Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 21 (2007) 266–281
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

The common vector approach and its comparison with other subspace methods in case of sufficient data

M. Bilginer Gülmezoğlu [a,*], Vakıf Dzhafarov [b], Rifat Edizkan [c], Atalay Barkana [d]

[a] Eskişehir Osmangazi University, Electrical and Electronics Engineering Department, Eskişehir 26480, Turkey
[b] Anadolu University, Department of Mathematics, Eskişehir, Turkey
[c] Eskişehir Osmangazi University, Electrical and Electronics Engineering Department, Eskişehir, Turkey
[d] Anadolu University, Electrical and Electronics Engineering Department, Eskişehir, Turkey

Received 18 March 2005; received in revised form 5 June 2006; accepted 5 June 2006
Available online 17 July 2006

Abstract

This paper presents an application of the common vector approach (CVA), an approach mainly used for speech recognition problems when the number of data items exceeds the dimension of the feature vectors. The calculation of a unique common vector for each class involves the use of principal component analysis. CVA and other subspace methods are compared both theoretically and experimentally. TI-digit database is used in the experimental study to show the practical use of CVA for the isolated word recognition problems. It can be concluded that CVA results are higher in terms of recognition rates when compared with those of other subspace methods in training and test sets. It is also seen that the consideration of only within-class scatter in CVA gives better performance than considering both within- and between-class scatters in Fisher's linear discriminant analysis. The recognition rates obtained for CVA are also better than those obtained with the HMM method.

1. Introduction

Statistical and probabilistic approaches are commonly used in classification problems (Bishop, 1995; Deller et al., 1993; Fukunaga, 1990; Rabiner and Juang, 1993). One of the statistical methods is based on subspace methods. The most well-known subspace method is Fisher's linear discriminant analysis (FLDA). FLDA is an important method for linear dimensionality reduction in statistical pattern classification with small and large sample size (Bishop, 1995; Fukunaga, 1990; Haeb-Umbach and Ney, 1992; Schukat-Talamazzini et al., 1995; Saon et al., 2000). Although FLDA is the linear transformation that maximizes the mean squared distance between the classes in lower-dimensional feature space, it is not optimal in respect to minimizing classification error rate in that space (Loog and Haeb-Umbach, 2000). Loog and Haeb-Umbach (2000) proposed a generalized

* Corresponding author. Tel.: +90 222 2393750x3261; fax: +90 222 2290535.
E-mail addresses: bgulmez@ogu.edu.tr (M.B. Gülmezoğlu), vcaferov@anadolu.edu.tr (V. Dzhafarov), redizkan@ogu.edu.tr (R. Edizkan), atalaybarkan@anadolu.edu.tr (A. Barkana).

0885-2308/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2006.06.002

version of FLDA that allows deemphasizing of the contributions of classes far apart from each other. In this criterion, the differences among class means are also considered. This is an extension of FLDA towards heteroscedastic data. A weighting function is used to define a generalized between-class scatter matrix (Loog and Haeb-Umbach, 2000; Loog et al., 2001). Saon et al. (2000) showed that under diagonal covariance gaussian modelling constraints, heteroscedastic discriminant analysis (HDA) alone actually degrades recognition performance.

Yang et al. (2002) emphasized that the Fisher criterion is not an absolute criterion and it should be associated with statistical correlation to assess the discrimination of a set of discriminant vectors. They stated that, in order to obtain a set of the most discriminatory discriminant vectors, Fisher criterion should be associated with orthogonal constraints so that the resulting features are uncorrelated.

The common vector approach (CVA) is a subspace method that eliminates unwanted information, such as environmental effects, personal and phase differences, and temporal variations from a spoken word. In the CVA method, a common vector that represents the common properties or invariant features of the word-class is calculated. CVA, which is a feature extraction method, is applied to each class separately considering the within-class scatter of the data only. A special application of CVA in speaker-independent isolated word recognition has also been recently introduced (Barkana et al., 1995; Gülmezoğlu et al., 1999; Gülmezoğlu et al., 2001; Keskin et al., 1995a; Keskin et al., 1995b). CVA has also been used in speaker recognition applications (Gülmezoğlu and Barkana, 1998).

