STUDIES ON SPEAKER RECOGNITION USING A FOURIER ANALYZER SYSTEM

B.Yegnanarayana, V.V.S.Sarma and D.Venugopal

Indian Institute of Science
Bangalore 560012

ABSTRACT

In this paper, we describe studies made on designing speaker recognition schemes using an interactive signal processing facility. The facility consists of a HP 2100S minicomputer based Fourier Analyzer System.This facility accepts speech input and provides a display of results at various stages of recognition procedure. Such a system permits the use of large design and test data sets with consequent advantages in performance evaluation. We describe attempts on isolating better features purely from speaker recognition point of view. Feature selection criteria, choice of code words, design of classifiers and performance assessment are discussed.

I INTRODUCTION

It is wellknown that design of any pattern recognition scheme is a highly interactive process[1]. In speaker recognition literature most of the studies deal with flow of data and decisions in one direction i.e., from the input pattern environment to the classifier output which is normally the end result we seek. The purpose of this paper is to show that a signal processing facility built around a minicomputer, which accepts speech input directly and provides display of results at various stages, provides the necessary interactive capability for designing speaker recognition schemes. Even though a little constrained by the small memory size, such a facility permits use of large design and test data sets with consequent advantages in performance evaluation. We describe studies made on speaker recognition schemes with special reference to feature selection, choice of code word and performance assessment. Finally we describe a speaker recognition scheme implemented on a minicomputer based signal processing facility. The scheme which uses a short code word, a long term average feature and a minimum distance classifier is shown to display very good recognition capability in a practical environment.

II BASIC SPEAKER RECOGNITION SCHEME

1. Signal Processing Facility. The basic recognition scheme studied in this paper is simulated on a dedicated interactive signal processing facility. The system is a Hewlett Packard 5451B Fourier Analyzer built around HP2100S microprogrammable minicomputer with 16K memory of 16 bit words. In Fourier mode, through keyboard commands, the system permits analog data inputs, data manipulation functions, signal processing functions and arithmetic operations. Keyboard programming facility allows a sequence of these operations to be performed automatically,without any software. The system operates in block mode with block sizes 2^N for N = 6,7,...64. Thus the smallest block size is 64 and the largest is 4096. Short segments of speech can be entered directly through the system's A/D converter. Input data and results are displayed on a storage scope which provides the required interactive capability.Throughout this study a sampling rate of 10 KHz is used.

2. Pattern Environment. It has been observed that background noise is the single most important factor that affects the reliability and repeatability of speech and speaker recognition systems[2]. The studies reported here are made by producing speech in the computer room into a shure microphone placed at a distance of about 5 cm from the mouth of the speaker. The overall noise level is about 60 dB in A weighting and 68 dB in C waiting. Design and testing is thus done in a real-life situation.

3. Feature Selection. Feature selection is made on the basis of following considerations : (i) the feature vector should be good from speaker recognition point of view and should be computed easily from the signal processing functions available through keyboard commands, (ii) the feature vector should be normalized with respect to speech information, (iii) the feature vector should be extracted over sufficiently long segments of speech so that environmental factors such as background noise have less effect on it unlike the features used for representing speech[3]. The feature selected in the present study is normalized autocorrelation function of the test utterance. The first 64 coefficients of the autocorrelation function are chosen as they reflect the gross spectral features of the long term spectrum of the utterance. It is known that such sp-

ectral features are more speaker dependent than temporal variations of parameters extracted from short segments of speech[4].

4. Selection of code word. Considerations in the choice of code word are the following : (i) the utterance should be short for ease of computation on the signal processing The limited memory of the system permits entry of about 4 sec. of speech data of 10 KHz sampling rate. (ii) The feature vector extracted from the word should have sufficient inter-speaker variability, (iii) For minimum intra-speaker variability the word should be such that it provides minimum flexibility in changing the manner of pronouncing it from trial to trial, (iv) Scope for pitch variations in different repetitions should be minimized. These considerations led to the choice of a single short code word. Studies with different code words are presented in a latter section. Results for the word "mum" are discussed in this section. The choice of this word is governed by the fact that it is predominantly nasal and nasal sounds are known to provide good speaker recognizing features[5].

5. Reference pattern. The reference pattern for each speaker is obtained by collecting 8 repetitions of the word "mum" from each speaker, computing the normalized autocorrelation for each utterance and taking the average.

6. Classification Scheme. Histogram plots of Euclidian distances between test feature and reference vectors show that minimum distance (Euclidian) classifier provides adequate accuracy of classification. The distance is computed as follows:

$$D_i = \sum_{K=1}^{d} (x_K - m_K^{(i)})^2$$

where d = 1,2,...,64 is the chosen dimension of the feature vector, x_K is the Kth component of the feature vector of the test utterance and $m_K^{(i)}$ is the Kth component of the reference vector of speaker i i.e., class C_i. The decision rule is

$$x \in C_i \quad \text{iff } D_i < D_j \quad \text{for } j=1,2,\ldots,M,\ j \neq i,$$

where M is the number of speakers for which the system is designed.

