_0 of each vowel for both word initial and final positions. In all cases the word final vowel has a greater F_0 than the initial vowel.

3. Text-to-speech system for Hindi

We are developing a text-to-speech system for Hindi based on a parameter concatenation model (Yegnanarayana *et al.*, 1990). Speech data for all the basic units are stored using parameters such as linear predictive coefficients (LPC), formants with their band widths, pitch and gain. Since speech has been modelled using parameters, voice characteristics can be manipulated and prosodic features can be incorporated by changing these parameters. This representation is highly flexible and needs much less storage compared to the waveform concatenation model.

The input to the text-to-speech system is stored in the form of ISCII (Indian Standard Code for Information Interchange) codes. The preprocessor scans the string of ISCII codes to locate abbreviations, numbers, dates and special symbols and replaces them by their expansion for spoken form. Basic units are extracted from the expanded text using a simple parser. The parser takes care of some language specific rules like word final short vowel deletion. For synthesizing speech, parameters of the basic units in the input text are concatenated using coarticulation rules that operate across adjacent basic units of speech. Studies are being made to obtain the coarticulation knowledge in the form of transition patterns of formants (Ramachandran, submitted). The pitch and gain contours are smoothed at the boundaries between the basic units. The pitch contour is modified to incorporate intonation rules 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 (LPCs and formants) are used to generate the speech waveform. Studies have been made to formulate rules for the change of duration of the basic units in different contexts (Rajesh Kumar, 1990). The quality of the

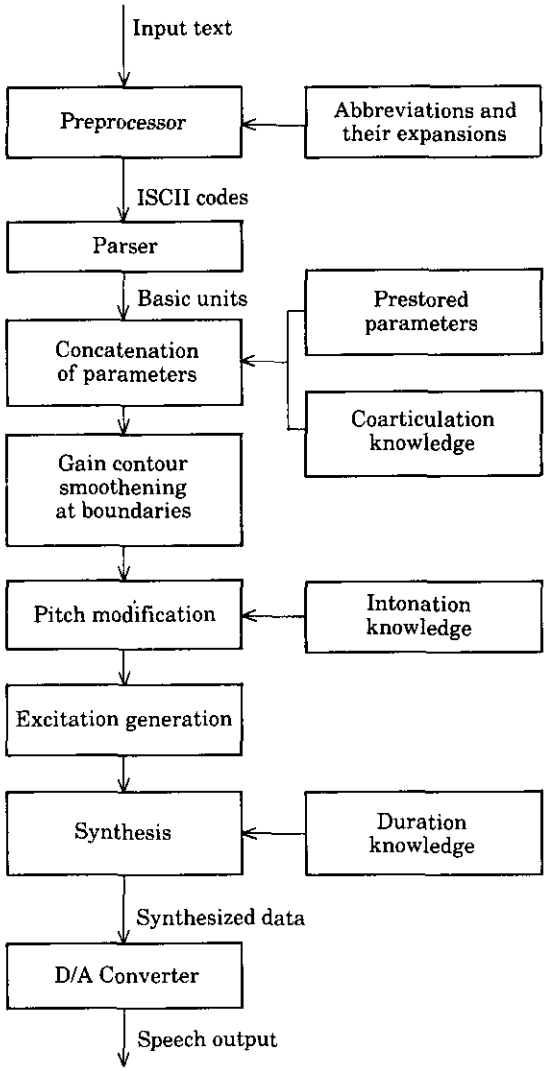


Figure 8. Block diagram of a text-to-speech system for Hindi.

synthesized speech improves significantly after incorporating the prosodic features. Figure 8 shows the block diagram of the current text-to-speech system for Hindi.

4. Representation and activation of knowledge

The intonation knowledge obtained from the analysis of natural speech has to be coded into a suitable form in order to incorporate it in a text-to-speech system. Our system is based on a production system approach. Here knowledge is represented using IF–THEN rules. 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 mostly independent of each other and the order of declaration of rules is not important.

