NEAREST NEIGHBOUR DECISION RULE FOR VOWEL AND DIGIT RECOGNITION

T.K. Raja and B. Yegnanarayana

Department of Electrical Communication Engineering
Indian Institute of Science
Bangalore-560012, India

ABSTRACT

Minimum distance to mean is usually used as a classification rule in speech and speaker recognition studies. In this paper it is shown that the nearest neighbour decision rule gives significant improvement in classification score for vowel and digit recognition schemes. Autocorrelation coefficients of lags two to five sampling instants are used to form the feature vector. Four samples per class have been used. Minimum squared Euclidean distance of the test vector from the nearest reference is chosen as the classification rule. For sustained vowels the recognition score is cent percent. For the same feature the minimum distance to mean gives 70 % recognition score. When the reference samples of a given speaker is tested over the vowels spoken by different speaker(up to 10), this scheme gives the recognition score of about 95 %. For digits without any time warping the recognition score of about 86 %. to 92 % is obtained.

I. INTRODUCTION

The success of speech recognition by machine depends on the selection of an appropriate feature and the classification algorithm. Ideally, the feature selected should be widely separated in the feature space, while be least affected by the environmental conditions, inter and intera speaker variations. Several studies on speech recognition, in particular, vowel and digit recognition schemes, mainly use the spectral information. This spectral information may be formants or the linear prediction coefficients (LPC's)[1,2]. Autocorrelation coefficients[3] of signals derived from filter banks and zero crossing rate (ZCR)[4] have also been used. Most of the speech recognition schemes use the minimum Euclidean distance to means as the criterion for classification. This criterion is although computationally simple, it assumes equal variance and unimodal Gaussian distribution for the features. Several other classification rules based on Itakura measure[5], log spectral distance[6] have also been reported. In this paper, we shall discuss the performances of vowel and digit recognition schemes based on the autocorrelation as the feature and the nearest neighbour decision rule (NNDR)[7] for classification. For more information on speech recognition systems, refer D.R. Reddy's paper [9].

II. THE FEATURE AND THE CLASSIFICATION RULE

It is well known the autocorrelation and the power spectrum of the signal are form a Fourier transform pair since the speech information is characterised by the spectral envelope, it appears logical that the autocorrelation coefficients could as well be used as feature. The LPC's are derived from autocorrelation coefficients by solving the autocorrelation normal equations.[8] Hence, instead of using LPC's and other features derived from LP analysis, one can as well use directly the autocorrelation coefficients as a feature. In this study, we have used the normalized autocorrelation coefficients of lags two to five sampling instants. The normalization is necessary to normalize the gain variations in the speech signal. And also this feature is found to be less susceptible to additive white noise and inter and intera speaker variations. The figure 1 shows the autocorrelation feature for the vowels /e/ approaches its true value when the signal to noise ratio(SNR) is greater than 12 dB and this character remains the same for all the vowels.

Any general classification algorithm is required to determine the weight vectors of the discriminant functions [7]. The efficient method of estimating the weight vectors is based on the availability of large number of training samples in order to obtain the convergence of weight vectors. The procedure involved in estimating the weight vectors are computationally expensive. The Mahalanobis distance criterion is also equally complex as it is required to estimate the sample mean and variance which also needs large number of training samples. On the other hand, the minimum Euclidean distance to mean criterion is simple but it requires the

CH1285-6/78/0000-0731$00.75@1978IEEE

samples to be unimodal, equal variance Gaussian distribution. In all these cases the features belong to different classes are to be well separated in feature space. The NNDR can be thought of as a compromise between these two extremes. This rule only assumes the samples form a well separated clusters in the feature space. In addition, the NNDR does not require the tedious computational procedures of estimating any of the parameters already discussed.