7. Results. The prototypes of 8 speakers are are stored in appropriate blocks of size 64 and the scheme is tested by independent test set. Each of the speakers has been asked to utter the code word once in each turn and the decision is noted. The recognition decision is displayed on the scope by initials of the speaker. The result of one such experiment is shown as confusion matrix in Table 1. The dimension of the feature vector d is chosen as 32 in this experiment.

TABLE-1. Confusion Matrix for an Eight Speaker Recognition Experiment.

Speaker recognized as / True	BYN	VVS	TVA	DVG	TKR	BNP	HSC	HMD
BYN	10							
VVS		10						
TVA			9					1
DVG				9				1
TKR					10			
BNP		1	1			8		
HSC						1	9	
HMD								10

III DIMENSION OF FEATURE VECTOR

In all pattern recognition problems there is a continuous effort to reduce the dimension of feature vector. This is not only due to obvious reduction in computational effort but also because of the performance degradation if more than required number of features are used. Studies are made to determine which of the first sixtyfour coefficients of the normalized autocorrelation function contribute significantly to the speaker discrimination ability of the scheme. For speaker identification good features are those for which over a number of repeated utterances the inter-speaker variation is large and the intra-speaker variation is small. For the purpose of evaluating the features a statistic 'S' defined as the ratio of the average inter-speaker variation and the average intra-speaker variation is used.

$$S = \frac{\frac{1}{M}\sum_{j=1}^{M} (\mu_{jL} - \bar{\mu}_L)^t (\mu_{jL} - \bar{\mu}_L)}{\frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N} (X_{ijL} - \mu_{jL})^t (X_{ijL} - \mu_{jL})}$$

where X_{ijL} is the L-dimensional feature vector corresponding to the ith utterance of the jth speaker.

μ_{jL} is the L-dimensional mean feature vector corresponding to the jth speaker and is given by

$$\frac{1}{N}\sum_{i=1}^{N} X_{ijL}$$

μ_L is the L-dimensional mean feature vector averaged overall speakers and utterances and is given by

$$\frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M} X_{ijL}.$$

Obviously the feature set that has a larger S contributes more to speaker discrimination than a feature set with lower S. The 64 autocorrelation coefficients are divided into sets of 8 features each (for example 0-7, 8-15, 16-23, 23-31,...) and for these sets taken one at a time and two or more at

a time the statistic S is computed as shown in Table-2. The computation is done for sixteen repetitions of the code word by each of the six speakers. It can be seen that a feature set consisting of coefficients in

Feature set	0-7	0-15	0-23	0-31	0-63	8-15
The statistic S	1.24	2.43	2.36	2.00	0.96	2.93
Feature set	8-23	8.31	15-23	15-31	23-31	
The statistic S	2.54	2.18	2.29	1.81	1.54	

TABLE-2. The S-statistic for different feature sets (code word : MUM)

the range 8-31 seem to display good speaker discrimination compared to the coefficients in the range from 0-7 or 32-63. The poor performance of feature set consisting of values in the range 32-63 may be due to pitch variations between repetitions which are reflected in that region of autocorrelation function.

IV CHOICE OF CODE WORDS

Using the first 32 coefficients of the normalized autocorrelation function as the components of feature vector different words are studied for their speaker discrimination abilities. The statistic S defined in the previous section is used as basis for comparison. A word with larger S is a better code word than one with a smaller S. Six words namely CHUM, BUMPER, SMALL, HUG, SING, SUNG have been used for this study. The values of S are computed using eight repetitions of t the words by each of eight speakers and the results are given in Table 3.

TABLE-3. The S-statistic for different code words

Code word	SMALL	BUMPER	SUNG	CHUM	SING	HUG
The statistic S	0.830	1.170	1.414	1.432	1.475	2.217

The table shows that the word HUG has highest speaker discrimination capability and the word SMALL the lowest. This study throws light on the choice of suitable code words. It is interesting to note that while producing the word HUG the vocal tract is nearly stationary except for the abrupt change due to voiced stop consonant. The utterance is short and hence there is very little scope for intra-speaker variability.

V. FEATURE SET FOR SPEAKER VERIFICATION

In a speaker verification system the test feature vector is compared with the reference feature vector of the speaker corresponding to the particular label under which he wants to be verified. The speaker is accepted or rejected depending on whether the distance between test and reference vectors falls within the prescribed threshold or not. There can arise two types of errors in this problem. In the first one a speaker wanting to be verified under his own label is rejected and in the second type the speaker being verified under a false label is accepted. Sixteen utterances by each of six speakers are considered in this study. From the sixteen repetitions by each speaker the sixteen distances of feature vector from his reference vector are computed. The largest distance of each speaker from his own reference is chosen as threshold thereby eliminating the type 1 error. The total number of type 2 errors resulting when sixteen repetitions of each of the remaining five speakers are tested on the system under the label of a given speaker are given in Table 4 for different feature sets. The results obtained using a threshold that allows a maximum of one of type-1 errors are given in Table 5. It can be seen that feature sets consisting of coefficients in the range 0-23 or 0-31 give better speaker verification performance.

