PERFORMANCE OF ISOLATED WORD RECOGNITION SYSTEM FOR CONFUSABLE VOCABULARY

S. Raman

Department of Electrical
Engineering

B. Yegnanarayana

Department of Computer
Science and Engineering

Indian Institute of Technology
Madras 600036 INDIA

ABSTRACT

In this paper we discuss some of the limitations of the existing isolated word speech recognition system (IWSR) when applied to confusable vocabulary. For our study we have chosen a subset of Hindi stop consonants as the confusable word set. The members of this set differ among themselves primarily in the short leading consonantal part and at the interface of the consonant and the following dominant vowel part. We adopt a signal-dependent approach for parameter extraction and matching strategy. This approach gives better performance compared with the conventional approach, but the performance still falls far short of the desired goal of 100% recognition. Refined signal processing suitable for appropriate segments of speech appear to be the way out of this problem. We discuss our studies in this direction. We use a new measure, called Performance Index, to evaluate the changes in performance due to innovations carried out on small data sets.

I. INTRODUCTION

Studies on isolated word speech recognition system for different vocabularies have raised doubts about the achievable recognition accuracy with these systems, especially when dealing with large vocabularies [1]. A highly reliable working system these days is one which deals with a small vocabulary of words having minimum confusability among them. Our objective is to develop systems which can effectively deal with confusable words also. The strategy for recognition should be based on focussing our attention on the differences in the characteristics of the words in the vocabulary. In literature, one finds the E-set (B,C,D,E,G,P,T,V,Z,3) of the alpha-digit vocabulary as one example of the confusable word set. Several refinements of the conventional approach have been suggested to deal with such confusable word sets. Parameter optimization has been proposed to enhance those aspects of the signal which contribute significantly to the detection of phonetic differences [2]. Templates are created to maximally represent the vocabulary items [3]. Spectral and temporal features to describe the acoustic cues have been evaluated [4]. Refinements in dynamic time warping, integration of feature-based and template-based system [5] and emphasis of transients over steady-state portions [6] have been attempted. Some of these refinments have reduced the error rate for the E-set from 37% to 10% [5].

Although the E-set appears to be acoustically similar due to their common vowel ending, the leading portions of the words differ significantly in their excitation characteristics. Hence any scheme which reduces the dominant effect of the vowel portion should result in increase in the recognition performance. This does not necessarily indicate that the scheme has superior features for distinguishing the leading portions. A more effective test on the capability of a recognition scheme will be on its ability to handle stop consonants whose characteristics are not easy to extract [7].

In this paper we report our studies on the recognition of Hindi stop consonants, which represent a vocabulary more confusable than the E-set. The members of this set are: क (ka), च (cha), ट (ta), त (tha) and प (pa). These are voiceless stop consonants characterized by different places of articulation. Acoustically, they differ among themselves in their short leading consonantal part and at the interface with the following vowel part. If the distinguishing primary features are not captured in these short durations, there is almost no chance of recovery from error in the recognition based on the remaining portions of the utterance. We discuss the results of signal-dependent matching [8] on this vocabulary set. Section II deals with a discussion on signal-dependent matching algorithm and its performance on the E-set and the

CH1945-5/84/0000-0183 $1.00 © 1984 IEEE

Hindi consonants set. Further experiments on the consonants set are described in Section III.

II. SIGNAL-DEPENDENT MATCHING

In the conventional approach the speech information in each frame is represented in parametric form for further processing and matching. The issues in matching are the nonlinear time registration of the test utterance with a reference word and the distance computation between a test frame and a reference frame. The recognition accuracy depends on various design choices like the size and nature of vocabulary, speaker dependence, background noise, etc., For a given specification, there are other factors that limit the performance of the systems. These factors include ambiguity in endpoint detection, fluctuations in the parameter contours, inadequacy of noise normalization and fixed matching strategy.

