FEATURE SELECTION USING GENETICS-BASED ALGORITHM AND ITS APPLICATION TO SPEAKER IDENTIFICATION

M.Demirekler

E.E. Eng. Department,M.E.T.U.,
06531,Ankara,Turkey.

A.Haydar

E.E. Eng. Department,E.M.U.,
Gazi Mağosa,Mersin 10, Turkey.

ABSTRACT

This paper introduces the use of genetics-based algorithm in the reduction of 24 parameter set (i.e the base set) to a 5,6,7,8 or 10 parameter set, for each speaker in text-independent speaker identification. The feature selection is done by finding the best features that discriminates a person from his/her two closest neighbors. The experimental results show that there is approximately 5% increase in the recognition rate when the reduced set of parameters are used. Also the amount of calculation necessary for speaker recognition using the reduced set of features is much less than the amount of calculation required using the complete feature set in the testing phase. Hence it is more desirable to use the subset of the complete feature set found using the genetic algorithm suggested.

1. INTRODUCTION

Speaker identification is a process of determining the identity of an unknown speaker among several speakers by comparing the input voice sample with the known voices and choosing the one that best matches the input voice sample. In order to increase the recognition rate, different algorithms and different features are used in the literature[1]. One of the commonly used feature set is a part of the cepstral coefficients. In this work, speech is represented by 12 LPC derived cepstral coefficients and 12 Δ-cepstral coefficients.

Optimal feature selection is one of the important problems and has been studied by several researchers[2][3]. Our aim here is to select the optimal features from 24 parameter set in the sense that the separability of a speaker from the two closest neigbors is maximized. In general, each speaker has a different optimal feature space [4]. The discriminating power of each parameter depends on the speaker identification algorithm used. This work is based on the novel speaker identification algorithm given in [5]. Two different sets of experiments were conducted. In a particular experiment of the first set, the number of parameters selected, i.e, the cardinality of the parameter set is fixed either to 5, or 6, or 7, or 8, or 10, however the parameter selected for each speaker may be different. In the second set, we remove this restriction. Hence, for the second type of experiments, for each speaker, not only the parameters selected may be different, but also the number of parameters selected (i.e the cardinality of teh parameter set) could vary. In the next section, the training algorithm that is used in speaker identification is explained.

2. TRAINING

In the training phase, for each speaker, exactly two Gaussian distributions i.e two mean vectors and two covariance matrices are calculated. The first mean vector, μ_i, and covariance matrix, Σ_i, are obtained using the data of the i^{th} speaker and the second mean, $\bar{\mu}_i$, and the covariance matrix, $\bar{\Sigma}_i$, are calculated using the collection of data from the rest of the speakers in the identification system. For convenience , we use the compact notation

$$M^i = (\mu_i, \bar{\mu}_i, \Sigma_i, \bar{\Sigma}_i) \qquad (1)$$

to indicate the model of the i^{th} speaker. By this method we obtain 2N number of mean vectors and covariance matrices when we have N speakers in the identification system. Any single feature vector , $x_j(k)$, the k^{th} vector of the j^{th} speaker, is classified according to the decision rule which is stated as;

$$(x_j(k) - \mu_i)^T \Sigma_i^{-1} (x_j(k) - \mu_i) -$$

$$(x_j(k) - \bar{\mu}_i)^T \bar{\Sigma}_i^{-1} (x_j(k) - \bar{\mu}_i) \begin{cases} > 0 & \text{select } \bar{s}_i \\ < 0 & \text{select } s_i \end{cases}$$

$$1 \leq i \leq N, \qquad 1 \leq j \leq N$$

Here s_i is the set that contains only the i^{th} speaker and $\bar{s}_i$ is the complement of this set i.e it contains all speakers except the i^{th} speaker. Any single feature if it is in class s_i and is rejected if it is in class $\bar{s}_i$. This process is repeated by the training data of each speaker in the identification system using the model of the i^{th} speaker. In particular , for the training data of speaker j, all feature vectors are classified according to decision rule given above by using model M^i and the value b_{ij} is calculated as follows

$$b_{ij} = \frac{\text{number of accepted feature vectors by model } M^i}{\text{total number of feature vectors of speaker j}}$$