CVA is particularly very practical when the number of feature vectors[1] (m) in the training set is less than or equal to the dimension (n) of each feature vector, that is, when $m \leqslant n$ (Gülmezoğlu et al., 1999; Gülmezoğlu et al., 2001). We can say that the number of data items, i.e., the number of feature vectors (m), in each class is insufficient when $m \leqslant n$. This case is usually true for many pattern classification problems.

Let the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m \in \mathbf{R}^n$ be the feature vectors for a certain word-class C in the training set. Eigenvector decomposition is then applied to the within-class scatter matrix of the set of training data. The n-dimensional feature space spanned by all eigenvectors can be divided into $(m-1)$ dimensional difference subspace $\mathbf{B}$ and $(n-m+1)$ dimensional orthogonal indifference subspace $\mathbf{B}^{\perp}$, with the result that the direct sum of these two subspaces would cover the whole feature space. In the insufficient case $(n-m+1)$ eigenvalues will be zero. The indifference subspace $\mathbf{B}^{\perp}$ is spanned by the eigenvectors corresponding to the zero eigenvalues and its compliment is the difference subspace.

In this paper, we suggest a method to find a unique common vector for each class, especially when the number of feature vectors in the training set exceeds the dimension of each feature vector ($m > n$). This is the case when the number of data items in the training set is sufficient to calculate the inverse of the within-class scatter matrix. This inverse does not exist for the insufficient data case. The common vector will be the zero vector in the sufficient data case because indifference subspace disappears, or the same thing can happen since none of the eigenvalues of the within-class scatter matrix is going to be zero in the sufficient case. Thus, the main objective in the sufficient case is to show that indifference subspaces would not disappear as long as some of the eigenvalues are very small when compared with others in the within-class scatter matrix. In the sufficient case, the common vector is obtained from the projection of the class-mean vector onto the indifference subspace.

It is obvious that the scatter of the classes in any database will affect the performance of classifiers. In Fig. 1, all the feature vectors are correctly classified with the CVA method even though they are not with FLDA. Meanwhile, FLDA can give better results than CVA for different class scatters, as shown in Fig. 2.

Our purpose is to see which one of the subspace methods will work better in the TI-digit database. Since isolated digits would have very high dimensions after feature selections, one would not be able to see the scatters of the classes in these feature spaces. Therefore, it is impossible to tell which one of the methods would work better without actual realization. Since one may wonder about the efficiency of subspace methods in comparison with the well-known HMM, we have also provided the recognition rates of the HMM method and compare it with our proposed subspace method.

[1] The vector made by concatenating the acoustic parameters (i.e., the frames) associated with an observation of a spoken word is called the feature vector throughout this paper. Thus, one utterance of a word generates a single feature vector for that word.

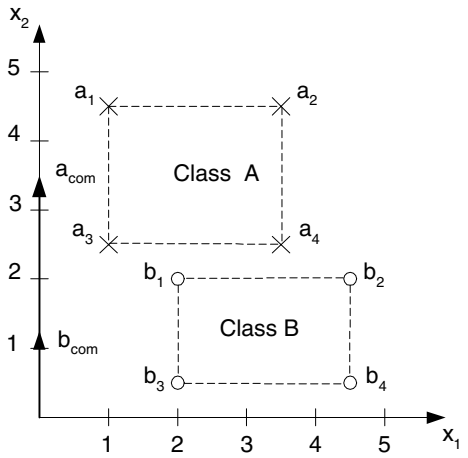

Fig. 1. Scatters of two classes in two-dimensional subspace in which FLDA fails at one of the points and CVA works at all the points.