To overcome some of these limitations, a signal-dependent matching strategy, which makes use of the signal knowledge for parameter extraction and matching, has been proposed [8]. In this method the number and type of parameters are not fixed apriori, but are selected based on the nature of the input speech segment. The measure of performance described in [8] is used in our studies. The measure reflects the changes in the performance due to innovations in the system design and it is useful while working with limited sets of data on a modest computational facility. Table-1 shows the improvement in the Performance Index (PIX) values for signal-dependent approach over those for the conventional approach. The results for three sets of vocabulary [8] are given in Table-1. Under identical conditions of frame size, parametric representations, etc., the PIX value obtained for Hindi stop consonants is only 25.8%, even for signal-dependent matching. This clearly shows that the Hindi set is more confusable than the E-set.

Each Hindi consonant word consists of two parts, namely, the short leading consonantal part and the following dominant vowel part. The consonantal part, being short and transitory in nature,can be represented more effectively by a small number of spectral parameters over small segments of data. The small frame size provides the required temporal resolution. On the other hand, the vowel part requires more spectral resolution than the consonantal part. Thus the parameters and the frame rates have to be adaptively chosen based on the signal knowledge. Table-2 shows that by increasing the temporal resolution and decreasing the spectral resolution in the consonantal part, the PIX can be increased significantly. The effect of increased temporal resolution is reflected in the overall PIX because the duration of the consonantal part (including silence) is of the same order as that of the following vowel part in this case.

When the duration of the consonantal part is small compared with the vowel part, the overall PIX is dominated by the influence of the vowel part. This can be seen from the results in Table-3 for three sets of Hindi consonants, all ending with the same vowel. A much higher PIX value is obtained for the consonant part alone, indicating the higher discrimination of these parts compared with the vowel parts. Although the PIX value for the consonantal part is more compared with that for the entire utterance, the recognition of the consonants is not more than 60% for the three sets of data given in Table-3. On the other hand, for a word set consisting of the same consonantal part but with different vowel endings, the overall PIX value is around 80% and the recognition of the vowels is 100%. The interesting point is that the PIX for the whole utterance is dictated by the vowel parts for the data sets in Table-3 and Table-4. In addition Table-4 illustrates that representation of the vowel part by linear predictor coefficients yields better results than if the representation is by log melspectral values. The distance between two linear prediction spectra is computed using cepstral distance [9]. The influence of the consonantal part on the overall PIX can be seen in Table-4 when the frames are overlapped with 192 samples. The PIX value for the vowel part changes little, whereas the PIX value for the overall distance has decreased from 86.5% to 75.1% due to the influence of the consonantal part.

III. NEED FOR A DIFFERENT APPROACH FOR CONFUSABLE WORD RECOGNITION

The performance of signal-dependent matching for the confusable words set shown in Table-3 is not high enough to design a practically useful system. The PIX values for a practical system should be in the range of 80% to 100%. A different approach to recognition of confusable words consists of identifying and isolating those parts of utterance which primarily characterize the consonantal parts. It is necessary to deemphasize the transition regions which carry only the secondary clues for the consonant discrimination. The disadvantage of using the secondary clues for consonant discrimination is that these clues are different for different vowel endings.

To develop a recognition scheme for the confusable word set, we manually isolated the leading consonantal part by observing the speech waveform and by listening to different portions of the speech utterance. In our experiments we have found that the leading consonantal part can always be isolated. We have also verified through listening tests that the isolated consonant part does carry the discriminating information required for recognition.

Having identified the consonantal part, it should be appropriately represented to reflect the characteristics of the consonant. The representation requires good temporal resolution to capture the articulatory dynamic characteristics of the particular consonant. In addition to the parametric representation, the excitation characteristics and the prosodic features such as duration should be used to describe the consonantal part. We are currently working on methods to represent the leading consonantal parts for recognition purposes.