This NxN matrix , B , represents the percentage of feature vectors of training data accepted by each speaker model and we normalize this matrix using diagonal elements. For the ideal case the resultant B matrix must be the identity matrix. Large deviations from the identity matrix shows close relationship between the speakers. Note that b_{ij} defined by the above procedure is the normalized value of the number of feature vectors of the j^{th} speaker which are classified as the i^{th} speaker's feature vector according to the i^{th} model . At this point, it may be argued that a proper selection of the feature set, so the feature vector, may decrease the similarity between the two speakers. So from now on the aim will be the optimum reduction of the feature set. For this purpose let us rewrite the following preliminary definitions. Let t_{all} denote an ordered set of features and z_{all} denote its cardinality. Let t be a subset of t_{all} , $t \subset t_{all}$ and z denote its cardinality. The optimal feature selection is based on the discriminative power of each feature. Discriminating a person from his two closest neighbours is aimed rather than discriminating it from all the other speakers in the set. A matrix similar to the matrix B described above is used to find the closest neighbours. Hence the argument of the maximization problem,i.e indices j and k that maximizes

$$\max_{j,k} \frac{\hat{b}_{ij} + \hat{b}_{ik}}{2} \qquad (2)$$

is selected as the two closest neighbours for model i. Note that $\hat{B} = [\hat{b}_{ij}]$ is used in Eq.(2) instead of the matrix $B = [b_{ij}]$. There are two basic differences between B and $\hat{B}$. The first one is that $\hat{B}$ is obtained like B but by using a new reduced vector set where each vector is a sub-vector of the previous full vector. So $\hat{B}$ depends on the selected subset t. The second one is that the rows of the matrix $\hat{B}$ is not normalized by its diagonal elements. The criterion that must be maximized to find the features that have the most discriminating power is written as

$$\max_{t_i \subset t_{all}} (\hat{b}_{ii} - \max_{\substack{j,k \\ j \neq k \neq i}} \frac{\hat{b}_{ij} + \hat{b}_{ik}}{2}) \qquad (3)$$

t_i is the set of features selected from t_{all} for the i^{th} speaker and $\hat{b}_{il}$ can be defined as follows

$$b_{il} = \sum_{k=1}^{T_l} u((x_l(k) - \hat{\bar{\mu}}_i)^T \hat{\bar{\Sigma}}_i^{-1} (x_l(k) - \hat{\bar{\mu}}_i) -$$

$$(\dot{x}_l(k) - \dot{\mu}_i)^T \hat{\Sigma}_i^{-1} (x_l(k) - \hat{\mu}_i))/T_l \qquad (4)$$

where T_l is the number of feature vectors used as the training data of l^{th} speaker and u is the unit step function. In Eq.(4) , $\dot{x}(k)$, $\dot{\mu}_i$, $\hat{\bar{\mu}}_i$, $\hat{\Sigma}_i$, $\hat{\bar{\Sigma}}_i$ are used instead of $x(k), \mu_i, \bar{\mu}_i, \Sigma_i$, and$\bar{\Sigma}_i$, respectively. Let us use the compact notation

$$\hat{M}^i = (\dot{\mu}_i, \hat{\bar{\mu}}_i, \hat{\Sigma}_i, \hat{\bar{\Sigma}}_i)$$

to indicate the model of the i^{th} speaker using the reduced set of features, t. Note that the aim of the maximization problem given by Eq.(3) is to increase the difference between the diagonal element of the new $\hat{B}$ matrix (obtained using the reduced feature vector) and the average of the two nearest off diagonal elements. The resultant $\hat{B}$ matrix obtained using the reduced set is normalized during the testing phase. This late normalization causes some kind of nonoptimality in the parameter selection.

For the second experiment set, where the number of parameters could also vary for each speaker, the criterion given in Eq.(3) is modified as

$$\max_{\substack{t_i \subset t_{all} \\ z_i \in \{5,6,7,8,10\}}} (\hat{b}_{ii}(t_i) - \max_{\substack{j,k \\ j \neq k \neq i}} \frac{\hat{b}_{ij}(t_i) + \hat{b}_{ik}(t_i)}{2}) \qquad (5)$$

Here, for speaker i, the number of feature is not fixed and can be selected from a set $S = \{5, 6, 7, 8, 10\}$ in such a way that a person is discriminated from his/her two closest neigbours. An exhaustive search of all possible t_i sets requires the evaluation of the training algorithm for $\binom{z_{all}}{z}$ times. For $z_{all} = 24$ and $z = 6$, this number is $\frac{24!}{6!18!} = 134596$ which shows the difficulty of such a search. To overcome this difficulty the genetics-based algorithm described below is used.
Our algorithm uses mutation to obtain new generations. The main idea here is to start with a feature vector of size 24 and choose only some entries of this complete feature vector to describe a speaker. Actually the so called "complete feature vector" is formed by concatenating 12 cepstral coefficients by 12 Δ-cepstral coefficients. It is believed that using only the most representative features, the better discrimination of a particular speaker from the others can be achieved. So

gorithm is to find these entries of the complete feature vector. For this purpose, a binary vector of size 24 is generated. An entry of this binary vector that contains '1' shows that the corresponding feature (i.e. the corresponding entry of the feature vector) is selected as feature that models the speaker.