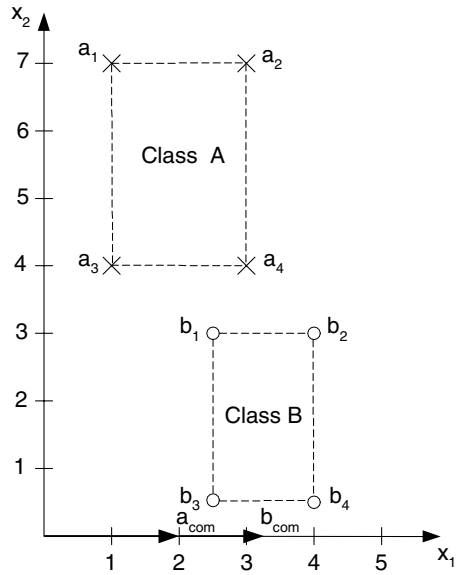

Fig. 2. Scatters of two classes in two-dimensional subspace in which FLDA works and CVA fails at four points.

The relation between the CVA and principal component analysis (PCA) has been shown in our previous work (Gülmezoğlu et al., 2001). Obtaining the eigenvalues and the eigenvectors of the within-class scatter matrices for normal distributions is known as PCA or Karhunen–Loeve transforms (KLT) (Bishop, 1995; Marrison, 1967; Parsons, 1986; Tou and Gonzales, 1974; Kuhn et al., 2000; Lee and Landgrebe, 1993; Landgrebe, 2002). PCA suggests the elimination of the feature components along the direction of the eigenvectors of the smallest eigenvalues and keeping those components along the direction of eigenvectors of the largest eigenvalues in order to reduce the number of dimensions in the feature space.

PCA is also used in many of subspace methods. Since PCA produces lower-dimensional subspaces, some subspace methods (Kohonen et al., 1979; Kohonen et al., 1980; Kohonen et al., 1984) are given in this paper: SELFIC (self-featuring information comparison) (Watanabe et al., 1967); CLAFIC (CLA may implement the

class) (Oja, 1983; Watanabe et al., 1967); and SIMCA (soft independent modelling of class analogy) (Oja, 1983; Wold, 1976) are compared with CVA both theoretically and experimentally in this paper. CVA seems to yield better results in the speech recognition rates compared with the aforementioned subspace methods. This indicates that the least varying directions are more important in the pattern classifiers than the significantly varying directions (Landgrebe, 2002).

In Section 2, short reviews of the computation of other subspace methods such as FLDA, HDA, CLAFIC, SELFIC, SIMCA and CVA in the insufficient data case are given. In Section 3, the derivation of the common vector for a certain class in the sufficient data case is presented. In Section 4, CVA for sufficient data case is compared with the insufficient data case. Section 5 includes the theoretical comparison of CVA and other subspace methods. Finally, the results of the experimental study on isolated word recognition are given in Section 6.

2. Review of the previous subspace methods

2.1. FLDA and HDA

In Fisher's criterion, LDA is used to solve two-class problems by maximizing the ratio of the between-class scatter matrix $\mathbf{S}_B$ to within-class scatter matrix $\mathbf{S}_W$ in the lower-dimensional space (Bishop, 1995; Saon et al., 2000; Loog and Haeb-Umbach, 2000; Loog et al., 2001). The maximization criterion is expressed as:

$$J(W) = \mathrm{tr}\{(W^{\mathrm{T}}\mathbf{S}_W W)^{-1}(W^{\mathrm{T}}\mathbf{S}_B W)\}. \tag{1}$$

Maximization is achieved by an eigenvector decomposition of $\mathbf{S}_W^{-1}\mathbf{S}_B$ and by taking eigenvectors corresponding to nonzero eigenvalues. For a multi-class problem, the Fisher criterion is clearly suboptimal (Loog et al., 2001). The decomposition, however, adds weight to the contribution of individual class pairs to the overall criterion, in order to improve LDA. The weighting scheme is called the approximate pairwise accuracy criterion (aPAC) (Loog et al., 2001). In order to find a subspace in which a projection of the class means preserves the class distances in such a way that class separability is maintained as well as is possible, Loog et al. (2001) defined the between-class scatter matrix $\mathbf{S}_B$ as:

