# IMPROVEMENTS ON ISOLATED WORD RECOGNITION USING SUBSPACE METHODS

*M. Bilginer Gülmezoğlu[1], Rifat Edizkan[1], Semih Ergin[1] and Atalay Barkana[2]*

[1]Department of Electrical and Electronics Engineering, Osmangazi University
Batı-Meşelik, 26480, Eskişehir, Turkey
phone: + (90) 222-2393750, fax: + (90) 222-2200535, email: {bgulmez, redizkan, sergin}@ogu.edu.tr

[2]Department of Electrical and Electronics Engineering, Anadolu University
Muttalip, Eskişehir, Turkey
phone: + (90) 222-3213550, email: atalaybarkan@anadolu.edu.tr

## ABSTRACT

The purpose of this study is to investigate the effects of different forms of between-class scatter matrices on multi-class problems. Two different between-class scatter matrices are defined in Fisher's linear discriminant analysis (FLDA) and the classification rates better than that of classical FLDA are obtained for TI-digit database. In this study, the criteria that give separate subspaces for each class are also proposed. It is seen that considering only the within-class scatter in the classification gives better results than that of considering both the within- and between-class scatters for TI-digit database.

## INTRODUCTION

In the speech recognition, Hidden Markov Model (HMM), Neural Networks (NN) and subspace methods are widely used. One of the well known subspace methods is FLDA. FLDA is an important method for linear dimension reduction in statistical pattern classification and speech recognition with small and large vocabulary applications [1-5]. In comparison with HMM, FLDA uses more simple training techniques and decision criterion.

Loog and Umbach proposed a generalized version of FLDA which allows to deemphasize the contributions of classes which are far apart from each other [6]. This criterion also considers differences in class covariances, thus being an extension of FLDA towards heteroscedastic data [5,6].

FLDA is also widely used together with other classification methods. For instance, it is used together with HMM in online handwriting [7] and speech [8] recognition, and together with Maximum Likelihood in isolated word recognition [9].

Raudys and Duin pointed out that the pseudo-Fisher linear classifier is the "diagonal" Fisher linear classifier in the subspace of the principal components corresponding to nonzero eigenvalues of the sample covariance matrix [10].The pseudo Fisher classifier plainly ignores the directions with zero eigen values. But Jing et al. [11] stated that according to Direct LDA's theory, some of the eigenvectors of within-class scatter matrix corresponding to the largest eigenvalues should be discarded and the process should keep the remainders, especially those eigenvectors corresponding to the zero eigenvalues.

One of the disadvantages of the LDA is to give an inadequate hypothesis when the boundary between two classes is nonlinear [12]. Yang et al. [13] emphasized that Fisher criterion is not an absolute criterion, and it should be associated with the statistical correlation together to assess the discrimination of a set of discriminant vectors. They stated that in order to obtain a set of most discriminatory discriminant vectors, Fisher criterion should be associated with the orthogonal constraints which can make sure the resulting features to be uncorrelated.

It is obvious that the scatter of the classes in any database affect the classification performance of subspace methods. Let us consider two classes having scatters as shown in Fig.1. In this case, the FLDA correctly classifies the feature vectors, but the Common Vector Approach (CVA) fails. However the scatter of classes given in Fig. 1 can not be encountered in speech database. For this reason, the aim of this paper is to improve the recognition rates of isolated words by considering within-class scatters together with between-class scatters defined in different forms. The experimental studies on TI-digit database indicate that the recognition rates are increased when improved windowing method is applied and proposed subspace methods are used.

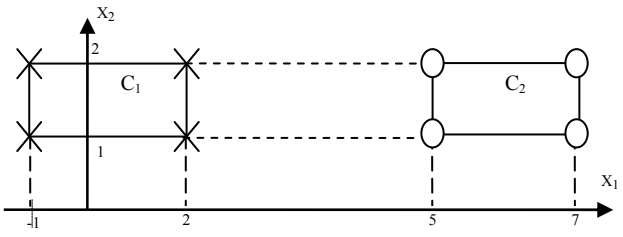


Figure 1.

