SYNTHESIZING INTONATION FOR INDIAN LANGUAGES

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 examine the importance of intonational knowledge and its acquisition, representation and activation in an unrestricted text-to-speech system for Indian languages, particularly for Hindi. The declarative sentences in Hindi show a declining pattern of fundamental frequency (F_0) contour whereas interrogative sentences show a rising F_0 contour. This 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 declining tendency resets at syntactic boundaries. The values of resetting depends on the strength of the syntactic boundaries. Experiments on inherent F_0 of segments are conducted and the contextual variations of the segments in connected speech are analyzed. The knowledge obtained from the analysis is coded in a production system format. The intelligibility and naturalness of speech output increases after the incorporation of intonational knowledge.

1. INTRODUCTION

Text-to-speech systems involve conversion of an input text into a speech waveform. Humans use several knowledge sources such as phonetics, phonology, morphology, syntax, semantics and pragmatics to produce speech. 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 speech signal corresponding to the basic units of speech does not produce intelligible and natural sounding speech. Rules that govern prosodic aspects of sentences and discourse have to be incorporated. In this paper emphasis is given to acquisition and incorporation of intonational knowledge in a text-to-speech system for Hindi.

The intonational pattern is defined as the pitch pattern over time. An utterance may convey different meanings due to change in intonation, even if it is composed of the same segmental phonemes. The intonational pattern plays an important role in encoding information for the listener, because it not only conveys information about the syntactic and stress patterns, but it also helps to indicate the speaker's gender, psychological state and attitude towards what is being spoken [1]. The primary function of intonation is to signify the prominence of the most important words in a sentence or discourse and is signaled by either significant rise or fall in pitch, the perceptual correlate of fundamental frequency (F_0) [2].

2. PROPERTIES OF INTONATIONAL PATTERNS

For the present analysis, we preferred reading style to spontaneous speech. A corpus of 500 sentences was read out by two male speakers of Hindi. From the analysis of this database we observed the following major properties of intonational patterns.

2.1. Declination/rising tendency of F_0 contour

Declination tendency of F_0 contour is a characteristic attribute of declarative sentences. The F_0 and acoustic amplitude fall towards the end of a sentence. Physiologically this can be explained in terms of articulatory control. F_0 is directly proportional to the subglottal air pressure and the tension of laryngeal muscles. The subglottal air pressure falls at the end of expiration. When the tension of laryngeal muscle is not deliberately increased, the F_0 also decreases at the end of expiration. This pattern of articulatory activity causes the declination tendency for the declarative sentences in normal speech [3].

The general properties of F_0 can be summarized as follows: (1) F_0 values will fluctuate between two abstract lines - a top line and a base line, drawn near or through all maximum and minimum F_0 values in a sentence respectively. (2) F_0 range (difference between maximum and minimum F_0 values in a word) will decrease with respect to time. That is, the top line and the base line converge and both lines monotonically decrease [4]. (3) In a declarative sentence the maximum value of F_0 will be located in the stressed syllable of the first meaningful word itself. This value is almost constant in any declarative sentence. (4) Throughout the sentence there will be a repeated succession of F_0 rises and falls. In connected speech the content word together with the following post positions, if any, form a pitch accent group called a prosodic word under certain conditions. (5) In a declarative sentence F_0 contour tapers off at the end.

Speech waveform and corresponding F_0 contour for a natural utterance is given in Fig. 1.

In Hindi, the intonation pattern is not same for all interrogative sentences. Questions expecting yes/no answers have a continuous rise in 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 (eg: /kya/) the intonation pattern exhibits dual nature. That is, the F_0 contour decreases up to the question-word and then it rises. In our experiments we noticed that the question-word has the lowest F_0 values in the sentence. Even though the overall F_0 contour is different for questions, the local attributes such as fall-rise pattern remains same. Fig. 2. shows the speech waveform and F_0 contour for a yes/no type question.

Experiment 1: Prediction of intermediate peaks and valleys

We selected 10 simple declarative sentences uttered by two adult male speakers. In all the sentences the number of content words was held constant. For each speaker the maximum and minimum values of F_0 peaks are relatively constant and the intermediate peaks occur around the line which joins the above two values.

valleys also exhibit the same characteristics. By exploiting this nature using a mathematical expression we can 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

$$p_1 = p_0 + (t_1 - t_0)*(p_0 - p_n)/(t_0 - t_n).$$

Table 1 shows the actual intermediate peaks, predicted peaks and the percentage errors for each sentence for one speaker. In all the cases the error is less than 10%.

