final report for

PHONE CLASSIFICATION USING LINEAR DISCRIMINANT ANALYSIS AND DECISION TREE APPROACHES

submitted to fulfill the semester project requirement for

EE 8990: Pattern Recognition

May, 5 1999

submitted to:

Dr. Nicolas Younan

Department of Electrical and Computer Engineering
413 Simrall, Hardy Rd.
Mississippi State University
Box 9571
MS State, MS 39762

submitted by:

Vishwanath Mantha, Yufeng Wu

Electrical and Computer Engineering Department
Mississippi State University
Box 9671
Mississippi State, Mississippi 39762
Tel: 601-325-8335
Fax: 601-325-8192
email: {mantha, wu}@isip.msstate.edu

ABSTRACT

The aim of this project is to perform phone classification on the Mel Frequency Cepstral Coefficients (MFCCs) generated by a speech recognition front-end. We have used two classification techniques, namely Linear Discriminant Analysis and Decision Trees. The former is a linear technique where as the latter is a nonlinear one. The motivation for the application of the nonlinear techniques was to observe some reduction in the misclassification error rate. Our experiments have shown that the decision tree approach does well on the training data but the performance degrades in case of the test data due to over-training.

INTRODUCTION

Linear Discriminant Analysis and Decision Trees are two well-founded, commonly used classification techniques. Since previous work shows that they have enjoyed success in many application domains, such as scenic beauty estimation, generation of proper noun pronunciation, etc., a logical progression is to apply them to classification of phonetic segments of speech data. Another motivation is that many existing methods do not achieve good performance when dealing with this particular classification problem. Also we hope to investigate some new functionalities of these two traditional methods.

Linear Discriminant Analysis (LDA)[1] helps us discriminate between different classes based on a linear classification rule. It typically uses a linear transformation which can either be implemented in a class-dependent or class-independent fashion. The effectiveness of LDA is restricted in that it fails to construct nonlinear decision regions. The advantage of using the Decision tree [2] approach is that we need not make any such assumption about the classification rule but they do require a large amount of training data to model a distribution that is representative of the problem space.

In this work we present preliminary efforts to apply LDA and Decision Trees to phone classification as a first step towards integration into a complete speech recognition system[3]. demonstrate the efficacy of these schemes using a subset of the OGI Alphadigit corpus which consists of telephone-bandwidth alphadigits strings and is described later.

CLASSIFICATION TECHNIQUES

In this section we give a comprehensive description of the two classification techniques we implemented during the course of this project.

Linear Discriminant Analysis

Linear Discriminant Analysis is the common technique used for multigroup data classification and dimensionality reduction. LDA maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal separablity. It uses a linear transformation which can either be implemented in a class-dependent or class-independent fashion.

A within-class scatter matrix defines the scatter of samples around their respective class means

and is computed using:

$$S_w = \sum_{i=1}^{L} P_i E\left\{(X - M_i)(X - M_i)^T\right\}$$

Where M_i is the mean of the ith class and P_i is the relative occurrence of members of class i in the training data. A between-class scatter matrix defines the scatter of the input data around the global mean and is computed as:

$$S_b = \sum_{i=1}^{L} P_i (M_i - M_o)(M_i - M_o)^T$$

Where M_o is the global mean, M_i is the individual mean of class i and P_i is the occurrence probability for class i in the training data. The overall mixture scatter matrix is obtained by the covariance matrix of all samples and is computed using the following equation:

$$S_m = E\left\{(X - M_o)(X - M_o)^T\right\} = S_w + S_b$$

The optimizing criterion to obtain the LDA transform is a combination of within-class scatter, between-class scatter and the mixture-scatter. The criteria commonly used are:

$$criterion = \ln\left|inv(S_w) \times S_b\right|$$

$$criterion = S_b - \mu(S_w - c)$$

In our implementation, we used the first of these to optimize the class separation in the transform space.

The transformation matrix is formed by the eigenvectors corresponding to the dominant eigenvalues of the optimizing criterion. An eigenvector of a transformation represents a 1-D subspace of the vector space in which the transformation is applied. A set of these eigenvectors whose corresponding eigenvalues are non-zero are all linearly independent and invariant. Thus, any vector space can be represented in terms of linear combinations of the eigenvectors.

LDA for data classification can be implemented in two ways: Class-dependent and Class-independent transformation. The class-dependent approach involves maximizing the ratio of between-class covariance to within-class covariance for each class separately. This results in L transformation. each corresponding to one maximizing the ratio of between-class scatter to the within-class scatter across all classes simultaneously. In the class-independent approach the

optimizing criterion is used to define a single transformation.

