DETECTION OF WORD BOUNDARIES IN CONTINUOUS HINDI SPEECH USING PITCH AND DURATION

G.V.RAMANA RAO
Department of Computer Science and Engineering
Indian Institute of Technology
Madras 600 036, INDIA.
Email : ramana@iitm.ernet.in

Reliable detection of word boundaries in continuous speech is an important problem in speech recognition. Many studies established the importance of prosodic knowledge in detecting word boundaries. In this paper we report a word boundary hypothesisation technique based on the durational knowledge for Hindi. Recently another technique using pitch patterns was proposed for Hindi. We have also shown in this paper that combining the duration and pitch knowledge leads to significant improvements in the overall detection of word boundaries.

1. INTRODUCTION

Word boundary hypo‿hesisation is an important problem in continuous speech recognition. If word boundaries can be recognised accurately, many of the techniques developed for Isolated Word Recognition(IWR) systems can be adapted for continuous speech recognition. In most of the speech recognition systems built, the word boundary hypothesisation is done as part of the lexical analysis. However, in absence of word boundaries, the dictionary search requires large amounts of computer storage and time. Studies on English(Harrington and Johnstone, 1987) showed that in absence of word boundaries the number of possible word strings matching an utterance at a mid class level, may exceed 10 million. Similar results were obtained for Hindi by us. Hence in continuous speech recognition systems it becomes necessary to perform word boundary hypothesisation before performing lexical analysis.

Despite its importance, the problem of word boundary hypothesisation has only recently attracted attention. Harrington (Harrington, Watson and Cooper, 1989) examined the use of phoneme sequence constraints for word boundary detection in English, but in the presence of ambiguities in the phonemes(as in the context of speech recognition), he found them to be of limited use. However, he found that a strong/weak classification of syllables can lead to a good word boundary detection. For Hindi, Ramana Rao (Ramana Rao and Yegnanarayana, 1991) reported the use of some language clues for hypothesising word boundaries. Based on simulation studies, he reported that these clues can be used at fairly large error levels in the phoneme strings. In a recent paper, Ramana Rao (Ramana Rao, 1992) also reported the use of some spectral clues to detect word boundaries. Rajendran and others (Rajendran, Madhu Kumar and Yegnanarayana, 1992) also reported a word boundary detection technique using pitch patterns for Hindi. In this paper we report our results on using duration knowledge to hypothesise word boundaries.

We also show that combining duration and pitch knowledge sources can lead to a better recognition of word boundaries.

The paper is organised as follows. In section 2, we discuss some of the features of the Hindi language and show that duration can be used to make a distinction between the word internal and word final vowels. In section 3, we describe the technique proposed by Rajendran and others (Rajendran, Madhu Kumar and Yegnanarayana, 1992) , and describe the results on our data. We then show that by adding durational constraints, one can significantly improve the detection algorithm. In section 4, we discuss the results.

2. DURATIONAL KNOWLEDGE IN WORD BOUNDARY HYPOTHESISATION

Several studies on English showed that duration can be used to hypothesise some of the word boundaries in speech. Lea (Lea, 1980) showed that interstress intervals can be used to detect some word boundaries. Lengthening of a vowel can also be used as a clue to word boundaries. For Hindi, we have found that a simple long/short classification can lead to a good word boundary detection.

The proposed technique is based on the following features of Hindi: (1) In Hindi, very few words end in a short vowel, (2) In any Hindi text, vowels occur twice as often as consonants in the position before a word boundary. This is because, in any text case markers and other function words occur nearly 40% of the total words and most of these end in vowels. For the remaining words the vowels and consonants occur roughly equal times before a word boundary. Thus in the entire text, vowels occur 65-70% before word boundaries and the consonants occur the rest. Now if one combines the above two features, one can conclude that long vowels occur 65-70% before word boundaries. Thus a simple classification of vowels based on length would detect nearly two thirds of the word boundaries.

However long vowels can also occur in word internal positions. However more short vowels occur word internally than long vowels. Hence a long/short vowel classification will detect many word boundaries and make fewer errors.

A good estimate of the performance of the above method can be obtained by studying the distribution of vowels in large texts. Using a text of 800 sentences containing 17,600 vowels and 10800 word boundaries, we found that long vowels are 10800 in number. Among these, long vowels in word final position numbered 8,200. Hence detection of long vowels will detect nearly 80% of the word boundaries with errors around 25%. Similar results can be expected from any other texts.

We applied the above technique on a speech data consisting of 670 vowels extracted from three speakers. The text was 40 sentences. Of these speakers 1 and 2 spoke 10 sentences each and speaker 3 spoke 20 sentences. From these utterances, the vowels were segmented manually. These were then classified into long/short vowels using a simple threshold. The results are shown using two measures, hit rate and false alarm rate. The hit rate gives the percentage of word boundaries detected and the false alarm rate specifies the percentage error in the word boundary hypotheses generated by our technique. They are defined as follows.

$$\text{hit rate} = \frac{\text{No. of word boundaries hypothesised}}{\text{No. of word boundaries in utterance}}$$

$$\text{false alarm rate} = \frac{\text{No. of erroneous hypotheses}}{\text{No. of word boundary hypotheses}}$$

Table 1 shows the results for the three speakers. From these, one can see that the proposed technique performs the word boundary hypothesisation fairly well.

