STUDIES ON PATTERN RECOGNITION APPROACH TO VOICED-UNVOICED-SILENCE CLASSIFICATION

V.V.S. Sarma and D. Venugopal

Indian Institute of Science
Bangalore 560 012, INDIA

Abstract

A pattern recognition approach for deciding whether a given segment of speech should be classified as voiced speech, unvoiced speech or silence based on a set of five measurements of the signal is given by Atal and Rabiner [1]. In this paper, we demonstrate that it is possible to achieve this classification with much less computational effort. These computational savings are mainly achieved by adopting a scheme based on the concept of variable decision space, using only three features and by avoiding the time consuming linear prediction analysis.

I Introduction

The need for deciding whether a given segment of speech should be classified as voiced speech (V), unvoiced speech (UV) or silence (S) arises in many speech communication and understanding systems. Atal and Rabiner [1] have recently formulated this problem as a 3-class pattern recognition problem and demonstrated the usefulness of this approach for several applications. The attractive features of this scheme is that it effectively delinks V-UV-S classification from pitch analysis. They use five features: the zero-crossing rate (N_z), the speech energy (E_s), the correlation between adjacent speech samples (C_1), the first predictor coefficient from a 12-pole linear prediction analysis (a_1) and the energy in the prediction error (E_p). Rabiner and Sambur [2] used the same set of features under various recording conditions for connected digit recognition. Siegel and Steiglitz [3] also formulate V/UV decision as a pattern recognition problem, using five features; RMS value, N_z, peak amplitude, E_p and ratio of high to low frequency energy (HILO) in the signal and propose a nonparametric linear classification scheme using three of the features RMS, E_p and HILO.

It is well known that the design of any pattern recognition scheme is highly iterative process [4]. In this paper, we demonstrate that the basic scheme of Atal and Rabiner can be improved further leading to considerable decrease in computational effort without increasing the error. This is achieved by using only 3 features and by employing a classification scheme based on the concept of variable decision space. The three features chosen are E_s, N_z and C_1 thus avoiding the need for linear prediction analysis. The variable decision scheme approach solves the 3-class problem as a sequence of two 2-class problems.

II Speech Measurements and Decision Algorithms

A Measurements

The speech measurements are made using a HP 5451 A Fourier Analyzer system incorporating a HP 2100S minicomputer. The speech signal was given to the system's A/D converter via a microphone in a fairly noisy environment. The microphone was kept close to lips to enhance the signal to noise ratio. The average signal to noise ratio for voiced speech was 43 dB and for unvoiced speech 12 dB. The analog speech signal was sampled at 10 kHZ and each sample was quantized with an accuracy of 14 bits. No high pass filtering was done to remove the hum or noise components. They formed the background noise and constituted the silence class of the signal. The speech was formatted into blocks of 128 samples (12.8 m sec) as the signal processing operations on the system are carried out on fixed blocks of lengths 64, 128 etc. For each block we define $s(n)$, $n=1,2,\ldots,128$ to be n^{th} sample in the block. The following three features were computed for each block of samples.

1. Log energy E_s, defined by

$$E_s = 50\,\text{dB} + 10\log_{10}\left(\sum_{n=1}^{128} s^2(n)\right) \qquad (1)$$

2. Normalized autocorrelation coefficient at unit sample delay defined by

$$C_1 = \left[\sum_{n=1}^{128} s(n)\,s(n-1)\right] \Big/ \sum_{n=1}^{128} s^2(n) \qquad (2)$$

3. The number of zero crossings in the block (N_z) after filtering out the dc term. There are two reasons for the choice of only the three measurements E_s, N_z and C_1 among the five considered by Atal and Rabiner [1]. Firstly, E_s, N_z

CH1285-6/78/0000-0001$00.75@1978IEEE

and C_1 can be obtained with much less computational effort compared to E_p and a_1. It is easy to see that each of the first three features need just N arithmetic operations for a frame of N samples. On the other hand, a_1 and E_p require LP analysis needing approximately $(MN+\frac{1}{6}M^2)$ operations where M is the order of linear predictor. Secondly, it is well known in statistical literature that increasing the number of features in a pattern recognition problem does not always increase the power of discrimination between classes [5-7]. Rao [5] shows that the classifying ability of features depends upon Δ_p^2, the true Mahalanobis-squared distance between the classes using p features and that the power of discrimination will increase if p+q features are used instead of p features only when $\Delta_{p+q}^2 - \Delta_p^2$ is of a certain order of magnitude. Young [6] shows this effect for Gaussian features and plots the probability of error $p(E/N,R_i)$ as a function of number of features and cluster radius R_i, a statistically defined neighbourhood around the cluster corresponding to class i (see Fig.1). More recently Van Ness and Simpson [7] considered the basic question of how many features should be used for a particular discriminant algorithm, given fixed number of labeled samples on which training data is obtained for each class. They consider Gaussian populations and five algorithms namely linear discrimination with unknown means and known covariances, linear discrimination with unknown means and unknown covariances, quadratic discrimination and two non-parametric Bayes-type algorithms and find the increase in Δ^2 necessary to justify increasing the dimension of the observation vector for specified classification accuracies and specified sizes of training data sets.