Figure 1 outlines the flow of the training and testing procedure in our system.

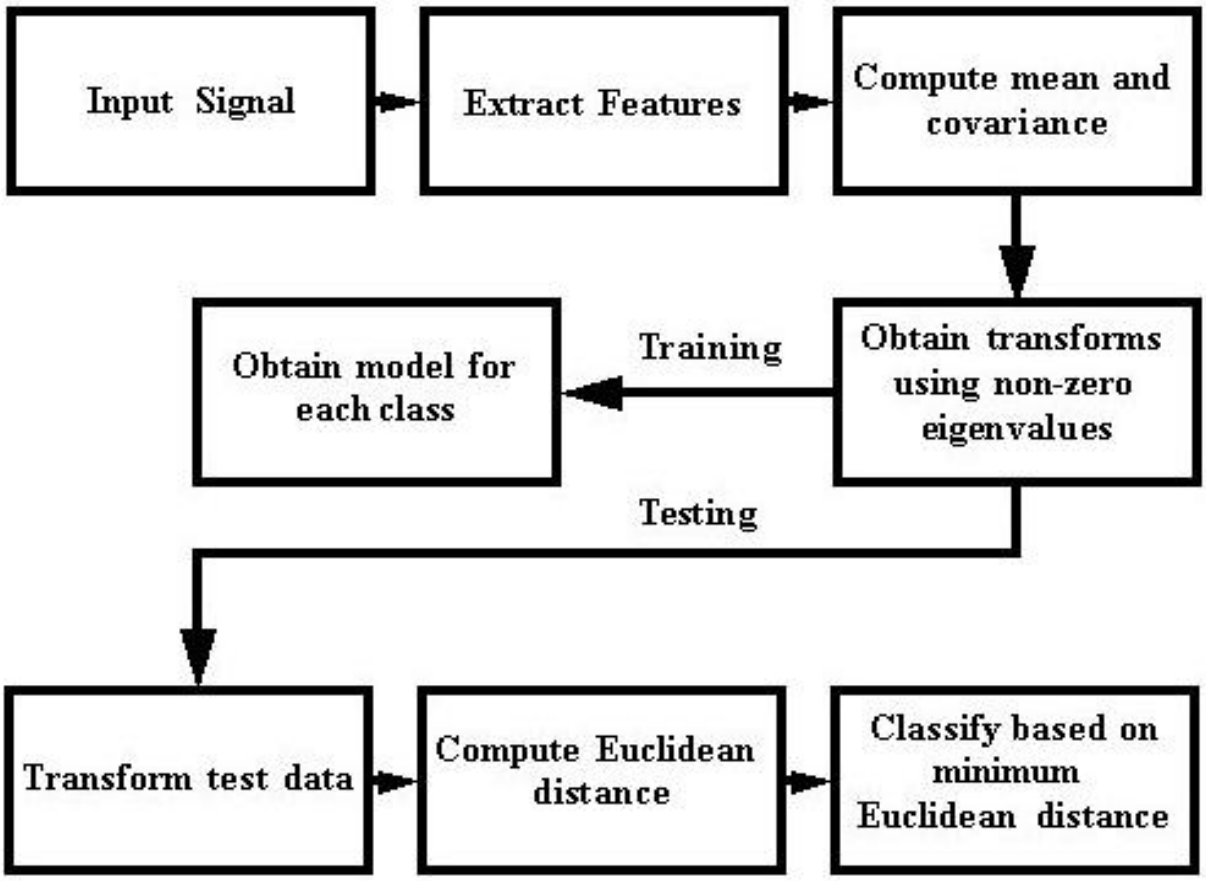

Figure 1. A block diagram shows the steps involved in the classification of LDA

Decision Trees

Binary trees give an interesting and often illuminating way of looking at data in classification problems. In the last few decades, considerable research has been conducted on the use of decision trees to solve classification problems. One important feature of decision trees is their capability to break down a complex decision-making or classifying problem into a set of simplified problems. The purpose of a decision tree classifier is to draw a conclusion through the breaking down and solving of these simple problems that achieve the desired solution of the original complicated problem.

To construct a decision tree, the tree is first grown to completion so that the tree partitions the training sample into terminal regions of all one class [4]. Tree construction uses the recursive partitioning algorithm, and its input requires a set of training examples, a splitting rule, and a

stopping rule.