The basic principle of NNDR is that its assigns class to the test sample(or test set) to which its nearest neighbour in the design samples(or design set) belongs. The mathematical description of this techniques is as follows:

$$d_{ij}^2 = \| X_{ij} - X_t \|^2 \qquad (1)$$

where d_{ij}^2 is the squared Euclidean distance between the reference sample and the test sample. X_{ij} is the ith sample reference belong to C_jth class and X_t is the test sample. Both X_{ij} and X_t are of same dimension.

$$\text{If,} \quad (d_{ij}^2)_{min} = \| X_{ij} - X_t \|^2 \qquad (2)$$

then $X_t \in C_j \quad \forall\, i,j$.

III. EXPERIMENT

Experiments have been conducted to evaluate the autocorrelation feature with NNDR as a classification algorithm using the HP Fourier analyser system 5451B. The speech data is entered into the system through its built in A/D conversion unit. The Schur microphone is used as a transducer and is kept at a distance not more than 3 inch from the mouth. Both design and testing is done in the computer room environment. The back ground noise is about 60 dB. Throughout this study a sampling rate of 10 kHz and Hanning window is used. The number of speakers participated in this experiment are 10 (8 male + 2 female). For sustained vowel (/a/,/i/, /u/,/e/, and/o/)recognition, the autocorrelation coefficients (R(2) to R(5)) are computed from the stable portion of the signal of duration 25.6 m.sec. Only four samples per class for all the vowels per speaker is computed and is stored in the paper tape. The dimensionality of the feature space is 4. When dipthongs /AI/, and /AU/ (in Indian languages they are called vowels) are included in already discussed vowel recognition scheme, the contour of autocorrelation coefficients (R(2) to R(5)) for four successive frames each of duration 25.6 msec. are used. The dimensionality of the feature space is 16 and the number of classes are seven. Since the dipthongs are articulated in sequence first by /A/ and then by /I/ or /U/ as the case may be and hence it is necessary to consider the initial portions of the utterances. Only four samples per class for all the vowels per speaker is computed and is stored in the paper tape. For digit recognition, the autocorrelation coefficients (R(2) to R(5)) are computed in each frame of duration 25.6 msec. for successive 16 frames. Thus the dimensionality of the feature space is 64. In this case only 3 samples per class for all the 10 classes per speaker is stored in the paper tape. In all these cases the dc power is removed from the over all power spectrum. The sequence of feature extraction procedure is shown in the block diagram of Figure 2.

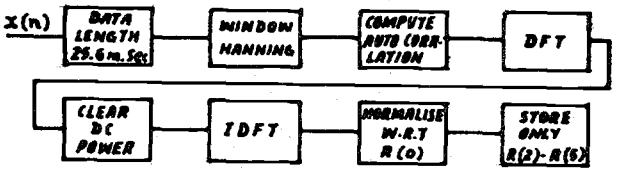

Fig.2: Block diagram of feature extraction cycle

The DFT and IDFT are used to remove the DC power which varies from utterance to utterance. The testing of this recognition scheme was spread over a period of nine months.

IV. RESULTS

A. Sustained Vowel Recognition:

Each speaker's reference data on the paper tape is transferred to the system's memory one at a time. 100 utterance of each vowel from all the speakers are tested. The test sequence procedure is shown in the block diagram of Figure 3. The

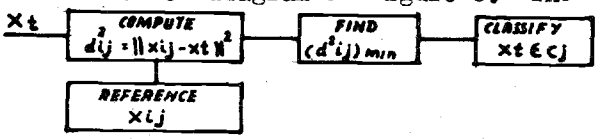

Fig. 3: Block diagram of recognition cycle

results of the experiment is shown in Table-1, and is cent percent if the reference and the test vectors are from the same speaker, otherwise; it is 96.2 % on an average taken over speakers. The recognition score remains unaltered if the SNR is greater than 15 dB. The noise derived from the random noise source is added to the signal.

B. Vowel Recognition including Dipthong:

The experimental procedure described in para IVA is repeated and the results are shown in Table-2 and is 100 % if the

reference and test vectors are from the same speaker, otherwise; it is 89.9 %. on an average taken over the speakers.