TABLE-4. Total Number of Type-2 Errors (max. 80) for a speaker verification system (Type-1 error is completely eliminated code word: MUM)

Feature set	0-7	0-15	0-23	0-31	0-47	0-63
Speaker						
DVG	38	15	3	7	9	61
VVS	52	23	10	18	15	27
BYN	13	3	5	3	0	0
TVA	24	2	0	0	3	7
HSC	61	39	11	3	5	0
BMP	63	45	47	34	48	73

TABLE-5. Total Number of Type-2 Errors (max. 80) in a speaker verification system. (Max. one type 1 error is permitted. code word: MUM)

Feature set	0-7	0-15	0-23	0-31	0-47	0-63
Speaker						
DVG	22	8	3	7	9	48
VVS	51	15	3	16	5	26
BYN	12	0	0	0	0	0
TVA	19	2	0	0	2	2
HSC	7	16	2	2	3	0
BNP	44	43	24	34	48	71

VI SEQUENTIAL SPEAKER IDENTIFICATION

An improvement in the performance of speaker identification can be achieved by using a string of code words presented in a sequential manner to the system instead of using a single code word. In this scheme a given number (say m) of speakers from to-

tal population are selected based on the nearest distances of their reference vectors from test vectors for a code word. If a set S_1 of m speakers are chosen from the first code word and another set S_2 of m speakers are chosen from the second code word the intersection of S_1 and S_2 should give the speaker to be identified. In case the intersection contains more than one speaker, the speaker with average minimum distance is identified as true speaker. More than two words also can be used in this procedure.

An experiment is conducted using the code words 'mum' and 'nun' on sixteen speakers using the first 32 coefficients as a feature vector. A set of nearest 4 speakers are selected for each word. Result of eight trials of the experiment is illustrated in Table 6. Table 7 gives the total number of errors obtained when the code words are used separately and in a sequential manner for 16 speakers in 108 trials. The results indicate the significant improvement in performance of speaker recognition scheme when two code words are used in a sequential mamner.

TABLE-6. Formation of Sets in a Sequential Speaker Recognition Experiment for Eight Test-Trials.(for speaker 2)

Test trial No.	Set of speakers for MUM	Set of speakers for NUN	Speaker identified as
1	2, 9, 4, 6	2, 8, 13, 16	2
2	9, 2, 6, 3	2, 8, 13, 11	2
3	6, 2, 9, 3	2, 8, 13, 6	2
4	9, 6, 2, 3	6, 3, 2, 5	6
5	9, 6, 2, 3	2, 8, 13, 6	2
6	2, 9, 4, 3	2, 6, 3, 8	2
7	2, 9, 3, 4	6, 2, 3, 9	2
8	2, 6, 9, 3	2, 8, 6, 12	2

TABLE-7. Total number of Errors in a 16 speaker Recognition Experiments (Total Number of Trials 108)

Cord word.	MUM alone	NUM alone	Using MUM and NUN in a sequentialmanner
Total Number of Errors	27	24	7

VII CONCLUSIONS

Studies made on a simple speaker recognition scheme simulated on a minicomputer-based signal processing facility are presented in this paper. The main conclusions are : (i) good recognition capability can be obtained with a single carefully selected short code word using normalized autocorrelation function as a feature. This may be compared with human recognition capability of recognizing a speaker where the duration of utterance plays an important role, (ii) when the feature selection problem is systematically approached an extremely simple minimum distance classifier is adequate for speaker recognition in small populations, (iii) speaker recognition using a string of code words in a sequential manner provide high recognition performance, and (iv) an interactive signal processing facility offers considerable flexibility in the design and performance evaluation of automatic speaker recognition schemes.

ACKNOWLEDGEMENTS

The authors wish to thank Prof.B.S.Ramakrishna for his interest and encouragement and their colleagues T.V.Ananthapadmanabha, H.M. Dante and G.R. Dattatreya for their help in conducting the experiments.

REFERENCES

1. L.Kanal, Patterns in pattern recognition, IEEE Trans. Information Theory, IT-20, (6), 697 (1974).
2. D.R. Reddy, Speech recognition by machine, Proc. IEEE, 64 (4), 501 (1976).
3. B. Yegnanarayana, Effect of noise and distortion in speech on parametric extraction, Proc. IEEE First International Conference on Acoustics, Speech and Signal Processing, Philadelphia, Pa., 336 (1976).
4. H. Hollien, W. Majewski and P. Hollien, Speaker identification by long term spectra under normal, stress and disguise conditions, J. of Acoustical Society of America, 55, (S 20), (1974).
5. I. Pollack, J.M. Pickett and W.H. Sumby, On the identification of speakers by voice, J. of Acoustical Society of America, 26, 403 (1954).