$$\mathbf{S}_B = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C} p_i p_j (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^{\mathrm{T}}, \tag{2}$$

where $\mathbf{m}_i$ and $\mathbf{m}_j$ are the mean vectors of classes i and j, respectively, while p_i and p_j are their priori probabilities. The term $(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^{\mathrm{T}}$ in Eq. (2) is actually the between-class scatter matrix for the classes i and j in a two-class model. Using this decomposition in Fisher criterion, it can be seen that C-class Fisher criterion can be decomposed in $(1/2)(C(C-1))$ two-class Fisher criteria. This criterion is referred to as pairwise Fisher criterion (Loog et al., 2001). Loog et al. (2001) generalized Fisher criterion by introducing a weighting function depending on the Mahanalobis distance.

The function of the Mahanalobis distance matrix is to approximate pairwise accuracy when used in the heteroscedastic discriminant analysis (HDA) (Loog and Haeb-Umbach, 2000). Saon et al. (2000) pointed out that HDA alone actually degrades the recognition performance.

2.2. CLAFIC, SELFIC and SIMCA

We will now provide an explanation of other subspace methods used for pattern recognition purposes in this section. These subspace methods were introduced by Watanabe (Watanabe et al., 1967) and have been widely used by Kohonen (Kohonen et al., 1979, 1980, 1984) with all work summarized by Oja (1983). SELFIC and CLAFIC methods are basically the same methods since the feature vectors are first normalized in both methods, the only difference being that the average feature vector of the class is subtracted from each feature vector in SELFIC, whereas this average vector is not subtracted at all in CLAFIC. Furthermore, derivation of the CLAFIC method is followed by the metric (Watanabe et al., 1967; Oja, 1983)

$$\mathbf{F}_{\text{clafic}} = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left|\mathbf{a}_i^{\mathrm{T}}\mathbf{v}_j\right|^2 = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{v}_j^{\mathrm{T}}\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}}\mathbf{v}_j, \tag{3}$$

where m is the total number of feature vectors within one class and $\mathbf{a}_i \in R^n$ is the feature vector[2] for $i = 1,\ldots,m$. The same metric can be written as:

$$\mathbf{F}_{\text{clafic}} = \frac{1}{2}\sum_{j=1}^{n}\mathbf{v}_j^{\mathrm{T}}\mathbf{Q}\mathbf{v}_j, \tag{4}$$

where $\mathbf{v}_j$'s are the orthonormal basis vectors of the whole space, and $\mathbf{Q}$ is the class correlation matrix, that is

$$\mathbf{Q} = \sum_{i=1}^{m}\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}}. \tag{5}$$

The necessary condition for the extremum of the metric $\mathbf{F}_{\text{clafic}}$ yields

$$\mathbf{Q}\mathbf{u}_{j_{\text{cor}}} = \lambda_j\mathbf{u}_{j_{\text{cor}}} \quad j = 1,2,\ldots,k-1,k,\ldots,n, \tag{6}$$

where $\lambda_{j_{\text{cor}}}$'s are the eigenvalues and $\mathbf{u}_{j_{\text{cor}}}$'s are the eigenvectors of the correlation matrix $\mathbf{Q}$, that is, $\mathbf{v}_j$'s turn out to be the eigenvectors $\mathbf{u}_{j_{\text{cor}}}$ of $\mathbf{Q}$.

The value of the metric $\mathbf{F}_{\text{clafic}}$ will be (Oja, 1983)

$$\mathbf{F}_{\text{clafic}} = \frac{1}{2}(\lambda_{1_{\text{cor}}} + \lambda_{2_{\text{cor}}} + \cdots + \lambda_{k-1_{\text{cor}}} + \lambda_{k_{\text{cor}}} + \cdots + \lambda_{n_{\text{cor}}}), \tag{7}$$

where $\lambda_{j_{\text{cor}}}$'s are the eigenvalues of $\mathbf{Q}$ and all $\lambda_{j_{\text{cor}}}$'s are in descending order.