C. Digit Recognition:

The experimental procedures described in para IVA is repeated but the number of utterance per digit per speaker is 50. The results of the experiments are shown in Table 3 to 6. The recognition score for languages Hindi, Telugu, English, Kannada, and Tamil respectively are 86.8 %., 87.6 %., 88.1 %., 90.7 %., and 92.2 % if the reference and test vectors are from the same speaker, otherwise; it is 80.6 %., 81.1 %., 81.9 %., 82.3 %., and 83.6 % on an average taken over the speakers. The Table 3 to 4 is for the Tamil language which shows the highest recognition rate and the Table 5 to 6 is for the Hindi language which shows the lowest recognition rate.

The same experiment is repeated by using Fortran IV programming and IBM 360 computer for language Tamil only. In this case the reference samples per class is 16. Four utterances per speaker per class from one female and three male speakers were taken for reference. The number of test utterances per class per speaker is 10. The recognition rate has gone upto 98.1 % on an average taken over speakers.

V. DISCUSSION ON RESULTS

In the case of sustained vowel recognition under additive white noise condition, it is interesting to note that only the distance between classes increase uniformly but still retaining the minimum distance to the class that is correctly classified under no additive white noise condition.

In the case of digit recognition, the recognition score looks to be better for the languages English, Kannada, and Tamil. It may be due to the reason that all the speakers are from the Tamil origin and also all the speakers are well conversant with English and Kannada It may also be due to the reason that for the languages Kannada and Tamil, all the digit words are end with the same character namely /u/ and hence the last character is redundant. As there is a possibility of any digit word utterance exceeding the duration of 409.6 msec., the loss of information at the end for languages Tamil and Kannada digit words may not affect the recognition capability. But this is not the case in other languages.

In the above discussion, it is mentioned that the NNDR works better than the minimum distance to mean criterion. The reason is that the NNDR needs only well separated clusters in the feature space. The situation where the minimum distance to mean criterion fails even if the cluster are well separated can be explained by refering to Fig. 4. Let A_1 and A_2 be the two different cluster

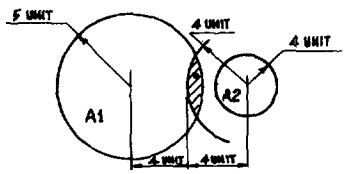

Figure:4. Cluster area for class C1 & C2

areas belong to class C_1 and C_2 if the variance are unequal and unimodal Gaussian distribution in two dimensional feature space. The minimum distance to mean (centre of the cluster) criterion always misclassifies the samples falling in the shaded region. But the NNDR always classifies correctly even if any of the samples falls in the shaded region since it considers only the minimum distance to individual samples in the clusters.

VI. CONCLUSION

This study shows the importance of NNDR together with Autocorrelation feature in speech recognition. If the reference samples that falls only on the periphery of the clusters but well separated on the periphery are selected, the recognition performance can be improved and the memory size to store the reference can also considerably be reduced.

ACKNOWLEDGEMENT

The authors wish to thank Prof.B.S. Ramakrishna, Dr. V.V.S. Sarma and Mr. T.V. Ananthapadmanabha for many useful discussions.

	/a/	/i/	/u/	/e/	/o/
/a/	100				
/i/		100			
/u/			90	7	3
/e/		5	1	94	
/o/			4		96

Table-1: Vowel recognition(sustained) Reference and test vectors from different speakers.

REFERENCES

1. G.M. White and B.R.Neely,'Speech Recognition Experiment with Linear Prediction,Bandpass filtering and Dynamic programming',IEEE Trans.Acoust.,Speech, and Signal Proc.,Vol.ASSP-24,No.2, pp.183-188, April 1976.

2. M.R.Sambur,and L.R.Rabiner,'A statistical decision approach to the recognition of connected digits',IEEE Trans. Acoust.,Speech,and Signal Proc.,Vol. ASSP, p.24, No.6, Dec. 1976.