## 2. THEORY

In this paper, the effects of different forms of between-class scatter matrices on the multi-class problems are investigated and new subspace methods are proposed.

This study is a part of the project (200315053) supported by the Research Fund of Osmangazi University.

### 2.1. The Study on the FLDA

Fisher's optimization criterion is one of the methods that consider both within- and between-class scatters. Fisher has used LDA in order to solve multi-class problems by maximizing the ratio of between-class scatter matrix to within-class scatter matrix in low dimensional space [1,14].

Fisher's criterion to be maximized is expressed as:

$$\boldsymbol{J(W) = Tr\left\{(W^T \mathbf{S}_W W)^{-1}(W^T \mathbf{S}_B W)\right\}} \qquad (1)$$

where $\boldsymbol{S}_w$ and $\boldsymbol{S}_B$ are the within- and between-class scatter matrices respectively, and $\boldsymbol{W}$ represents the projection matrix [1]. The dimension of the subspace obtained from the optimization criterion in Eq. (1) is one less than the number of classes. In this subspace, the recognition rates of the classes are not high. This limitation is a result of the definition of the between-class scatter matrix. The previous studies which optimize the criterion in Eq. (1) can be summarized as follows:

i) The condition of $\boldsymbol{w}_i^T \boldsymbol{S}_t \boldsymbol{w}_j = 0, \quad \forall i \neq j$ is used instead of the condition $\mathbf{w}_i^T \mathbf{w}_j = 0, \quad \forall i \neq j,$ $i,j = 1, 2, \ldots, d$, [13]. Here, $\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_B$.

ii) In [14], the between-class scatter matrix is written in terms of the differences of averages of classes.

iii) The discrimination of the between-class scatter matrix is increased by multiplying that matrix with a weight function which is defined by eigenvalues of the between-class scatter matrix [6].

### 2.2. Finding a Unique Subspace by Minimizing $\boldsymbol{F}_1 = \left|\boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}\right| / \left|\boldsymbol{W}^T \boldsymbol{\Phi}_{B_T} \boldsymbol{W}\right|$ Criterion

Let *m* represents the number of feature vectors of each class, *n* represents the number of elements in each feature vector and $\boldsymbol{a}_i^c$ ($i$=1, 2,.., $m$) represents a feature vector. In this section, the optimization criterion $\boldsymbol{F}_1$ is defined to find a unique subspace that represents all classes. In this criterion, $\boldsymbol{S}_w$ denotes Fisher's total within-class scatter matrix. Total between-class scatter matrix $\boldsymbol{\Phi}_{B_T}$ is defined in two different forms:

i) $\boldsymbol{\Phi}_{B_T}$ is defined in a form that the distance between average of each class and the average of the rest of classes is maximized:

$$\boldsymbol{\Phi}_{B_T}^1 = \sum_{c=1}^{C} \left(\boldsymbol{a}_{ave}^c - \boldsymbol{a}_{r,ave}^c\right)\left(\boldsymbol{a}_{ave}^c - \boldsymbol{a}_{r,ave}^c\right)^T \qquad (2)$$

where $\boldsymbol{a}_{ave}^c$ and $\boldsymbol{a}_{r,ave}^c$ represent the average of feature vectors in class $c$ and the average of feature vectors of the rest of classes respectively. Since the determinant of $\boldsymbol{\Phi}_{B_T}^1$ is zero, the criterion to be maximized is defined as $\left|\boldsymbol{W}^T \boldsymbol{\Phi}_{B_T}^1 \boldsymbol{W}\right| / \left|\boldsymbol{W}^T \mathbf{S}_w \boldsymbol{W}\right|$. The eigenvectors associated with the largest eigenvalues of the $\boldsymbol{S}_w^{-1} \boldsymbol{\Phi}_{B_T}^1$ will be used in the recognition process.

