An F-Ratio Based Optimization Technique for Automatic Speaker Recognition System

Goutam Saha, Sandipan Chakroborty, Suman Senapati

Abstract—**Speaker recognition needs an efficient feature extraction process and appropriate speaker model developed from these features. The work uses Fisher's F-Ratio to find discriminative ability of each individual coefficient in any feature extraction algorithm. The reduced speaker model is developed by eliminating low performing coefficients without degrading recognition ability. The result is presented on Artificial Neural Network(ANN) based model for text-dependent recognition with two different phrases.**

I. Introduction

Personal identity verification is an important requirement for controlling access to protected resources. The identity can be verified by the Biometric features like fingerprints, retinal pattern, voice, hand geometry, handwriting, etc. The biometric attributes differ from person to person and recognition system identify different individuals based on these differences. Over last three decades several recognition systems [1] are employed by using different kind of biometric traits for applications like secure access to buildings, computer systems, laptops, cellular phones, and ATMs. Each system has pros and cons relative to accuracy and deployment. But as far as simplicity and cost are concerned Automatic Voice Recognition or Automatic Speaker Recognition(ASR) [2] system is one of the better recognition systems in today's world. For telephone based applications, there is no need for special signal transducers or networks to be installed at application access points since a telephone gives one access anytime from anywhere. Even for non-telephone applications, soundcards and microphones are low cost alternative and readily available. The speaker recognition area has a long and rich scientific basis with over 30 years of research, development and evaluations.

Automatic Speaker Recognition(ASR) involves recognizing a person from his spoken words. The goal is to find a unique voice signature [3] to discriminate one person from another. A speaker recognition system consists of two distinct modules, a Speech Parameterization Module and a Statistical Modeling Module. Speech Parameterization Module is basically the feature extraction ([4]-[5]) process by which speaker or speech specific characteristics can be obtained. The output of these Speech Parameters are fed to Statistical Modeling Module to produce a speaker model.

The effectiveness of the speaker verification depends mainly on the accuracy of discrimination of speaker models developed from acoustic feature methods. Each feature method comprises of several coefficients, however all the coefficients may not be equally good for discrimination purpose. Also, redundancy within the coefficients increases modeling complexity without improving discrimination performance. Thus there is a need in ASR application for optimization rules by which time, computational complexity, storage requirement both in feature extraction and modeling schemes can be minimized. Optimization in a feature space is earlier attempted in [6], which is based on heuristic search, among a set of potential features, for the feature subset that gives the minimal experimental Equal Error Rate. The work proposes a technique where coefficients extracted from a feature method is subjected to Fisher's F-Ratio test ([7]-[8]). The higher F-Ratio score of a coefficient represents higher discrimination ability of that coefficient. The optimization is done by discarding poorer coefficients, and feeding them to an ANN based speaker model. The work shows application of this technique with two popular feature extraction algorithms, Linear Predictive Coefficients(LPC) [9] and Mel Frequency Cepstral Coefficients(MFCC) ([10]-[11]). However, any other feature method and speaker model can be used for this purpose. The result is presented for text-dependent speaker recognition on two different phrases with 30 speakers.

II. Theoretical Background

A. F-Ratio

F-Ratio [7] is a statistical measure in the analysis of variance where multi-cluster data are available(Fig. 1). If there are k number of clusters, and if each cluster consists of n number of data points then

$$\mathrm{F-Ratio} = \frac{\text{Variance of means between the clusters}}{\text{Average Varaince within the clusters}} \tag{1}$$

If x_{ij} is an i^{th} element of j^{th} class then the mean of the j^{th} class μ_j can be expressed mathematically as,

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \tag{2}$$

The mean of all μ_j s is called the global mean of the data and can be expressed as $\overline{\mu}$. Analytically one can write

$$\overline{\mu} = \frac{1}{k}\sum_{j=1}^{k} \mu_j \tag{3}$$

The authors are with the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Khragpur, India, Kharagpur-721 302. Email: gsaha@ece.iitkgp.ernet.in, sandipan@ece.iitkgp.ernet.in, speech_ece@rediffmail.com

0-7803-8909-3/04/$20.00 ©2004 IEEE

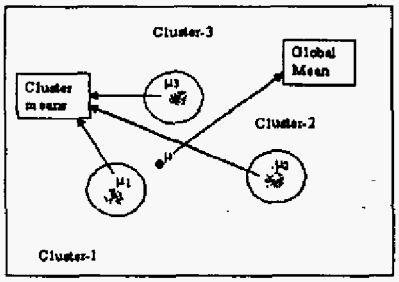

Fig. 1. Diagram showing multi-cluster data

So equation no. 1 for F-Ratio can be re-written as

$$\text{F}-\text{Ratio}=\frac{\frac{1}{k}\sum_{j=1}^{k}\left(\mu_j-\overline{\mu}\right)^2}{\frac{1}{k}\sum_{j=1}^{k}\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}-\mu_j\right)^2} \qquad (4)$$

F-Ratio will increase if the clusters move away from each other or the clusters in their positions shrink. Speaker recognition is also a multi-cluster data analysis problem where data represent some parameter value of different individuals calculated by some feature method. Depending on these parameter values, F-Ratio is calculated per feature(coefficients) in a feature method.