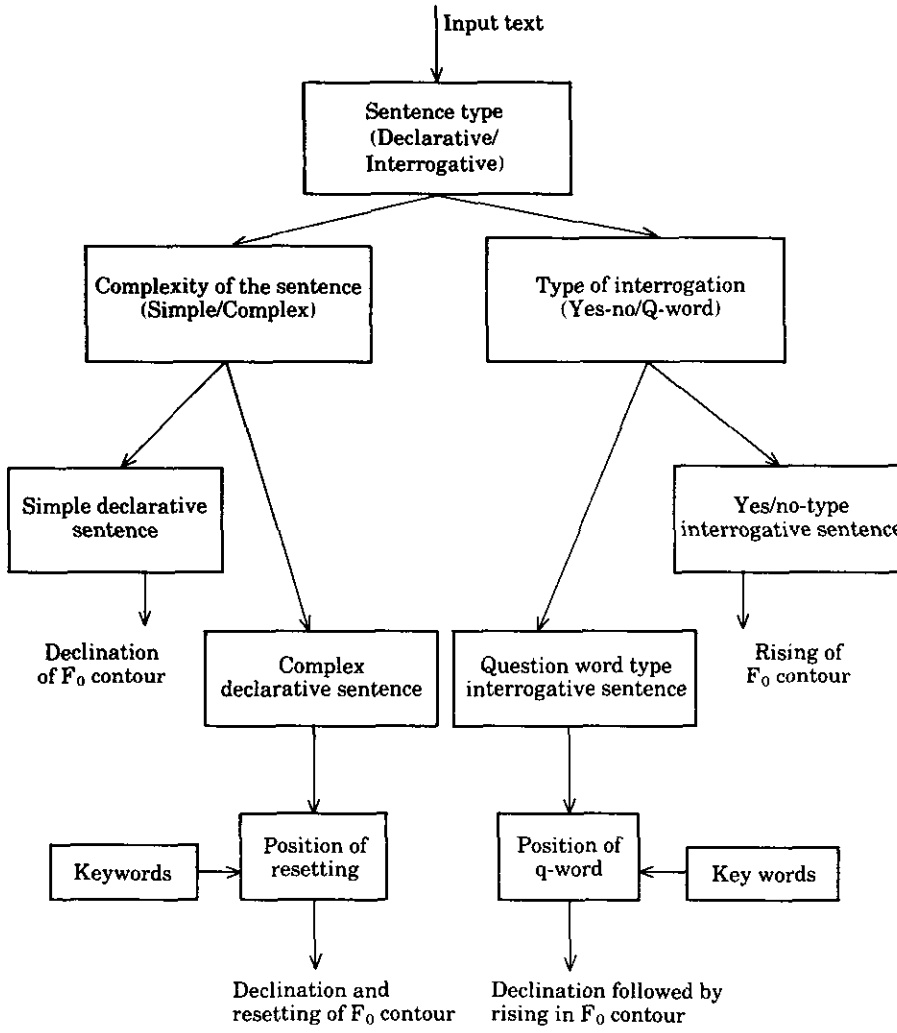


Figure 9. Block diagram of an intonation parser to determine the types of sentences and to assign appropriate F_0 contour.

Depending upon the type of sentence to synthesize, the global properties of the F_0 pattern will get modified. For example, F_0 patterns of simple declarative sentences show declining tendency, while *yes/no* type interrogative sentences show a continuous rise. In complex declarative sentences and in *question word* type questions F_0 pattern exhibits a dual nature. The local attributes remain unchanged in all these cases. So in order to activate this knowledge accurately, we need an intonation parser to determine the type of sentences and the position of pattern change if any. Figure 9 shows the block diagram of the intonation parser. The activation of the knowledge is achieved by means of a rule-based inference engine with a forward chaining control strategy. The valleys and peaks of all the prosodic words can be determined by the inference engine. In order to get a smooth pitch contour for synthesis, the valleys and peaks are joined in the voiced portions by spline curves (deBoor, 1978). The voiced/unvoiced decision is obtained from

the prestored parameters of the basic units. The acoustic–phonetics of characters and the inherent F_0 of vowels will introduce small variations in the F_0 contour at each character level. These variations can also be accounted for by a set of rules.

In order to test the improvement in perceptual quality, several sentences of different types were synthesized with and without the incorporation of intonation rules. Evaluation was based on subjective measurements of native and non-native speakers of Hindi. All listeners agreed on the point that the sentences synthesized with intonation rules sounded more natural and intelligible than the sentences synthesized without intonation knowledge. Figure 1b shows the synthesized speech and the corresponding unmodified F_0 contour for the sentence used in Fig. 1a. After activating the intonation knowledge, the F_0 contour gets modified as shown in Fig. 1c. Similarly, Figs 2b, 3b and 6b show the synthesized speech and the F_0 contours for the sentences used in Figs 2a, 3a and 6a respectively.