ii) As another approach, $\boldsymbol{\Phi}_{B_T}$ is defined in a form that the distance between each feature vector of any class and the average of the rest of classes is maximized:

$$\boldsymbol{\Phi}_{B_T}^2 = \sum_{c=1}^{C} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^c - \boldsymbol{a}_{r,ave}^c\right)\left(\boldsymbol{a}_i^c - \boldsymbol{a}_{r,ave}^c\right)^T \qquad (3)$$

In this approach, usually the determinant of $\boldsymbol{\Phi}_{B_T}^2$ is not zero. Therefore the minimization of $\left|\boldsymbol{W}^T \mathbf{S}_w \boldsymbol{W}\right| / \left|\boldsymbol{W}^T \boldsymbol{\Phi}_{B_T}^2 \boldsymbol{W}\right|$ criterion and maximization of $\left|\boldsymbol{W}^T \boldsymbol{\Phi}_{B_T}^2 \boldsymbol{W}\right| / \left|\boldsymbol{W}^T \mathbf{S}_w \boldsymbol{W}\right|$ criterion give the same results. The eigenvectors associated with the smallest eigenvalues of the $\left(\boldsymbol{\Phi}_{B_T}^2\right)^{-1} \boldsymbol{S}_w$ will be used in the recognition process.

### 2.3. Finding Separate Subspaces for Each Class by Minimizing $\boldsymbol{F}_2 = \left|\boldsymbol{W}^T \boldsymbol{\Phi}_w^c \boldsymbol{W}\right| / \left|\boldsymbol{W}^T \boldsymbol{\Phi}_B^c \boldsymbol{W}\right|$ Criterion

In this section, the optimization criterion $\boldsymbol{F}_2$ is defined to find separate subspaces for each class. In this criterion, the within-class scatter matrix $\boldsymbol{\Phi}_W^c$ is defined as:

$$\boldsymbol{\Phi}_w^c = \sum_{i=1}^{m} \left(\boldsymbol{a}_i^c - \boldsymbol{a}_{ave}^c\right)\left(\boldsymbol{a}_i^c - \boldsymbol{a}_{ave}^c\right)^T \qquad (4)$$

and the between-class scatter matrix $\boldsymbol{\Phi}_B^c$ is defined in two different forms:

i) First of all, the $\boldsymbol{\Phi}_B^c$ is defined as:

$$\boldsymbol{\Phi}_B^c = \sum_{i=1}^{m} \left(\boldsymbol{a}_i^c - \boldsymbol{a}_{r,ave}^c\right)\left(\boldsymbol{a}_i^c - \boldsymbol{a}_{r,ave}^c\right)^T \qquad (5)$$

The eigenvectors associated with the smallest eigenvalues of the $(\boldsymbol{\Phi}_B^c)^{-1} \boldsymbol{\Phi}_w^c$ will be used in the recognition process.

ii) Secondly, the between-class scatter matrix is defined in a form that the distance between feature vectors of any class and the average of each of the other classes is maximized.

$$\boldsymbol{\Phi}_{B_I}^c = \sum_{\substack{j=1 \\ j \neq c}}^{C} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^c - \boldsymbol{a}_{ave}^j\right)\left(\boldsymbol{a}_i^c - \boldsymbol{a}_{ave}^j\right)^T \qquad (6)$$

The eigenvectors associated with the smallest eigenvalues of $(\boldsymbol{\Phi}_{B_I}^c)^{-1} \boldsymbol{\Phi}_w^c$ will be used in the recognition process.