To give a better axplanation of the genetics-based algorithm developed to reduce the feature space, the problem is simplified to a constant 'z' case ($z_i = z$ for all i). The algorithm basically uses mutation and can be defined as follows:

Step 1 (Initialization): Form a binary vector of size z_{all} which contains exactly z number of 1's. The positions of these 1's denote that the corresponding feature is selected for the reduced set. Call this vector $V_{initial}$. Set $V_{final} = V_{initial}$. Calculate the value of the objective function using $V_{initial}$, say MaxVal.

Step 2 (Mutation): Obtain a new binary vector, V, from $V_{initial}$ by interchanging a randomly selected '1' and '0' entries of $V_{initial}$. By this process M (M=25 in our application) number of new candidate sets for the reduced set are generated using $V_{initial}$.

Step 3 (Selection of a new $V_{initial}$): Calculate the objective function for these new M vectors and choose the one that has the maximum value, say MaxNew. Replace $V_{initial}$ with the vector that corresponds to value MaxNew. If MaxNew > MaxVal, replace V_{final} with the vector that corresponds to value MaxNew and equate MaxVal to MaxNew. Go to step 2 (elitist approach).

Iterate Step 2 and Step 3 for 24 times. V_{final} gives the selected subset t. Note that in this algorithm, to generate the new population, only mutation is used.

3.TESTING

Testing is performed after different number of features are selected for each person. From now on, the same symbol B will be used instead of the normalized $\hat{B}$ and B described in the above sections. In the testing phase, using each model, N fitness values are obtained for the test data and normalized using the normalizing constants of each model used in the normalization of B. This new vector is denoted by $\bar{c}_{test}$, which is very similar to a column of the matrix B. The N values, $\bar{c}_{test}(i)$, $1 \leq i \leq N$, are used to eliminate the candidate speakers one at a time till only one candidate is left. Elimination is based on the normalized matrix, B, and the vector $\bar{c}_{test}$. This elimination process is described below.

At the first step of the elimination process, speaker 1 is compared with speaker 2 and the winner of this step is compared with speaker 3. In general, the winner of the n^{th} step is compared with speaker n+1. For the comparison of the i^{th} speaker with the k^{th} speaker, we are using the following distance measure in order to decide which one is going to be eliminated.

$$((b_{ii} - \bar{c}_{test}(i))^2 + (b_{ki} - \bar{c}_{test}(k))^2)-$$

$$((b_{ik}-\bar{c}_{test}(i)^2+(b_{kk}-\bar{c}_{test}(k)^2) \begin{cases} > 0 & \text{choose speaker } k \\ < 0 & \text{choose speaker } i \end{cases}$$

The unknown speaker is identified as the final candidate.

4. EXPERIMENTAL RESULTS

The speech signal was sampled at a rate of 8 kHz, segmented into 22.5 ms nonoverlapping frames, preemphasized and Hamming windowed. A complete evaluation of the system is conducted for 15 male speakers selected randomly from the SPIDRE database. SPIDRE database contains continuous free speech recorded in four different telephone conversations. Two sets of experiments were conducted. In the first set, we have use the first three sessions(approximately 6000 feature vectors) for training and last session for testing. Each test utterance has a length of 11.25 sec. In the second experiment set, where the number of parameters, used for each speaker, could also vary, 3000 feature vectors that corresponds to 67.5 sec. are used for the training and each test utterance has a length of 4.5 sec.

First of all, text-independent speaker identification experiments for the case when the cardinality of the parameter set is same for all the speakers were performed. The results obtained for the cases when 5 ,6 ,7 ,8, 10 and all 24 parameters are used in the recognition system are given in Table I. Last column of Table I shows the number of test utterances of each speaker and the last row of Table I shows the overall recognition rates in percentage. Note that the recognition rates given for each speaker are in percentages.

Table II. Parameters selected for Spk. 2

z=5	1	2		4	5	6					
z=6	1	2		4	5	6	7				
z=7	1	2		4	5	6	7		13		
z=8	1	2		4	5	6	7	10	13		
z=10	1	2	3	4	5	6	7		13	14	21

Spk	Recognition rates						no of data
	5	6	7	8	10	24(all)	
1	100	99 3	100	100	99 3	100	134
2	100	100	100	100	100	93 7	127
3	100	100	100	100	100	100	81
4	98 8	67 9	65 4	60 5	48 2	80 3	162
5	100	100	100	100	100	100	66
6	100	100	100	100	100	100	99
7	100	100	100	100	100	100	66
8	61 5	61 5	63 1	61 5	82.8	68	122
9	86 7	93 3	90 7	98 7	96	98 7	75
10	64.8	81 7	64 8	83 1	69	62	71
11	85 1	87 8	87 8	86 5	89 2	90 5	74
12	86 6	90 7	89 7	87 6	89 7	92 8	97
13	84 8	98 3	79 7	78	76 3	76 3	59
14	100	98.4	98 4	95 1	95 1	95 1	61
15	68 6	79	58	69 4	61 3	67	124
Total rec	89	88 4	84 7	85 9	84 9	87 3	