The partitioning of the tree is determined by the splitting rule and the stopping rule determines if the examples in the training set can be split further. If a split is still possible, the examples in the training set are divided into subsets by performing a set of statistical tests defined by the splitting rule. The test that results in the best split is selected and applied to the training set which divides the training set into subsets. This procedure is recursively repeated for each subset until no more splitting is possible.The splitting rules usually involve an exhaustive search in finding the best split. A statistical value is obtained for every possible split of all attributes at each node. Figure 2 summarizes the decision trees algorithm in our implementation

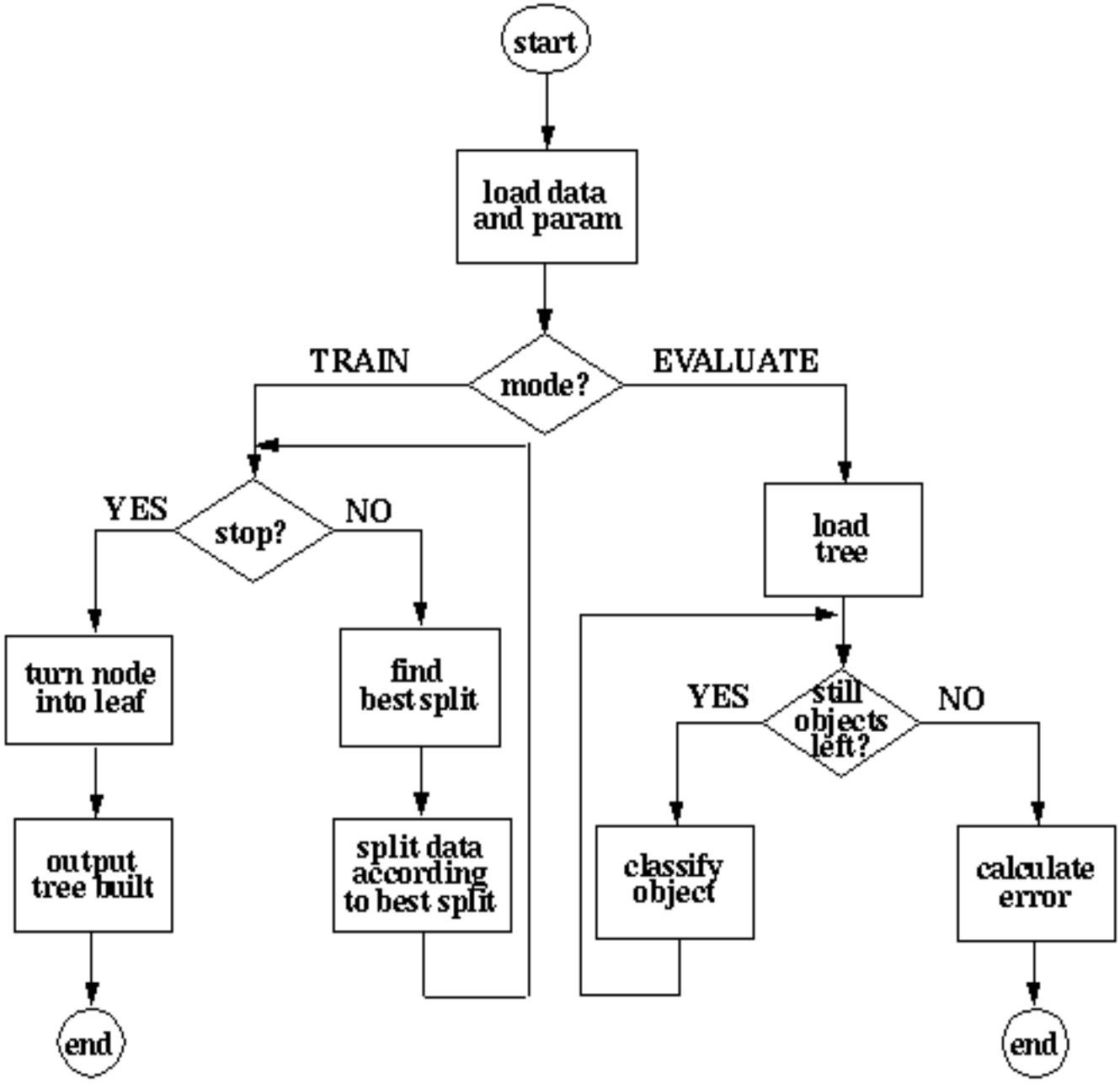

Figure 2. The flow diagram for the decision tree based classification

DATABASE

We have used the OGI Alphadigit corpus[5] to test the performance of the classification schemes. The OGI Alphadigit corpus is a telephone database collected using a T1 interface with over 300 subjects reading a list of either 19 or 29 alphanumeric strings. These strings averaged 6 word in length and each list was designed to balance the phonetic context of all letter and digit pairs. There were 1102 unique propounding strings.