## 3. EXPERIMENTAL STUDY

In the experimental study, the TI-digit database is used. After end-point detection, the speech frames are pre-emphasized and two different windowing methods are applied:

i) Each repetition is divided into frames with 256 samples. After the Hamming window with 64 overlap is applied, 11 root-melcep parameters are calculated. Then these parameters are stacked in order to construct the feature vector for each repetition of each digit. After this process, the dimensions of the feature vectors are extended to 407 (dimension of the longest vector in the training set) by padding random values. The scatter matrix for each digit turns out to be a 407x407 matrix. For sufficient case ($m>n$) [15], the scatter matrix and its eigenvalues and eigenvectors are calculated by using the $m$=427 feature vectors in each class. The eigenvectors corresponding to different number of the smallest eigenvalues ($n$-$k$+1) are used in the recognition process.

ii) In the second windowing method, each repetition is divided into 8 frames. The Hamming window is applied to each frame. The overlap between the frames is set to ¼ of the number of samples in each frame. 33 root-melcep parameters are calculated and stacked in order to construct the feature vector with the size of 330 for each repetition of each digit. This proposed method is called Variable Frame Length (VFL) method [16].

### 3.1. The study on FLDA

The nonzero eigenvalues of $\boldsymbol{S}_w^{-1}\boldsymbol{S}_B$ are used to maximize the Fisher's optimization criterion in Eq. (1). The decision criterion is expressed as follows:

$$K^* = \underset{1\le c\le C}{argmin} \left\| \boldsymbol{P}_f (\boldsymbol{a}_x - \boldsymbol{a}_{ave}^c) \right\|^2 \qquad (7)$$

where $\boldsymbol{P}_f$ represents the projection matrix of all classes. If unknown feature vector $\boldsymbol{a}_x$ belongs to the class $c$, the distance should be minimum. The experimental studies can be summarized as follows:

i) In the recognition process, when the feature vectors with the dimension of 407 are used, the recognition rates of 96.5% and 92.3% are obtained for the training and test sets respectively. Besides this; when the optimizations defined in parts (i-iii) in the subsection 2.1 are applied to Fisher's criterion, it is observed that the recognition rates are not increased [17].

ii) A subspace from within-class scatter matrix $\boldsymbol{S}_w$ and another subspace from between-class scatter matrix $\boldsymbol{S}_B$ are defined and the experimental studies are repeated using the feature vectors with the length of 330. The results obtained from $\boldsymbol{S}_w$, $\boldsymbol{S}_B$ and their combination using the Borda Count method [18,19] are given in Table 1.

Table 1. The recognition rates obtained from $\boldsymbol{S}_w$, $\boldsymbol{S}_B$ and Borda Count method.

| | Training Set (%) | Test Set (%) |
|---|---|---|
| $\boldsymbol{S}_w$ | 92.5 | 90.3 |
| $\boldsymbol{S}_B$ | 90.9 | 87.6 |
| Borda Count | 96.5 | 95.1 |

### 3.2. The Studies for the Criterion $\boldsymbol{F}_1$

In this part, if $\boldsymbol{\Phi}_{B_T}^2$ is considered as the between-class scatter matrix, the following decision criterion is used in the recognition process.

$$K^* = \underset{1\le c\le C}{argmin} \frac{\left\| \boldsymbol{P}_1 (\boldsymbol{a}_x - \boldsymbol{a}_{ave}^c) \right\|^2}{\left\| \boldsymbol{P}_1 (\boldsymbol{a}_x - \boldsymbol{a}_{r,ave}^c) \right\|^2} \qquad (8)$$

where $\boldsymbol{P}_1$ is the projection matrix of all classes. If $\boldsymbol{\Phi}_{B_T}^1$ is considered as the between-class scatter matrix, only numerator of Eq. (8) is used in the recognition process. The results obtained for two different between-class scatter matrices given in Eqs. (2) and (3) are given in Table 2.