B. Linear Predictive Coefficients(LPC)

Linear prediction(LP) analysis [9] of speech is historically one of the most important speech analysis techniques. LP model is based on the source-filter model which is constrained to be an all-pole linear filter. Here the linear prediction of the present sample is expressed as a weighted sum of past samples:

$$\widehat{S_n}=\sum_{i=1}^{p}a_i s_{n-i} \qquad (5)$$

and the all pole model is

$$H(z)=\frac{1}{1-\sum_{i=1}^{p}a_i z^{-i}} \qquad (6)$$

where, $\widehat{S_n}$ is the linear prediction of the present sample based on a weighted sum of past samples, S_n is the past sample(s), and $a_i(i=1,2,\ldots,p)$ are the prediction coefficients. We have used first 14 LPC coefficients to characterize speakers in the present work.

C. Mel-Frequency Cepstral Coefficients(MFCC)

The psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale ([10]-[11]). Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The mel-frequency scale is a linear frequency spacing below 1kHz and a logarithmic spacing above 1kHz. As a reference point, the pitch of an 1kHz tone, 40dB above the perceptual hearing threshold, is defined as 1000mels. Therefore we can use the following approximate formula to compute the mels for a given frequency f in Hz:

$$\text{mel}(f)=2595\cdot\log_{10}(1+f/700) \qquad (7)$$

One approach to simulate the subjective spectrum is to use a filter bank, spaced uniformly on the mel scale. That filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval. The modified spectrum of $S(\omega)$ thus consists of the output power of these filters when $S(\omega)$ is the input. The number of mel spectrum coefficients, K, is typically chosen less than 20. Note that this filter bank is applied in the frequency domain, therefore it simply amounts to taking those triangle-shape windows(Fig. 2) on the spectrum. A useful way of thinking about this mel-wrapping filter bank is to view each filter as an histogram bin(where bins have overlap) in the frequency domain. Finally the log mel spectrum is converted back to time. The result is called the mel frequency cepstrum coefficients(MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients(and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform(DCT). Therefore if we denote those mel power spectrum coefficients that are the result of the last step are $\widetilde{S_k}, k=1,2,\ldots,K$, we can calculate the MFCC's, $\widetilde{c_n}$, as

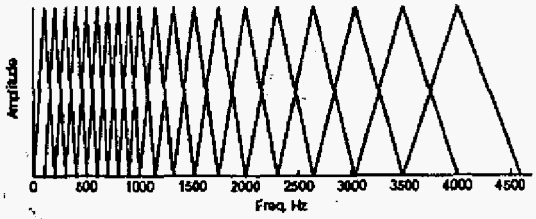

Fig. 2. Triangular Filter Banks for MFCC

$$\widetilde{c_n}=\sum_{k=1}^{K}\left(\log\widetilde{S_k}\right)\cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right] \qquad (8)$$

We have used first 18 MFCC coefficients discarding the d.c term to characterize speakers in the present work.

III. Method

A. Preprocessing of Speech Signal

The continuous speech signal is first digitized with a sampling frequency of 8kHz. After removing the silence periods it is divided into frames of 256 sample size with 50% overlap. Each frame is multiplied with a Hamming window function, $w(n)$, where

$$w(n)=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right),\ 0\le n\le N-1 \qquad (9)$$

This minimizes signal discontinuity at the edges and spectral distortion that arises from framing.

B. Feature Extraction

Features are extracted from the preprocessed signal framewise and averaged over all the frames of an utterance. For a feature method generating c no. of coefficients we have an array of c coefficients for each utterance of a speaker. Thus with a speaker database size of $m \times u(m =$ no. of speakers, $u =$ no. of utterances per speaker) we have $m \times u$ no. of coefficient array for use in speaker model.