EXPERIMENTAL DESIGN

We first used the ISIP front-end[6] to extract the required features for the phones. The features we used were 12 mel scaled cepstral coefficients (MFCC's). These are the standard features used by most state-of-the-art phone classification as well as speech recognition systems. The number of phones in to which these would be classified were thirty. So this turns out to be a thirty-class twelve-dimensional classification problem.

The division of the data into training and test sets was done. The distribution of both these sets was identical to the distribution of the original data. Also the division was done randomly so that no bias is introduced.

Since the goal of our project was also to better understand how each of the classification algorithms that we had employed actually work, we wanted to be able to visually inspect the results. So we first performed the classification on a reduced two-dimensional two-class problem. We took the phones 'f' and 'uw' and classified them on the basis of the first and eight MFCCs. The reason we took these two phones was because one of them represented a consonant and vowel respectively. The coefficients were also chosen so that they are comparatively apart. Once the classification was done, we wanted to plot the decision region for both the cases.

Once we understood how LDA and DT based classification work, we wanted to apply them to the complete data set with all the features. This was done and the closed and open loop results were computed.

RESULTS

The results for the simplified two-class two-dimensional problem are shown in Figure 3. The black dots represent one class whereas the red dots represent the other. The green dots represent the data points that have been misclassified. The variable along the X axis the first MFCC coefficient and the variable along the Y axis is the eighth MFCC coefficient. We can notice that the discriminant function is a straight line in case of LDA whereas it is a nonlinear one in case of the decision tree based classification. The open loop error rates are also shown in Table 1.

The misclassification rates for the thirty-class twelve-dimensional case are also given in Table 2. Observe that DT's have a near zero classification error in all cases when tested on the training set i.e. closed loop error.The LDA results are comparable to the best error rates obtained from other linear classification techniques.