TABLE 1 : Prediction of intermediate peaks in declarative sentences

	actual		predicted		%error	
sentence	peak1	peak2	peak1	peak2	peak1	peak2
1	127.63	123.69	129.14	123.64	-1.19	0.04
2	132.88	117.13	135.56	124.88	-2.02	-6.62
3	147.37	125.00	141.42	132.56	4.04	-6.04
4	135.44	121.07	138.32	128.41	-2.13	-6.06
5	134.19	127.63	137.64	127.69	-2.57	-0.05
6	132.88	121.07	136.84	123.78	-2.98	-2.24
7	132.88	121.07	135.69	125.00	-2.12	-3.25
8	142.06	118.44	135.04	127.89	4.94	-7.98
9	144.69	135.50	152.50	140.25	-5.40	-3.51
10	157.81	144.69	154.29	139.33	2.23	3.70
mean	138.78	125.53	139.64	129.34	-0.72	-3.20
SD	8.62	8.09	7.49	5.81	3.16	3.46

Similar experiment was conducted for interrogative sentences also. The intermediate peaks can be predicted with high accuracy as in the case of declarative sentences. It is observed that the accuracy of the prediction of valleys is less compared with that of peaks. Table 2 shows the prediction statistics for intermediate F_0 peaks in yes/no type questions.

TABLE 2 : Prediction of intermediate peaks in yes/no type questions

	actual		predicted		%error	
sentence	peak1	peak2	peak1	peak2	peak1	peak2
1	176.18	226.04	175.82	216.16	0.20	4.37
2	198.49	186.68	188.37	202.97	5.10	-8.73
3	199.80	237.85	205.87	228.63	-3.04	3.88
4	199.80	239.42	203.34	242.38	-1.77	-1.24
5	202.42	277.22	198.97	217.46	1.71	21.55
6	203.73	190.17	193.68	224.83	4.94	-18.22
7	180.12	253.55	216.23	252.39	-20.05	0.46
8	207.67	215.54	189.32	174.13	8.84	19.22
9	220.79	250.97	213.49	237.42	3.31	5.40
10	193.24	236.54	199.05	223.84	-3.01	5.37
mean	198.22	231.40	198.41	222.02	-0.38	3.21
SD	12.21	26.62	11.58	20.83	7.51	11.10

Experiment 2: Comparison of range of F_0 of 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 the difference between the peak and the preceding valley (range of a prosodic word) decreases with respect to time. Table 3 shows F_0 range for each prosodic word in 10 simple declarative sentences. All sentences have equal number of prosodic words. From the table it is obvious that the range decreases with respect to the position of the prosodic word.

TABLE 3 : Comparison of F_0 ranges at prosodic word level in declarative sentences

	F_0	F_0	F_0	F_0
sentence	range1	range2	range3	range4
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

We conducted same experiments for yes/no type interrogative sentences. Here the top line and base line diverge and hence the range of prosodic word increases with respect to the position. In contrast with the observations in declarative sentences, here the final word has the maximum range and the first word has the minimum.

2.2. Fall-rise patterns

F_0 contours of sentences show a repeated succession of valleys and peaks called local fall-rise pattern. By analyzing large amount of data, we have observed some general features of local falls and rises of F_0 which are determined by the stress (pitch accent) pattern of the words. From the studies made on the F_0 contours of words in Hindi sentences the following observations are made: (1) For a content word, F_0 contour exhibits a regular pattern where a valley precedes each peak. (2) The valleys and peaks are mostly associated with the vowels which are the nuclei of the syllables in a word. As an exception, in a few cases some voiced consonants are associated with the peaks when these function as coda (the consonant that follows the vowel nucleus of syllable). (3) If the word is monosyllabic then the valley and the peak will be within the same syllable and hence F_0 will rise steadily. (4) In the case of disyllabic words and trisyllabic words the peak is found to be on the final syllable and the valley will be on the initial syllable. (5) A tetrasyllabic word is characterized by two peaks - one on the second

syllable and the other on the final syllable. These pitch accent rules are useful in synthesizing intonation patterns. The valley(s) and peak(s) for a prosodic word are assigned on similar lines of word stress rules. In such cases the case marker gets accented [5].