IV. SUMMARY AND CONCLUSIONS

Hindi stop consonants represent a typical confusable word set. Identifying the consonantal and vowel parts of the utterances of these words and applying different parametric representation and matching techniques resulted in some improvement in performance. But the improvement is not adequate to design a recognition system. A totally different approach is needed in which the leading consonantal part is isolated and handled separately for recognition. One reason for doing this is that we have obtained better performance for the consonantal parts of the utterances. Moreover, it is necessary to reduce the influence of the dominating vowel part. The justification for using the consonantal parts alone for recognition is that we have been able to obtain 100% recognition of the consonants in the listening tests when the consonantal parts were actually concatenated with different vowels. Our approach for the design of a recognition system consists of representing the different parts of speech signal appropriately, giving more importance to classification based on the characteristics of the primary consonantal part. The secondary clues from the vowel part may be considered as additional information. We are currently exploring some of these ideas to design a recognition system for confusable words.

REFERENCES

[1] L.R. Rabiner, A.E. Rosenberg, J.G. Wilpon and W.J. Keilin, 'Isolated Word Recognition for Large Vocabularies', BSTJ, Vol.61, No.10, Dec. 1982, pp. 2989-3005.

[2] S.B. Davis and P. Mermelstein, 'Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences', IEEE Trans. ASSP, 28(4), Aug. 1980, pp. 357-366.

[3] L.R. Rabiner, 'On Creating Reference Templates for Speaker Independent Recognition of Isolated Words', IEEE Trans. ASSP 26, Feb. 1978, pp.34-42.

[4] Günther Ruske, 'Automatic Recognition of Syllabic Speech Segments Using Spectral and Temporal Features', Proc. ICASSP 82, Vol.1, pp. 550-553.

[5] Gary L. Bradshaw, R. Reddy and Zongge Li, 'A Comparison of Learning Techniques in Speech Recognition', Proc. ICASSP 82, Vol.1, pp. 554-557.

[6] K. Elenius and M. Blomberg, 'Effects of Emphasizing Transitional or Stationary Parts of the Speech Signal in a Discrete Utterance Recognition System', Proc. ICASSP 82, Vol.1, pp. 535-537.

[7] H. Fujisaki and M. Tominaga, 'Automatic Recognition of Voiced Stop Consonants in CV and VCV utterances', Proc. ICASSP 82, Vol.3, pp. 1996-1999.

[8] B. Yegnanarayana, T. Sreekumar and S. Raman, 'Signal-dependent Matching for Speech Recognition', Paper presented at International Conference of System Man and Cybernatics Society IEEE, Bombay, Jan. 1984.

[9] A.H. Gray and J.D. Markel, 'Distance Measures for Speech Processing', IEEE Trans, ASSP 24, Oct 1976, pp. 380-391.

Table-1 : Performance Index for three sets of vocabulary by conventional and signal-dependent approaches

Approach	/B,C,D,E,G,P,T,V,Z/	/A,J,K /	/ 0,1,....,9 /
Conventional	70.25	85.63	84.50
Signal-dependent	84.26	97.33	93.47

Table-2 : Performance of signal-dependent matching techniques for confusable vocabulary

Expt.No	Frame size (No. of samples/frame)		No. of log melspectral parameters		PIX
	Consonant	Vowel	Consonant	Vowel	
1.	256	256	16	16	25.8
2.	256 (with overlap of 192)	256 (with overlap of 192)	16	16	50.5
3.	256 (with overlap of 192)	256	4	16	68.0

Table-3 : Performance Index for different parts of the utterance represented by log melspectral parameters

Set No.	Performance Index		
	Overall	Consonant	Vowel
1.	34.20	56.1	24.20
2.	35.90	58.4	30.75
3.	27.75	52.3	24.15

Table-4 : Performance Index for different parts of the utterance represented by appropriate parameters

Set No.	No. of samples		Parameters for				Performance Index	
	per frame	overlapping	Consonant		Vowel		Overall	Vowel
			Type	No.	Type	No.		
1.	256	-	log melspec.	4	log melspec.	8	78.90	79.65
2.	256	-	log melspec.	4	LPC	8	86.50	88.65
3.	256	192	log melspec.	4	LPC	8	75.10	89.85
4.	256	192	log melspec.	4	log melspec.	8	67.35	73.80