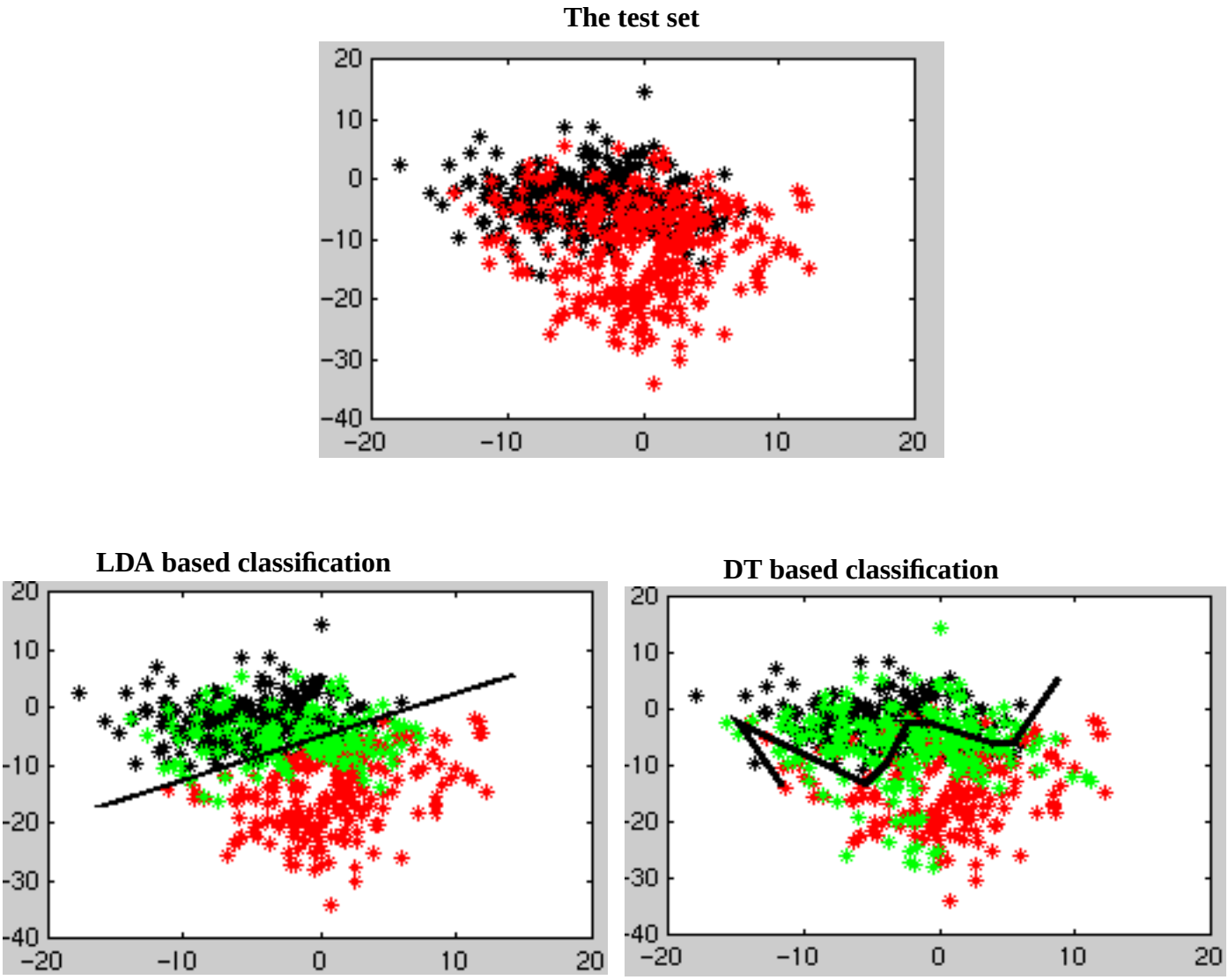

Figure 3. The test set and decision region plots obtained using LDA and DT

	LDA	DT
Closed loop error	20.15%	0.1%
Open loop error	25.62%	34.94%

Table 1: Misclassification rates for two-class two-dimensional problem

	LDA	DT
Closed loop error	57.17%	0.14%
Open loop error	68.49%	87.23%

Table 2: Misclassification rates for the whole dataset

SUMMARY AND CONCLUSIONS

Linear and nonlinear classification techniques were studied and applied to the phone classification problem. LDA was chosen as the linear classification technique and Decision tree based classification was employed for the nonlinear case. DTs seem to perform very well in case of closed loop testing but their performance is not satisfactory in case of open loop testing. The reason for this is overtraining resulting in their inability to generalize. The decision tree based approach does not stop training till all the elements are classified correctly leading to overtraining. The challenge lies in having an appropriate stopping rule during threshold.

LDA results are comparable to those obtained from other linear classifiers. It was evident that the decision region was not a linear one. So we can also conclude that nonlinear classification techniques should be used for phone classification to achieve a low misclassification error.

FUTURE WORK

As a continuation to this work, these classification techniques can be applied to better features extracted from the speech signal. The other common features that can be used are the delta coefficients and the mean in addition to the MFCCs. Also the DT approach can be employed but with better pruning strategies. To further understand the decision region, other nonlinear techniques such as Neural Networks and Support Vector Machines can also be employed.

ACKNOWLEDGMENTS

We would like to express our sincere thanks to Dr. Nicolas Younan and Dr. Joseph Picone for their support and technical help. The class was a pleasure and we immensely enjoyed working on this project. We would like to thank Aravind Ganapathiraju for his help in the design of the experiments.

REFERENCES

[1] S. Balakrishnama, A. Ganapathiraju and J. Picone, “Linear Discriminant Analysis for Signal Processing Problems”, *Proceedings of the IEEE Southeastcon*, pp. 78-81, Lexington, Kentucky, USA, March 1999.

[2] J. Ngan, A. Ganapathiraju and J. Picone, “Improved Surname Pronunciations Using Decision Trees”, *Proceedings of the International Conference on Spoken Language Processing*, pp. 3285-3288, Sydney, Australia, November 1998.

[3] J. Hamaker, A. Ganapathiraju, J. Picone and J, Godfrey, “Advances in Alphadigit Recognition Using Syllables”, *Proceedings of the IEEE International Conference on Acoustics, Speech and signal Processing*, pp. 421-424, Seattle, Washington, USA May 1998.

[4] W. Buntine, *A Theory of Learning Classification Rules*, Ph.D. thesis, University of Technology, Sydney, 1991.

[5] “Alphadigit v1.0”,available at *http://cslu.cse.ogi.edu/corpora/alphadigit/*, Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, USA, 1997.

[6] V. Mantha, R. Duncan, Y. Wu, J. Zhao, A. Ganapathiraju and J. Picone, “Implementation and Analysis of Speech Recognition Front-Ends”, *Proceedings of the IEEE Southeastcon*, pp. 32-35, Lexington, Kentucky, USA, March 1999.