Earlier studies [1,3] confirmed that no single feature is sufficient to make V/UV/S classification decision. On the other hand, considerations of computational time and accuracy of recognition require limiting the number of features. The problem of finding the optimum feature set is not difficult as the number of features is only five. The exhaustive search of all possible feature subsets yields the necessary optimal subset.[8]. It is necessary to try the 31 combinations by direct or indirect methods. The direct method involves using each possible subset in a recognition scheme and evaluating the probability of misrecognition while the indirect methods use some distance measure (cluster radius, divergence, Bhattacharya distance, etc) to evaluate feature subset. For the present problem of V/UV/S classification indirect methods have shown that 3 features are adequate. In Appendix are given the various cluster radii [6] for the data of Atal and Rabiner [1, Table 2]. It can be seen that N_z, E_s, E_p and N_z, E_s, C_1 are the best 3 feature subsets on the basis of cluster radius. The latter set was chosen becuase of the computational simplicity.

B Decision Algorithm

It is assumed that features for each class are from a multidirectional Gaussian distribution with known mean m_i and covariance W_i, where i=1,2,3 correspond to V, UV and S. The decision criterion is [1]: Decide class i if $d_i(x) = \frac{1}{2}(x-m_i)'W_i^{-1}(x-m_i) \leq d_j(x)$ for all $i \neq j$, where x is the feature vector from the sample to be classified.

C Variable Decision Space Approach

A decision scheme based on the concept of variable decision space [9] gives further improvement in classification accuracy. In conventional classification methods, description of classes are in the form of probabilistic expressions which depend on the same features for every class. Such methods, may be called the constant decision space methods, do not have proper mechanisms for selecting from the original set of variables, the subsets which are most suitable for describing each class. Also, to decide class membership, knowledge of the values of all variables describing an object of each class is required. In the present approach, only the most appropriate features characterizing each class are involved in decision making. A scheme based on this approach for V-UV-S classification is shown in Fig.2. In this scheme the test pattern vector is first tested for voiced class using only two features E_s and N_z. If d_1, the Mahalonobis-squared distance of the test pattern vector from voiced class mean pattern vector is minimum, the test segment is termed as voiced. If d_1 is not the minimum then the testing continues and the test vector is tested for silence using the feature C_1 alone. If d_3 the distance of the test pattern vector from the silence class mean pattern vector is minimum, then the test segment is assigned to silence class, otherwise it is assigned to unvoiced class.

III Experimental Results

We shall present in this section the experimental results of V/UV/S classification scheme using only 3 features E_s, C_1 and N_z. The classifier is designed on the basis of a training set consisting of utterances of 3 male speakers (VVS, DVG and HSC). The utterances used were: "Chester Bowles", "Justice", "Six Chickens" and "Watch Them". The classifiers are tested by an independent test set with utterances of 2 male and 1 female speaker (DVG, TVA and MTN). The test set utterances were "Hotch Potch", "Miss King", "She Sells" and "Cross Vote". The mean and covariance for the three classes is shown in Table 1. The classification results are shown in Table 2, for the training set and in Table 3 for the test set. The results demonstrate the usefulness of the proposed techniques.

Conclusions

Automatic segmentation of speech is an important first stage in many communication and speech and speaker recognition problems. It is therefore necessary to achieve this with minimum possible computational effort without degradation in performance. The main contribution of this paper is in this direction. The limitation mentioned by Atal and Rabiner [1] regarding the suitability of the algorithm for particular recording conditions still exists.