C. Neural Network Structure

Artificial Neural Network employed for speaker modeling uses Multi-layer Perceptron(MLP) mechanism with Back Propagation Algorithm. MLP has been successfully used to solve many complex and diverse classification problems ([12]-[13]). In our case, the problem is to classify the speech sample feature vectors into several speaker classes. The network consists of an input layer, one hidden layer and an output layer similar to Fig. 3. Here, the number of nodes in the input layer equals the feature dimensions whereas number of nodes in output layer is same as the number of speakers to be identified. The number of nodes in the hidden layer is taken three times the number of the nodes used in the input layer but not less than no. of output nodes. The network uses nonlinear hyperbolic tangent activation function with training goal set at 0.01, learning rate 0.05 and a momentum term 0.9. The training is done separately for each feature method and each phrase with mutually exclusive training and test data set.

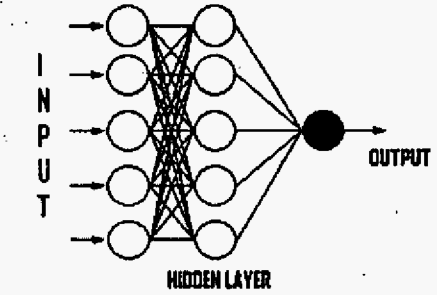

Fig. 3. A Neural Network with 5 input 5 hidden and 1 output node

IV. Results

Two separate tests are conducted for every feature method in which 30 speakers(22 male and 8 female) were asked to utter one combinational lock no.('24 – 32 – 75') and one code word('Indian Institute of Technology') 10 times each. A half of these i.e. 150 utterances are used for training and the other half (also 150 utterances) for testing.

Table I represents LPC coefficients arranged in order(descending) of their discrimination ability given by F-Ratio score for both the phrases. Same is done for MFCC coefficients in Table IV. Recognition performances of optimized models with lower no. of coefficients are presented in Table II and III for two phrases with LPC coefficients. The reduced models discard coefficients with lower F-Ratio score given in Table I e.g. a reduced model of 13 coefficients discards the lowest performer in Table I i.e. coefficient 12 for combination lock phrase and coefficient 13 for code word. Similar exercise with MFCC coefficients is presented in Table V and Table VI.

TABLE I

LPC F-Ratio

Sl. No.	24-32-75		Indian Institute of Technology	
	Coefficient No.	F-Ratio	Coefficient No.	F-Ratio
1	8	23.6897	8	8.3819
2	9	11.4402	4	6.5127
3	4	10.3775	7	5.7346
4	6	9.9244	1	5.3775
5	7	9.6318	9	4.6924
6	2	9.1157	10	4.3243
7	5	8.9098	6	4.1458
8	1	8.2543	2	3.9360
9	3	7.2059	14	3.8002
10	14	6.5441	3	3.6695
11	10	4.9052	12	2.4966
12	11	3.4716	11	2.4168
13	13	3.2487	5	2.0454
14	12	2.6254	13	1.7642

TABLE II

LPC '24-32-75'

Sl. No.	No. of Coefficients Taken	Network Structure	No. of Incorrect Identification	% of Incorrect Identification
1	14(all)	14-42-30	2	1.33
2	13	13-39-30	2	1.33
3	12	12-36-30	2	1.33
4	11	11-33-30	4	2.66
5	10	10-30-30	4	2.66
6	9	9-30-30	5	3.33
7	8	8-30-30	5	3.33
8	7	7-30-30	5	3.33
9	6	6-30-30	6	4.00

TABLE III

LPC 'Indian Institute of Technology'

Sl. No.	No. of Coefficients taken	Network Structure	No. of Incorrect Identification	% of Incorrect Identification
1	14(all)	14-42-30	2	1.33
2	13	13-39-30	2	1.33
3	12	12-36-30	5	3.33
4	11	11-33-30	7	4.66
5	10	10-30-30	7	4.66
6	9	9-30-30	7	4.66
7	8	8-30-30	7	4.66
8	7	7-30-30	8	5.33
9	6	6-30-30	11	7.33

We find recognition performance for the combination lock phrase with MFCC coefficients is the best having only 1 incorrect identification out of 150 for full 18 coefficient

TABLE IV

MFCC F-RATIO

Sl. No.	24-32-75		Indian Institute of Technology	
	Coefficient No.	F-Ratio	Coefficient No.	F-Ratio
1	9	22.9665	7	28.0005
2	7	21.9534	10	18.4123
3	2	20.7265	11	15.9539
4	8	15.6969	9	14.7164
5	1	12.1135	12	10.9175
6	6	11.1953	14	8.8498
7	13	11.0250	6	7.7895
8	15	9.7758	15	7.5846
9	11	9.7311	2	6.6756
10	10	9.5797	8	6.6700
11	12	9.1917	13	6.3666
12	14	8.7274	16	6.1185
13	5	6.2580	5	5.3978
14	4	5.5897	1	4.8324
15	16	4.1424	4	4.1182
16	17	3.5797	17	3.1648
17	3	2.8952	3	1.7223
18	18	1.6888	18	1.5079