Table 2. The recognition rates (%) for the criterion $\boldsymbol{F}_1$.

| $\boldsymbol{F}_1 = \left\| \boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W} \right\| / \left\| \boldsymbol{W}^T \boldsymbol{\Phi}_{B_T} \boldsymbol{W} \right\|$ | | | |
|---|---|---|---|
| $\boldsymbol{S}_w^{-1}\boldsymbol{\Phi}_{B_T}^1$ | | $\left(\boldsymbol{\Phi}_{B_T}^2\right)^{-1} \boldsymbol{S}_w$ | |
| Train | Test | Train | Test |
| 97.6 | 95.9 | 97.92 | 95.82 |

### 3.3. The Studies for the Criterion $\boldsymbol{F}_2$

In this part, the following decision criterion is used in the recognition process.

$$K^* = \underset{1\le c\le C}{argmin} \left\| \boldsymbol{P}_2^c (\boldsymbol{a}_x - \boldsymbol{a}_{ave}^c) \right\|^2 \qquad (9)$$

where $\boldsymbol{P}_2^c$ is the projection matrix of the class c. The results obtained for two different between-class scatter matrices given in Eqs. (5) and (6) are given in Table 3.

Table 3. The recognition rates (%) for the criterion $\boldsymbol{F}_2$.

| $\boldsymbol{F}_2 = \left\| \boldsymbol{W}^T \boldsymbol{\Phi}_w^c \boldsymbol{W} \right\| / \left\| \boldsymbol{W}^T \boldsymbol{\Phi}_B^c \boldsymbol{W} \right\|$ | | | |
|---|---|---|---|
| $\left(\boldsymbol{\Phi}_B^c\right)^{-1} \boldsymbol{\Phi}_w^c$ | | $\left(\boldsymbol{\Phi}_{B_l}^c\right)^{-1} \boldsymbol{\Phi}_w^c$ | |
| Train | Test | Train | Test |
| 96.16 | 88.91 | 100 | 86 |

i) Instead of minimizing the ratio $\boldsymbol{F}_2$ for each class, the results which are obtained from minimizing $\left\| \boldsymbol{W}^T \boldsymbol{\Phi}_w^c \boldsymbol{W} \right\|$ for each class and maximizing $\left\| \boldsymbol{W}^T \boldsymbol{\Phi}_B^c \boldsymbol{W} \right\|$ for each class can be combined with the Borda Count method. The results obtained from $\boldsymbol{\Phi}_w^c$, $\boldsymbol{\Phi}_B^c$ and their combination using the Borda Count method are given in Table 4.

Table 4. The recognition rates obtained from $\mathbf{\Phi}_w^c$, $\mathbf{\Phi}_B^c$ and Borda Count method.

| | Training Set (%) | Test Set (%) |
|---|---|---|
| $\mathbf{\Phi}_w^c$ | 100 | 98.82 |
| $\mathbf{\Phi}_B^c$ | 92.8 | 89.5 |
| Borda Count | 98.3 | 97.4 |

When the results obtained using the eigenvectors associated with the minimum eigenvalues of $\mathbf{\Phi}_w^c$ and the results obtained using the eigenvectors associated with the maximum eigenvalues of $\mathbf{\Phi}_{B_1}^c$ are combined with the Borda Count method, the recognition rates of 98.32% for the training set and 97.36% for the test set are obtained.

## 4. CONCLUSION

In the speech recognition, the desired performance cannot be obtained for some scatters of classes when the subspace methods considering only within-class or between-class scatter are used. This situation is encountered especially when the number of classes increases. Therefore, between-class scatters can be considered together with within-class scatters in the speech recognition.

In this paper, between-class scatter matrices defined in four different forms are used for a unique subspace of all classes and separate subspace of each class. The effects of these matrices on the isolated word recognition are investigated for the TI-digit database. It is seen that the best results for the unique subspace are obtained for $\mathbf{\Phi}_{B_T}^2$. In this case, the recognition rates of 97.926% and 95.82% are obtained for the training and test sets respectively. These results are greater than the results given in literature for the FLDA method [13,20].

When separate subspaces from within-class and between-class scatter matrices are used for each class as in Table 4, and the results obtained from these subspaces are combined using the Borda Count method, the recognition rates of 98.3% and 97.4% are obtained for the training and test sets respectively. However, when separate within-class scatters are used for each class, highest recognition rates are obtained (100% for the training set and 98.8 for the test set) [15]. In conclusion, the results in Table 4 indicate that consideration of between-class scatters is unnecessary for TI-digit database.