5. Conclusion

In this paper we have discussed the acquisition and incorporation of intonation knowledge in a text-to-speech system for Hindi. Declarative sentences are characterized by the declining tendency of F_0 contour, whereas interrogative sentences are characterized by the rising F_0 contour when it is *yes/no* type, and declining followed by rising when the sentence is of *question–word* type. The pitch accent rules were assigned on the basis of the phonological patterns of words in the input text. Syntactic boundaries are characterized by pause and resetting of F_0. The value of resetting frequency depends on the strength of the syntactic boundary. Behaviour of the inherent F_0 of vowels was also studied. Based on these observations rules were framed and incorporated in a production system format. The speech output is intelligible and more natural after incorporating the intonation rules.

References

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

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

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

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

deBoor, C. (1978). *A Practical Guide to Splines*, Springer-Verlag, New York.

Delgutte, B. (1978). Technique for the perceptual investigation of F_0 contours with application to French. *Journal of the Acoustical Society of America*, **64**, 1319–1332.

Fujisaki, H. & Hirose, K. (1985). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan*, **(E)5**, 233–242.

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

Klatt, D. H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, **82**, 737–793.

Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G. & Scherer, K. R. (1985). Evidence for independent functions of intonation contour type, voice quality, and F_0 range in signaling speaker affect. *Journal of the Acoustical Society of America*, **78**, 435–444.

Lea, W. A. (1980). Prosodic aids to speech recognition. In *Trends in Speech Recognition* (Lea, W. A., ed.). Prentice-Hall, New Jersey, 166–205.

Lehiste, I. (1970). *Suprasegmentals*, M.I.T. Press, Massachusetts.

Lehiste, I. & Peterson, G. E. (1959). Vowel amplitude and phonemic stress in American English. *Journal of the Acoustical Society of America*, **31**, 428–435.

Lieberman, P. (1967). *Intonation, Perception and Language*, M.I.T. Press, Massachusetts.

Madhukumar, A. S., Rajendran, S. & Yegnanarayana, B. (1991). Significance of prosodic knowledge in a

text-to-speech system for Hindi. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, Aix-en-Provence, France, **3**, 494–497.

Madhukumar, A. S., Rajendran, S., Chandra Sekhar, C. & Yegnanarayana, B. (1991). Synthesizing intonation for speech in Hindi. In *Proceedings of the II European Conference on Speech Communication and Technology*, Genova, Italy, **3**, 1153–1156.

O'Shaughnessy, D. (1976). Modelling fundamental frequency, and its relationship to syntax, semantics and phonetics, Ph.D. Dissertation, MIT, Cambridge.

Pierrehumbert, J. (1981). Synthesizing intonation. *Journal of the Acoustical Society of America*, **70**, 985–995.

Ramachandran, V. R. (submitted). Role of coarticulation knowledge for a text-to-speech system for Hindi, M.S. Dissertation, Indian Institute of Technology, Madras.

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.

Thorsen, N. G. (1980). A study of the perception of sentence intonation—evidence from Danish. *Journal of the Acoustical Society of America*, **67**, 1014–1030.

Vaissière, J. (1983). Language independent prosodic features. In *Prosody: Models and Measurements* (Cutler, A. & Ladd, D. R., eds). Springer-Verlag, Berlin. 53–66.

Waibel, A. (1988). *Prosody and Speech Recognition*, Pitman Publishers, London.

Yegnanarayana, B., Murthy, Hema. A. & Ramachandran, V. R. (1990). Speech enhancement using group delay functions. In *Proceedings of the International Conference on Spoken Language Processing 1990*, Kobe, Japan, 301–304.

Yegnanarayana, B., Murthy, Hema. A., Sundar, R., Alwar, N., Ramachandran, V. R., Madjukumar, A. S. & Rajendran. S. (1990). Development of a text-to-speech system for Indian languages. In *Proceedings of the Knowledge Based Computing Systems 1990*, Narosa Publishing House, Bombay, 467–476.