Another method developed by Wold (1976), also called the SIMCA method, yields the following metric when minimized

$$\mathbf{F}_{\text{simca}_{\min}} = \frac{1}{2}\sum_{i=1}^{m}\|\mathbf{a}_i - \mathbf{a}_{\text{ave}} - \mathbf{P}\mathbf{a}_i\|^2 = \frac{1}{2}\sum_{i=1}^{m}\|\mathbf{P}^{\perp}\mathbf{a}_i - \mathbf{a}_{\text{ave}}\|^2, \tag{8}$$

where $\mathbf{P}$ and $\mathbf{P}^{\perp}$ are the projection matrices of the difference ($\mathbf{B}$) and indifference subspaces ($\mathbf{B}^{\perp}$), respectively, and they are obtained from the orthogonal eigenvectors of the within-class scatter matrix $\boldsymbol{\Phi}$ as given in Section 3 in Eqs. (15) and (16).

2.3. CVA in the insufficient data case

The feature vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ of a certain word-class C in the training set, which are assumed to be linearly independent, can be written as:

$$\mathbf{a}_i = \mathbf{a}_{i,\text{dif}} + \mathbf{a}_{\text{com}} + \boldsymbol{\epsilon}_i \quad \text{for} \quad i = 1,2,\ldots,m, \tag{9}$$

where the vector $\mathbf{a}_{i,\text{dif}}$ indicates inter- and intra-speaker differences as well as acoustical environmental effects and phase or temporal differences, and the vector $\mathbf{a}_{\text{com}}$ is the common vector of the word-class C, and $\boldsymbol{\epsilon}_i$ represents the error vector (Gülmezoğlu et al., 2001).

The following metric can be defined to minimize the sum of the squares of norms of the error vectors in order to obtain a solution for the common vector (Gülmezoğlu et al., 2001), that is,

$$\mathbf{F} = \frac{1}{2}\sum_{i=1}^{m}\|\boldsymbol{\epsilon}_i\|^2 = \frac{1}{2}\sum_{i=1}^{m}\|\mathbf{a}_i - \mathbf{a}_{i,\text{dif}} - \mathbf{a}_{\text{com}}\|^2, \tag{10}$$

where $\|\boldsymbol{\epsilon}_i\| = \langle\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_i\rangle^{1/2}$ denotes the Euclidean norm of the vector $\boldsymbol{\epsilon}_i$. The minimization of the metric $\mathbf{F}$ with respect to $\mathbf{a}_{\text{com}}$ yields a unique solution for $\mathbf{a}_{\text{com}}$ (Gülmezoğlu et al., 2001):

$$\mathbf{a}_{\text{com}} = \mathbf{a}_i - \mathbf{a}_{i,\text{dif}} \quad \forall\, i = 1,2,\ldots,m. \tag{11}$$

[2] For the sake of clarity in the notation, the vectors will be shown in boldface characters.

In Eq. (11), $\mathbf{a}_{i,\mathrm{dif}}$ is written as:

$$\mathbf{a}_{i,\mathrm{dif}} = \langle \mathbf{a}_i, \mathbf{u}_1 \rangle \mathbf{u}_1 + \langle \mathbf{a}_i, \mathbf{u}_2 \rangle \mathbf{u}_2 + \cdots + \langle \mathbf{a}_i, \mathbf{u}_{m-1} \rangle \mathbf{u}_{m-1}, \tag{12}$$

where $\langle \mathbf{a},\mathbf{z} \rangle$ denotes the scalar product of $\mathbf{a} \in \mathbf{R}^n$ and $\mathbf{z} \in \mathbf{R}^n$, and $\mathbf{u}_i$'s are the eigenvectors of the within-class scatter matrix. These span the difference subspace.