Speaker	No. of vowels considered	Hit rate	False alarm rate
1	193	75%	22%
2	170	89%	19%
3	307	83%	23%

TABLE 1 Results of word boundary detection using duration

3. WORD BOUNDARY HYPOTHESISATION USING PITCH AND DURATION

In this section, we present a new technique for hypothesising word boundaries in Hindi speech using pitch and duration of vowels. Recently it was suggested that each content word in continuous Hindi speech has a pitch pattern, namely the pitch frequency(F0) increases from left to right. Thus in the sentence 'narmada: nadi: ...' the word narmada: will have F0 increasing from left to right. Similarly the word nadi: will also have an increasing F0 from left to right. Since F0 in simple sentences falls from left to right, the fall will occur at the boundary between the two words. Thus by detecting such falls in F0 one can detect word boundaries. Strictly speaking one detects word final vowels. The word boundary hypothesisation algorithm is as follows:

1. Detect the peaks in the pitch contour of the utterance and hypothesise these vowels as word final vowels,
2. Hypothesise a vowel as a word final vowel if its F0 is larger than that of the next vowel.

We used this algorithm to detect word boundaries in the above mentioned speech data. The results are shown in Table 2.

Speaker	No. of vowels considered	Hit rate	False alarm rate
1	193	70%	18%
2	170	74%	15%
3	307	67%	33%

TABLE 2 Results of word boundary detection using pitch(F0)

From the results, it can be seen that the algorithm performs fairly well for speakers 1 and 2. But for speaker 3, the performance is quite poor. We examined the errors and found that in a majority of the cases, they correspond to short vowels. Hence we modified the above word boundary hypothesisation algorithm to include a third step to eliminate these errors. Thus our modified word boundary hypothesisation algorithm is as follows:

1&2 Same as in the old algorithm
3 Apply a duration threshold to eliminate most of the short vowels.

The results of the modified algorithm are shown in Table 3.

Speaker	No. of vowels considered	Hit rate	False alarm rate
1	193	68%	10%
2	170	73%	6%
3	307	64%	14%

TABLE 3 Results of word boundary detection using F0 and duration

They show a substantial improvement for all the three speakers, especially with speaker 3 for whom the false alarm rate has dropped to a one-third of the earlier case. Further improvements are obtained by using a threshold for the drop in F0.

While the above modified algorithm for word boundary hypothesisation works well, it is difficult to explain these results. One way to explain these results is by assuming that the fall in F0 that must occur after a word boundary needs a minimum duration. Thus if the vowel after a word boundary is a short one, the fall will not completely occur but will continue into the next vowel and hence the short vowel will have a larger F0 compared to its next vowel even if they are in the same word. However this is only a conjecture and we do not have enough results to support it.

4. SUMMARY AND CONCLUSIONS

The above results showed that durational knowledge can significantly aid in hypothesising word boundaries in continuous speech. We have proposed a modification to the word boundary hypothesisation algorithm using pitch and showed that the modified algorithm performs better than the earlier one. Currently we are trying to incorporate the third prosodic feature, intensity, into the word boundary hypothesisation algorithm.

To conclude we have proved that prosodic knowledge can be used to detect many word boundaries in continuous speech. However this knowledge is language dependent and its utility will have to be reexamined for other languages. One of our current tasks is to examine if similar knowledge can be extracted for other Indian languages and to observe the similarities among the various Indian languages in this area.

REFERENCES

Harrington J. and Johnstone A. (1987), 'The effects of equivalence classes on parsing phonemes into words in continuous speech recognition', Computer Speech and Language, 2, 273-288.

Harrington J., Watson G. and Cooper M. (1989), Word boundary detection in broad class and phoneme strings, Computer Speech and Language, 3, 367-382.

Lea W. (1980), Prosodic aids to speech recognition, pp166- 205, Trends in Speech Recognition, Prentice Hall, New Jersey.

Rajendran S., Madhu Kumar A.S. and Yegnanarayana B., Word boundary hypothesisation for continuous speech in Hindi based on pitch accent rules, Communicated to Computer Speech and Language.

Ramana Rao G.V. and Yegnanarayana B. (1991), Word boundary hypothesisation in Hindi speech, Computer Speech and Language, 5, 379-392.

Ramana Rao G.V.(1992), Detection of Word final vowels in speech using first formant energy, Proc. of the second regional workshop on Computer Processing of Asian Languages, Kanpur, India, 243-247.

* * * * * *