Fall-rise patterns are superimposed over each prosodic word whereas declination and resetting manifest throughout the sentence. Hence operationally they are distinguishable by virtue of their different domain. We can experimentally show that the magnitude of fall-rise pattern is directly related to the strength of the following syntactic boundary. F_0 contours in Fig. 1 and 2 show the local fall-rise pattern at prosodic word level.

2.3. Resetting of F_0 contour

Across major syntactic boundaries F_0 pattern gets modified. Physiologically frequency resetting can be explained in terms of 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. 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 F_0 pattern.

From our experiments on the *read sentences* of complex declarative sentences, we have observed certain features related to the resetting across syntactic boundaries. They can be summarized as follows: (1) If the first syntactic unit is longer than the second one, the resetting value becomes small. (2) If the first syntactic unit is shorter than the second one, the resetting value becomes more. (3) If the two clauses are nearly equal in length, the major parameter which determines the resetting is the pause between them. (4) The value of resetting directly depends on the pause between the syntactic boundaries. (5) In normal case the resetting value is less than the maximum F_0 value of the initial syntactic unit. (6) Within syntactic boundaries, the F_0 contour is similar to the F_0 pattern of a simple sentence. (7) The effect of resetting directly depends on the strength of syntactic boundary. It will even supercede all other rules related to the resetting values. Fig. 3. shows the effect of resetting of F_0 contour across syntactic boundaries. The sentence has two clauses, and hence one major boundary, and the resetting of F_0 coincides with this syntactic boundary.

Following study has been made based on the above observations.

Experiment 3: Intonation pattern in complex declarative sentences

We selected 15 complex declarative sentences in Hindi. Syntactic clauses can be identified from the text by the presence of the function words like /jō/, /jaisē/, /ki/, /vah/, /aur/ etc. Comparative duration of the first syntactic clause varies from 34.78% to 56.67% of the total duration of sentence. Similarly the duration of second syntactic clause varies from 26.11% to 54.35%. Conclusions from this analysis are as follows. The initial peak F_0 (first peak of first syntactic clause) is always around 180 Hz and is speaker dependent. All other significant peaks such as intermediate and final peaks of the syntactic clauses, resetting F_0 (first peak of the second syntactic clause) and tapering frequency at the end of the utterance can be related to the initial frequency. The final peak of the first syntactic clause is about 81% (ranging from 66% to 90% with standard deviation (SD) of 6.49) of initial frequency. Resetting frequency is around 92% (ranging from 88% to 97% with SD of 2.60) and the final peak of the second syntactic clause is around 72% (ranging from 61% to 84% with SD of 6.88) of the initial frequency. End tapering frequency is always constant, that is, around 57% of initial frequency. Pause between two syntactic clauses can be represented as a ratio of total duration of the utterance. We also observed that the pause between clauses are around 13% (ranging from 11% to 17% with SD of 1.82%) of total duration of the utterance. Resetting ratio (ratio of resetting value to the initial frequency) decreases with duration of the first syntactic clause while the ratio of the final frequency to the initial frequency of the first syntactic clause increases. Another obvious trend is the increase of resetting ratio and ratio of the final frequencies of the first and second clauses with respect to pause. In all these cases SD is very small implying that these relative values are consistent.

Table 4 shows the experimental results. %dur and %pause are the ratio of duration of the first syntactic clause and duration of the pause with respect to total duration of the utterance respectively. $\%F_{01}$, $\%F_{0r}$, $\%F_{0f}$ and %ta are the ratio of the final frequency of the first syntactic clause, the first and final frequencies of the second syntactic clause, and end tapering frequency with respect to the initial frequency of the first syntactic clause respectively.

TABLE 4 : Resetting of F_0 at syntactic boundaries in declarative sentences

sentence	%dur1	%pause	$\%f_{01}$	$\%f_{0r}$	$\%f_{0f}$	%ta
1	43.73	17.22	82.66	96.82	65.32	53.76
2	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

2.4. Inherent F_0

The F_0 of vowels in Hindi were studied in a frame sentence /mērā nām _____ hai/ by embedding test words. For this study we selected all possible 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 can be summarized as follows.