## REFERENCES


[1] C.M. Bishop, *Neural Networks for Pattern Recognition*, Address: Clarendon Press, 1995.

[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* Address: Academic Press, 1990.

[3] R. Haeb-Umbach and H. Ney, "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition", in *Proc. ICASSP*, San Fransisco, California, pp. 13-16, 1992.

[4] E. G. Schukat-Talamazzini, J. Hornegger and H. Niemann, "Optimal Linear Feature Transformations for Semi-Continuous Hidden Markov Models", in *Proc. ICASSP*, Detroit, Michigan, pp. 369-372, 1995.

[5] G. Saon, M. Padmanabhan, and R. Gopinath, "Maximum Likelihood Discriminant Feature Spaces", in *Proc. ICASSP*, Istanbul, Turkey, pp. 1747-1750, 2000.

[6] M. Loog and R. Haeb-Umbach, "Multi-Class Linear Dimension Reduction by Generalized Fisher Criteria", in *Proc. Int. Conf. on Spoken Languange Processing*, pp. 1069-1072, 2000.

[7] D.G.A. Dolfing and R. Haeb-Umbach, "Signal Representations for Hidden Markov Model Based On-Line Handwritting Recognition", in *Proc. ICASSP*, Munich, Germany, 1997, pp. 3385-3388.

[8] R. A. Sukkar and J.G. Wilpon, "A Two Pass Classifier For Utterance Rejection In Keyword Spotting", in *Proc. ICASSP*, Minneapolis, Minnesota, 1993, pp. 451-454.

[9] A. Hauenstien and E. Marschall, "Methods For Improved Speech Recognition Over Telephone Lines", in *Proc. ICASSP*, Detroit, Michigan, 1995, pp. 425-428.

[10] S. Raudys and P.W. Duin, "Expected Classification Error of the Fisher Linear Classifier with Pseudo-Inverse Covariance Matrix", *Pattern Recognition Letters*, vol. 19, pp. 385-392, 1998.

[11] Y. Jing, D. Zhang and Y. Yao, "Improvements on The Linear Discrimination Technique with Application to Face Recognition", *Pattern Recognition Letters*, vol. 24, pp. 2695-2701, 2003.

[12] H. Nomiya and K. Uehara, "Improvement of Linear Discriminant Analysis by Applying The Ensemble Method", *IPSJ SIGNotes Mathematical modeling and Problem Solving*, No.046

[13] J. Yang, J-Y. Yang and D. Zhang, "What's Wrong with Fisher Criterion", *The Journal of the Pattern Recognition Society*, vol.35, pp. 2665-2668, 2002.

[14] M. Loog, R.P.W. Duin and R. Haeb-Umbach, "Multi-Class Linear Dimension Reduction by Weighted Pairwise Fisher Criteria", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 762-766, 2001.

[15] M. B. Gülmezoğlu, V. Dzhafarov, R. Edizkan and A. Barkana, "The Common Vector Approach and Its Comparison with Other Subspace Methods in Case of Sufficient Data", *Submitted to Speech Communication*, 2004.

[16] S. Ergin , M. B.Gülmezoğlu and A. Barkana, "Use Of Improved Feature Vectors In Affine Transformation For Robust Speech Recognition", in *Proc. Int. Conf. on Information,* Çeşme, İzmir, 2004

[17] G. Saon, M. Padmanabhan, and R. Gopinath, "Maximum Likelihood Discriminant Feature Spaces", in *Proc. ICASSP*, Istanbul, Turkey, pp. 1747-1750, 2000.

[18] T.K. Ho, J.J. Hull and S.N. Srihari, "Decision Combination in Multiple Classifier Systems", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 66-75, 1994.

[19] D. Black, *The Theory of Commitees and Elections*, Adress: Cambridge University Press, 1963.

[20] V. Zivognovic and D. Noll, "Minimum Fisher Information Spectral Analysis", in *Proc. ICASSP*, Munich, Germany, 1997.