TABLE V

MFCC '24-32-75'

Sl. No.	No. of Coefficients taken	Network Structure	No. of Incorrect Identification	% of Incorrect Identification
1	18(all)	18-54-30	1	0.66
2	17	17-51-30	1	0.66
3	16	16-48-30	1	0.66
4	15	15-45-30	1	0.66
5	14	14-42-30	1	0.66
6	13	13-39-30	3	2.00
7	12	12-36-30	4	2.66
8	11	11-33-30	5	3.33
9	10	10-30-30	6	4.00

model(18 – 54 – 30). The same performance is shown even when model is reduced to 14 coefficients(14 – 42 – 30). If little sacrifice in performance is acceptable one can go for further reduction as found in Table V. The result shows similar optimization is also possible for code word phrase and LPC method. Note that the combination lock phrase performs relatively better in ASR application than the codeword phrase for both the feature methods.

TABLE VI

MFCC 'INDIAN INSTITUTE OF TECHNOLOGY'

Sl. No.	No. of Coefficients taken	Network Structure	No. of Incorrect Identification	% of Incorrect Identification
1	18(all)	18-54-30	2	1.33
2	17	17-51-30	2	1.33
3	16	16-48-30	3	2.00
4	15	15-45-30	3	2.00
5	14	14-42-30	3	2.00
6	13	13-39-30	4	2.66
7	12	12-36-30	7	4.66
8	11	11-33-30	7	4.66
9	10	10-30-30	7	4.66

V. CONCLUSION

An F-Ratio based optimization technique is presented for ASR application that ranks the coefficients extracted from a feature method according to their discrimination abilities. The work shows its use in reduction of feature space and in turn the size of a speaker model. The result is presented for two feature methods on ANN based speaker model, however it is not limited to these and any feature method in conjunction with a speaker model can benefit from such optimization technique.

VI. ACKNOWLEDGEMENT

The work is partly supported by Indian Space Research Organization(ISRO), Government of India.

REFERENCES

[1] R. D.Luis-Garcia, Carlos Albarto-Lopez, O. Aghzout and Juan Ruiz-Azola, "Biometric identification systems", *Signal Processing*, vol.83, Issue.12, pp.2539-2557, Dec.2003.

[2] J. P.Cambell,Jr., "Speaker Recognition:A Tutorial", *Proceedings of The IEEE*, vol.85, no.9, pp.1437-1462, Sept.1997.

[3] L. G.Kersta, "Voiceprint Identification", *Nature*, vol.196, no.4861, pp.1253-1257, Dec.29, 1962.

[4] M. R.Sambur,"Selection of Acoustic features for Speaker Identification", *IEEE Trans. ASSP*, vol.ASSP-23, no.2, pp.176-182, Apr.1975.

[5] J. J.Wolf,"Efficent Acoustic Parameters for Speaker Recognition", *The Journal of the Acoustic Soceity of America*, vol.51, no.6(Part 2), pp.2044-2056, Mar.1971.

[6] D. Charlet and D. Jouvet, "Optimizing feature set for speaker verification", *Pattern Recognition Letters*, 18(1997), pp.873-879.

[7] S. Pruzansky and M. V Mathews, "Talker-Recognition Procedure Based on Analysis of Variance", *The Journal of Acoustical Society of America*, vol.36, no.11, pp.2041-2047, Nov.1964.

[8] B. S.Atal, "Automatic Recognition of Speakers from their Voices", *Proceedings of The IEEE*, vol.64, no.4, pp.460-475, Apr.1976.

[9] B. S.Atal, "Effectiveness of Liner Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification", *The Journal of Acoustical Society of America*, vol.55, no.6, pp.1304-1312, Jun.1974.

[10] S. B.Davis and P. Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continously Spoken Sentences", *IEEE Trans. ASSP*, vol.ASSP-28, no.4, pp.357-365, Aug.1980.

[11] J. C.Moore and B. R.Glassberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", *The Journal of Acoustical Society of America*, vol.74; no.3, pp.750-753, Sept.1983.

[12] N. P.Archer and S. Wang, "Fuzzy set representation of neural network classification boundaries", *IEEE Trans.Systems, Man and Cybernetics*, vol.21, no.4, pp.735-742, 1991.

[13] J. Sauvola, H. Kauniskangas and K. Vainamo, "Automated document image preprocessing management utilizing grey-scale image analysis and neural network classification", *Sixth International Conference on Image Processing and Its Applications*, vol.2, pp.14-17, Jul.1997.