TABLE 1

Typical Means and Covariance Matrices for the Three Classes (Three Speakers used in Training set)

	E_s	N_z	C_1
1) Voiced (i=1)			
Mean	43.68	15.19	0.87
Covariance	224.07	10.94	-0.21
Matrix	10.94	25.92	-0.17
	- 0.21	- 0.17	0.003
2) Unvoiced (i=2)			
Mean	13.34	60.21	0.29
Covariance	50.11	- 12.70	-0.48
Matrix	- 12.70	355.01	-6.10
	- 0.48	- 6.10	0.14
3) Silence (i=3)			
Mean	0.78	19.43	0.95
Covariance	0.59	3.48	-0.0085
Matrix	3.48	63.38	-0.15
	- 0.0085	- 0.15	0.00081

TABLE 2

Confusion Matrix for the Three Classes for Speech Data in Training Set

a) Actual Class		V	UV	S
Identified as	V	179	1	0
	UV	3	106	4
	S	0	0	40
Total		182	107	44
b) Actual Class		V	UV	S
Identified as	V	179	1	0
	UV	3	105	3
	S	0	1	41
Total		182	107	44
c) Actual Class		V	UV	S
Identified as	V	-	-	-
	UV	-	106	2
	S	-	0	42
Total		-	106	44
d) Actual Class		V	UV	S
Identified as	V	179	1	0
	UV	3	106	2
	S	0	0	42
Total		182	107	44

a) All the three features E_s, N_z and C_1 have been used.
b) Only two features E_s, N_z have been used.
c) Only one feature C_1 has been used for UV/S classification.
d) Variable decision space scheme has been used.

TABLE 3

Confusion Matrix for the Three Classes for Speech Data in Test Set

a) Actual Class		V	UV	S
Identified as	V	160	0	3
	UV	11	103	12
	S	0	0	35
Total		171	103	50
b) Actual Class		V	UV	S
Identified as	V	165	0	0
	UV	6	101	13
	S	0	2	37
Total		171	103	50
c) Actual Class		V	UV	S
Identified as	V	-	-	-
	UV	-	99	2
	S	-	4	48
Total		-	103	50
d) Actual Class		V	UV	S
Identified as	V	165	0	0
	UV	5	99	2
	S	1	4	48
Total		171	103	50

a) All the three features E_s, N_z and C_1 have been used.
b) Only two features E_s, N_z have been used.
c) Only one feature C_1 has been used for UV/S classification.
d) Variable decision space scheme has been used.

References

1. B.S. Atal and L.R. Rabiner, 'A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition', IEEE Trans. Acoustics, Speech and Signal Proc., Vol. ASSP-24, pp. 201-212, June 1976.
2. L.R. Rabiner and M.R. Sabur, 'Some preliminary experiments in the recognition of connected digits', IEEE Trans. Acoustics, Speech and Signal Proc., Vol. ASSP-24, pp. 170-182, April 1976.
3. L.J. Seigel and K. Steiglitz, 'A pattern classification algorithm for the voiced/unvoiced decision', Conference Record, IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, pp. 326-329, April 1976.
4. L. Kanal, 'Patterns in pattern recognition: 1968-1974', IEEE Trans. Information Theory, Vol. IT-20, pp. 697-722, Nov. 1974.
5. C.R. Rao, 'Covariance adjustment and related problems in multivariate analysis', in Multivariate Analysis - I, Edited by P.R. Krishnayya, Academic Press, New York, 1966, pp. 87-103.
6. I.T. Young, 'The prediction of performance in multiclass pattern classification', Proceedings of the Second International Joint

Conference on Pattern Recognition, pp.56-57, August 1974.

7. J.W. Van Ness and C. Simpson, 'On the effects of dimension in discriminant analysis', Technometrics, Vol.18, No.2, pp.175-187, May 1976.
8. C.H. Chen, 'On information and distance measures, error bounds and feature selection', Information Sciences, Vol.10, No.2, pp. 159-173, 1976.
9. R.S. Michalski, 'A variable decision space approach for implementing a classification system', Proceedings of the Second International Joint Conference on Pattern Recognition, Copenhagen, pp.71-77, August 1974.

APPENDIX

Table A1: Cluster Radii for Various Feature Subsets (Data of Atal and Rabiner [1])