There is a correlation between the height of the vowel and its inherent F_0 [6]. If other factors are constant, high vowels /i,u/ exhibit higher F_0 than that of low vowel /a/. The study shows that in Hindi the difference between high vowels and low vowel is 15 to 23 Hz. Physiologically this can be explained as follows. In articulation of high vowels, the tongue is raised towards the roof of mouth. The muscles constituting the tongue are attached to the superior part of the hyoid bone and some laryngeal muscles are attached to inferior part. When the tongue is raised, the larynx tends to be pulled upwards and laryngeal muscles are stretched [7]. As mentioned earlier, the rise in laryngeal muscle tension results in the increase in the F_0. In our experiments we noticed that the inherent F_0 of mid vowels /e,o/ are nearly equal to high vowels than low vowel. The difference is 3 to 10 Hz from high vowel. The length of vowel has a definite correlation with F_0 contours: the longer the vowel 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 shifted to the vowel onset position. This effect is clearly seen in isolated utterances of test words. The peak of F_0 contour occurs at the onset of the vowel and this is about 4 Hz higher than the F_0 at 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 on the middle of the vowel nucleus.

Table 5 shows inherent F_0 of each vowel for both word initial and final positions. In all cases the final vowel has greater F_0 than the initial vowel.

TABLE 5 : Inherent F_0 of vowels

position of vowel	vowel							
	a	aa	i	ii	u	uu	e	o
initial	116.52	123.40	133.54	142.26	135.28	147.10	133.74	135.94
final	183.46	180.16	190.78	193.80	195.88	196.34	192.30	194.80

3. REPRESENTATION AND ACTIVATION OF KNOWLEDGE

We are developing a text-to-speech system for Indian languages based on parameter concatenation model. Speech data for all basic units are stored using parameters such as linear predictive coefficient(LPC)s, pitch and gain. Since speech has been modeled using parameters, voice characteristics can be manipulated and prosodic features can be incorporated by changing these parameters [8].

The intonational knowledge obtained from the analysis of natural speech has to be coded into a suitable form in order to incorporate this in the text-to-speech system. Our system is based on 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 independent of each other and the order of declaration of rules is not important. Besides the production system rules provide an easy way to give an explanation for the intermediate decisions taken.

The activation of the knowledge is achieved by means of a rule based inference engine with forward chaining control strategy. Depending upon the rule applied, the pitch contour of the synthetic speech will get modified. Fig. 4. shows the synthesized speech and the corresponding F_0 contour for the natural utterance shown in Fig. 1.

4. CONCLUSION

In this paper we briefly discussed the acquisition and incorporation of intonational knowledge in a text-to-speech system for Hindi. A large amount of speech data were collected and analyzed. The declarative sentences are characterized by the declining tendency of F_0 contour whereas the interrogative sentences are characterized by 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 and the magnitude of resetting depends on strength of the syntactic boundary. Inherent F_0 of vowels were studied. Based on these observations rules were framed and incorporated in production system format. The speech output becomes more intelligible and natural after incorporating intonational rules.

REFERENCES

[1]. Dennis H. Klatt, "Review of text-to-speech conversion for English", *J. Acoust. Soc. Am.*, vol 82(3), pp 737 - 793, 1987.

[2]. Glenn Akers and Mathew Lenning, "Intonation in text-to-speech synthesis - evaluation of algorithms", *J. Acoust. Soc. Am.*, vol 77(6), pp 2157-2165, 1985.

[3]. Philip Lieberman, *Intonation, Perception and Language*, M. I. T Press, Massachusetts, 1967.

[4]. Jacquiline Vasserie, "Language independent prosodic features", *Prosody - Models and Measurements*, ed. by A. Cutler, and D. R. Ladd, Springer Verlag, New York, pp 53-66, 1983.

[5]. A. S. Madhukumar, S. Rajendran and B.Yegnanarayana, "Significance of prosodic knowledge in a text-to-speech system for Hindi", to be presented at *XII International Congress on Phonetic Sciences,* Aix-en-provence, France, August, 1991.

[6]. Ilse Lehiste and G. E. Peterson, "Vowel amplitudes and phonemic stress in American English", *J. Acoust. Soc. Am.*, vol 31, pp 428 - 435, 1959.