3. R.F.Purton, 'Speech recognition using autocorrelation analysis', IEEE Trans. Audio Electro Acoust., Vol. AU-16, p.2235, 1968.

4. W.Begdel and J.S. Bridel, 'Speech recognition using zero crossing measurements and sequence information', Proc. Inst.Elec.Eng. Vol. 116, pp.617-623, 1969.

5. F.Itakura, 'Minimum prediction residual principle applied to speech recognition' IEEE Trans.Acoust.,Speech and Signal Proc., Vol. ASSP-23, pp. 67-72,Feb.1975.

6. A.H. Gray Jr. and J.D. Markel, 'Distance measures for speech processing', IEEE Trans.Acoust.Speech,and Signal Proc., Vol. ASSP-24, No.5, pp.380-391, Oct. 1976.

7. R.D. Duda and Hart, 'Pattern classification and scene analysis', John Wiley and Sons, New York, 1973.

8. J. Makhoul, 'Linear prediction: A tutorial review', Proc.IEEE, Vol.63, No.4, pp. 561-580, April 1975.

9. D.R. Reddy, 'Speech recognition by Machine', Proc. IEEE, Vol. 64,p.501, April 1976.

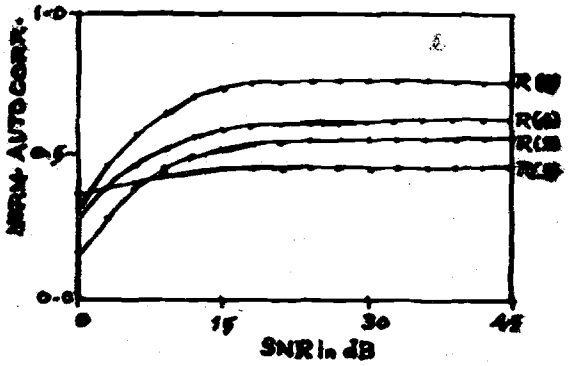

FIGURE:1.Noise in dB vs normalized Autocorrelation Coefficients R(2) to R(5)

	/A/	/I/	/U/	/E/	/AI/	/O/	/AU/
/A/	100						
/I/		100					
/U/			81	5		8	6
/E/		9	1	90			
/AI/		7		9	84		
/O/			5			93	2
/AU/	4		7			10	79

TABLE:2.Confusion Matrix(For Vowel including dipthongs,different speaker reference)

	0	1	2	3	4	5	6	7	8	9
0	45	1		4						
1	3	40		7						
2			47			3				
3		4		46						
4					45		5			
5				1		49				
6				2			48			
7								50		
8		5		1					44	
9		3								47

TABLE:3.Confusion Matrix(For Digit(Tamil) same speaker reference)

	0	1	2	3	4	5	6	7	8	9
0	40	3		6					1	
1		41		7						2
2			44			6				
3		4		45	1					
4					38		12			
5				3		47				
6				2	5		43			
7			1	2				42	5	
8		11		3		1			35	
9	1	5								44

TABLE:4.Confusion Matrix(For Digit(Tamil) different speaker reference)

	0	1	2	3	4	5	6	7	8	9
0	42		5				3			
1		39		2			9			
2	2		48							
3		2		44			4			
4					41	2		6	1	
5						46		3	1	
6				3			47			
7	3							40	7	
8				2				8	40	
9						1	1		1	47

TABLE:5.Cofusion Matrix(For Digit(Hindi) same speaker reference)

	0	1	2	3	4	5	6	7	8	9
0	34	2	9		4		5			
1	3	33		5			9			
2			48		2					
3		2		43			5			
4					35	2		9	4	
5					1	40		7	2	
6		1		7			42			
7						1		45	4	
8					3	5		6	36	
9					4					46

TABLE:6.Confusion Matrix(ForDigit(Hindi) different speaker reference)