No. of features	Features	Voiced	Unvoiced	Silence
	N_z	0.6452	0.8438	0.6452
	E_s	1.3551	1.1711	1.1711
1	C_1	0.6566	0.7765	0.6566
	a_1	1.6842	0.0925	0.0925
	E_p	2.1731	0.0711	0.0711
	N_z, E_s	3.0271	1.3154	1.3154
	$N_z, 1$	0.7178	0.9037	0.7178
	N_z, a_1	2.5407	1.3542	1.3542
2	N_z, E_p	2.1812	0.8703	0.8703
	E_s, C_1	3.116	1.2751	1.2751
	E_s, a_1	1.9190	1.1852	1.1852
	E_s, E_p	2.6355	1.7448	1.7448
	C_1, a_1	2.7962	1.3667	1.3667
	C_1, E_p	2.2588	0.7974	0.7974
	a_1, E_p	2.6134	0.1435	0.1435
	N_z, E_s, C_1	3.5440	1.3724	1.3724
	N_z, E_s, a_1	3.0622	1.6337	1.6337
3	N_z, E_s, E_p	3.6664	1.7998	1.7988
	N_z, C_1, a_1	2.8078	1.3926	1.3926
	N_z, C_1, E_p	2.3045	0.9329	0.9329
	N_z, a_1, E_p	2.8915	1.4178	1.4178
	E_s, C_1, a_1	3.4180	1.6036	1.6036
	E_s, C_1, E_p	4.0436	1.8025	1.8025
	E_s, a_1, E_p	2.7196	1.7910	1.7910
	C_1, a_1, E_p	3.0099	1.4391	1.4391
	N_z, E_s, C_1, a_1	3.8430	1.6534	1.6534
	N_z, E_s, C_1, E_p	4.0463	1.8844	1.8844
4	N_z, E_s, a_1, E_p	3.6676	2.1948	2.1948
	N_z, E_s, a_1, E_p	3.0268	1.4651	1.4651
	E_s, C_1, a_1, E_p	4.1610	2.2301	2.2301
5	N_z, E_s, C_1, a_1, E_p	4.1610	2.2738	2.2738

The cluster radius (R_i) gives a statistically defined cluster centre corresponding to class i.

$$R_i = \min_{j \neq i} \frac{d_{ij}\, d_{ji}}{d_{ij}+d_{ji}}$$

where $d_{ij} = [(\mu_j-\mu_i)' \sum_i^{-1} (\mu_j-\mu_i)]^{\frac{1}{2}}$ and the parameters of the i^{th} class belong to a multivariate Gaussian distribution with mean μ_i and covariance matrix Σ_i.

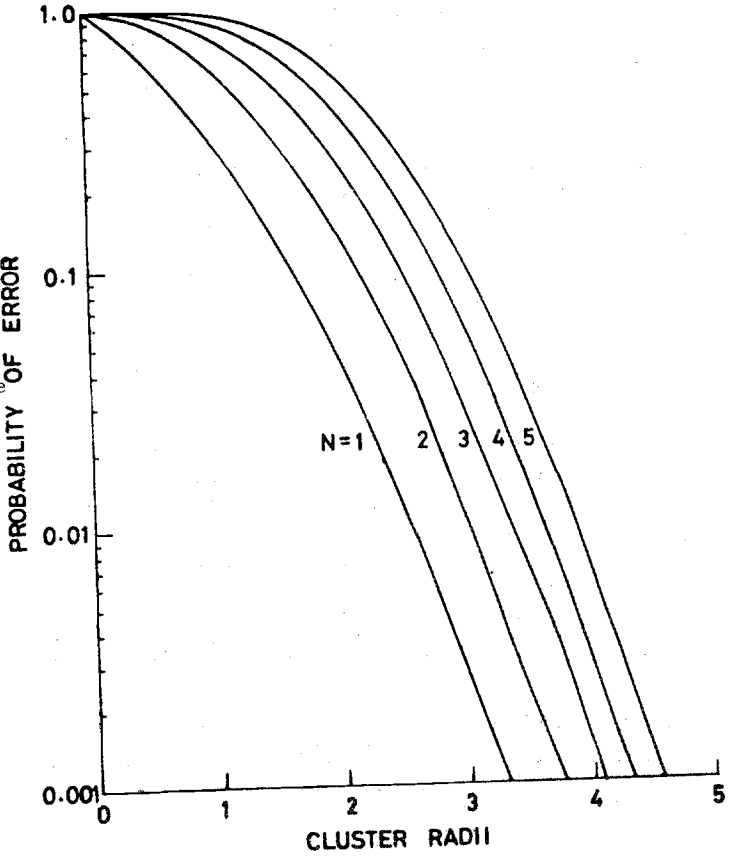

Fig.1. Probability of error Vs. Cluster radius

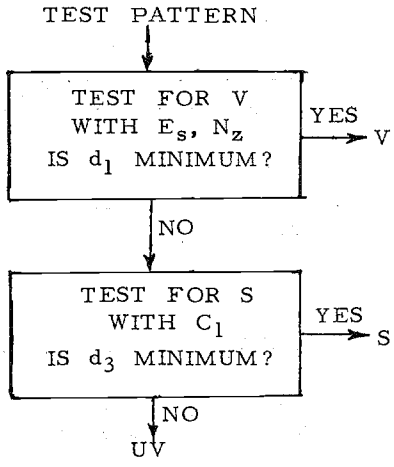

Fig.2. Variable decision space scheme for V/UV/S classification