[7] Isle Lehiste, *Suprasegmentals,* M. I. T Press, Massachusetts, 1970

[8]. B .Yegnanarayana, Hema A. Murthy, R. Sundar, N.Alwar, V. R. Ramachandran, A. S. Madhukumar, and S. Rajendran, "Development of a text-to-speech system for Indian languages", *Proc. Knowldge Based Computing Systems '90,* Pune, India, pp 467 - 476, 1990

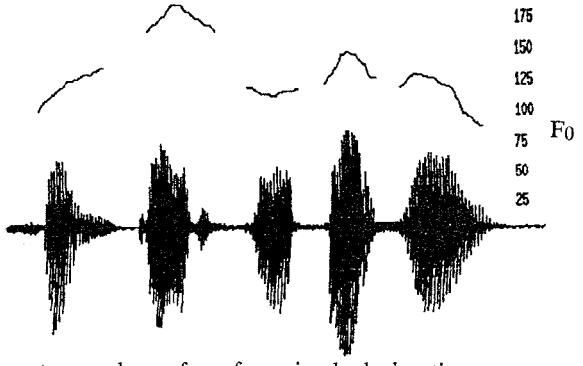

Fig. 1. F0 contour and waveform for a simple declarative sentence.
/šankar jātā hai/

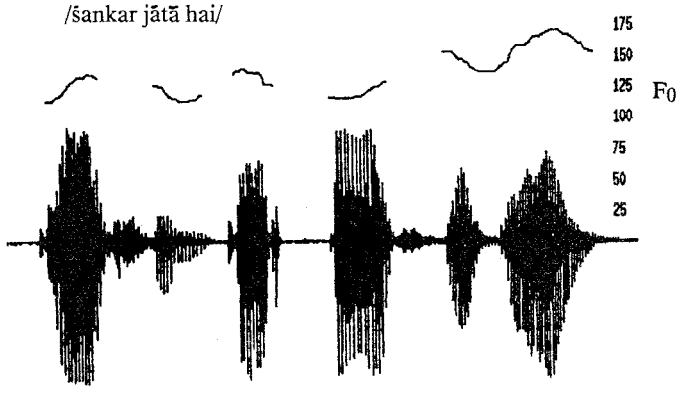

Fig. 2. F0 contour and waveform for a yes/no type interrogative sentence.
/kyā šankar pās hōgayā/

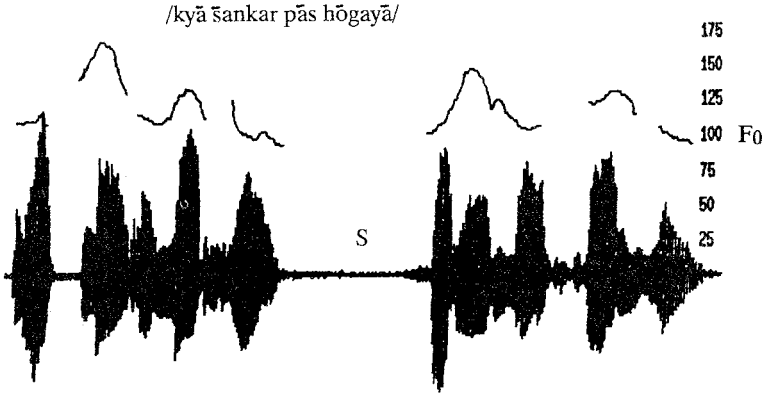

Fig. 3. Resetting of F0 contour at syntactic boundary (S).
/ātmā amar hai, sarīr nāšvān hai/

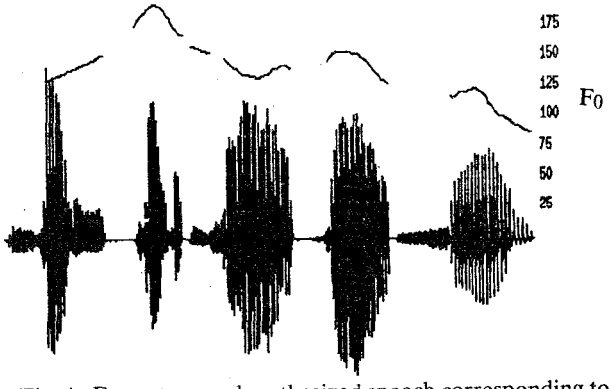

Fig. 4. F0 contour and synthesized speech corresponding to the natural utterance in Fig. 1.
/šankar jātā hai/