Table II shows the reduced set of parameters selected for Speaker 2 with cardinalities 5 ,6 ,7 ,8 and 10. The first 12 parameters (1 to 12) correspond to 12 cepstral coefficients and the remaining 12 parameters (13 to 24) correspond to 12 Δ-cepstral coefficients. The last row shows the total recognition rates. Note that the parameters corresponding to numbers 1,2,4,5,6 are always selected and some additional parameters are included when the system is forced to select the larger number of parameters.

Table III. Recognition rates for SI (in %)

Sp	Rec rate (all)	Rec rate (reduced)	ntest	Features selected
1	100	100	140	1,5,8,10,11, 12,13,23
2	93.2	100	133	1,2,4,5,6
3	100	100	87	2,4,5,10,13
4	79.8	98 2	168	1,9,10,11,21
5	100	100	72	4,5,6,8,11
6	100	100	105	3,4,5,6,7,8
7	100	100	72	3,6,11,13,18
8	68	77.3	128	1,3,5,6,7,8, 11,17,21,22
9	98.8	91.4	81	1,2,3,4,5,7,9,16
10	61	89.6	77	1,3,4,5,14,16,18,20
11	90	88.8	80	4,5,6,9,11,16, 17,19,20,21
12	92 2	90.3	103	1,2,3,4,7,8
13	75.4	98.5	65	4,5,9,16,18,21
14	95 5	100	67	2,3,4,9,19
15	67 7	69 2	130	1,3,4,6,7
	87 3	92.9		

In Table III, we tabulated the recognition rates for both "the all parameter case" and "the reduced parameter case" for each speaker. The column denoted by 'ntest' shows the number of test utterances used for each speaker. The last column of Table III shows the features selected from the ordered set, t_{all}, to form the reduced sets which contain different number of features (5,6,7,8 or 10) for each speaker.

5. CONCLUSIONS

In this study, a genetics-based algorithm is used to seing power for each speaker in text-independent speaker identification.

Two different sets of experiments were performed. In the first set, in a particular experiment, the cardinality of the reduced set is fixed. We observe that when the feature space is extended, in general, some new parameters are added to the previously found ones. This fact is shown in Table II. The experimental results also show that the number of parameters necessary to obtain the highest recognition rate vary from speaker to speaker. In order to prove this, the second experiment set is performed in which the number of parameters selected for each speaker could also vary. The experimental results show that approximately 5% increase in the recognition rate is achieved when this selection is made. Table III shows that there is a dramatic increase in the recognition rates of some speakers (e.g. 30% for speaker 10). There are decreases in the recognition rates of only 3 persons, one is around 7% and the other two are approximately 2%. Also, since an average of 6.5 parameters are used in the testing phase, instead of 24 parameters, the amount of calculation necessary for speaker recognition using the reduced set is much less than the amount of calculation using the complete set of features(ratio is proportional with $6.5^2/24^2$). Although the experiment performed is not enough to draw some conclusions about the selection of "universally good" features, we may say that cepstral parameters 3-4-5 are selected more frequently than the others, which is an indication of their high discriminating power. Δ-cepstrals, when compared with cepstrals, seem to be less important in the average. However, as Speaker 10's or Speaker 13's scores indicate, their roles are unavoidable for some speakers.

6. REFERENCES

[1] D.A.Reynolds, "Experimental Evaluation of Features for Robust Speaker Identification",IEEE Trans. Spe., Audio Process.,Vol.2, No.4,pp.639-643,1994.

[2] Bidsaria H.B., "Least desirable feature elimination in a general pattern recognition problem ",

[3] Morgera S.M. and Datta L., " Toward a fundamental theory of optimal feature selection : Part I " , IEEE Trans. Acoust., Speech, Signal Process., Vol.PAMI-6, pp.601-616, 1984. Pattern Recognition, Vol.20, pp.365-370, 1987.

[4] Cohen A. and Froind I. , "On Text-Independent Speaker Identification Using a Quadratic Classifier With Optimal Features ", Speech Communication, Vol.8, No.1 , pp.35-44, March 1989.

[5]Haydar A., Demirekler M.,Nakipoglu B. and Tuzun B., " Text Independent Speaker Identification Using bayes decision Rule and Vector Quantizer ", 5^{th} International Conf. on Advances in Communication and Control, pp.56-60, COMCON5, 1995.