The common vector represents the common properties or invariant features of the word-class. The common vector does not depend on the choice of the orthonormal basis vector set of difference subspace $\mathbf{B}$ (Gülmezoğlu et al., 2001). Therefore, the common vector is unique for each class and all error vectors ϵ_i are zero. Since all of the feature vectors within one class yield the same common vector, the recognition rate for this class will always be 100% in the training set under the condition $m \leqslant n$. When m approaches n, the common vector approaches zero. If m equals to n, the common vector will be very close to the zero vector. Therefore the insufficient data case can only be applied when m is smaller than n.

3. Derivation of the common vector for the sufficient data case

In this section, it is suggested that a unique common vector can be found for each class mainly when the number of feature vectors m in the training set of one class is larger than the dimension n of the feature vectors, that is, when $m > n$. This is called the sufficient data case.

Let the difference subspace $\mathbf{B}$ be spanned by the orthonormal basis vectors $\mathbf{v}_j \in \mathbf{R}^n$ for $j = 1, 2, \ldots, k-1$ ($k - 1 < n$), and let the indifference subspace $\mathbf{B}^{\perp}$ be spanned by the orthonormal basis vectors $\mathbf{v}_j \in \mathbf{R}^n$ for $j = k, k+1, \ldots, n$. We can choose k so that the sum of the smallest eigenvalues is less than some fixed percentage L of the sum of the entire set (Oja, 1983). Thus, we let k fulfill

$$\left(\sum_{j=k}^{n} \lambda_i \right) \Big/ \left(\sum_{j=1}^{n} \lambda_i \right) < L. \tag{13}$$

If $L = 5\%$, a good performance is obtained while retaining a small proportion of the variance present in the original space (Swets and Weng, 1996). $L = 5\%$ for indifference subspace was attained at a different number of eigenvalues for each class. The average number of these eigenvalues was equal to 360.

The value of k is also determined from the point, where the eigenvalues of the training data start to vary slowly upon plotting of the eigenvalues in descending order. In Fig. 3, this point approximately corresponds to $k = 48$ ($= 407 - 360 + 1$). The orthogonal projection matrix $\mathbf{P}$ on the difference subspace $\mathbf{B}$ will be

$$\mathbf{P} = \sum_{j=1}^{k-1} \mathbf{v}_j \mathbf{v}_j^{\mathrm{T}}, \tag{14}$$

and the orthogonal projection matrix $\mathbf{P}^{\perp}$ onto the indifference subspace $\mathbf{B}^{\perp}$ will be

$$\mathbf{P}^{\perp} = \sum_{j=k}^{n} \mathbf{v}_j \mathbf{v}_j^{\mathrm{T}}. \tag{15}$$

$\mathbf{P}$ and $\mathbf{P}^{\perp}$ are symmetrical, idempotent $n \times n$ matrices, that is, $\mathbf{P} + \mathbf{P}^{\perp} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The purpose of the decomposition of whole feature space into two subspaces is to eliminate some part of the whole space having large variations from the mean (Landgrebe, 2002).

One assumption for the difference vectors $\mathbf{a}_{i,\mathrm{dif}}$ may be:

$$\mathbf{a}_{i,\mathrm{dif}} = \mathbf{P}\mathbf{a}_i = \sum_{j=1}^{k-1} \langle \mathbf{a}_i, \mathbf{v}_j \rangle \mathbf{v}_j \quad \text{for} \quad i = 1, 2, \ldots, k, \tag{16}$$

that is, $\mathbf{P}^{\perp}\mathbf{a}_{i,\mathrm{dif}} = \mathbf{0}$. The difference vector $\mathbf{a}_{i,\mathrm{dif}}$ is the projection of the feature vector $\mathbf{a}_i$ onto the difference subspace $\mathbf{B}$ (Gülmezoğlu et al., 2001). This is similar to the case obtained for the insufficient data. The other assumption is on $\mathbf{a}_{\mathrm{com}}$, that is, $\mathbf{a}_{\mathrm{com}}$ has the components only in the indifference subspace $\mathbf{B}^{\perp}$, $\mathbf{P}\mathbf{a}_{\mathrm{com}} = \mathbf{0}$, a feature which is similar to the insufficient data case. From this assumption, the following can be written as:

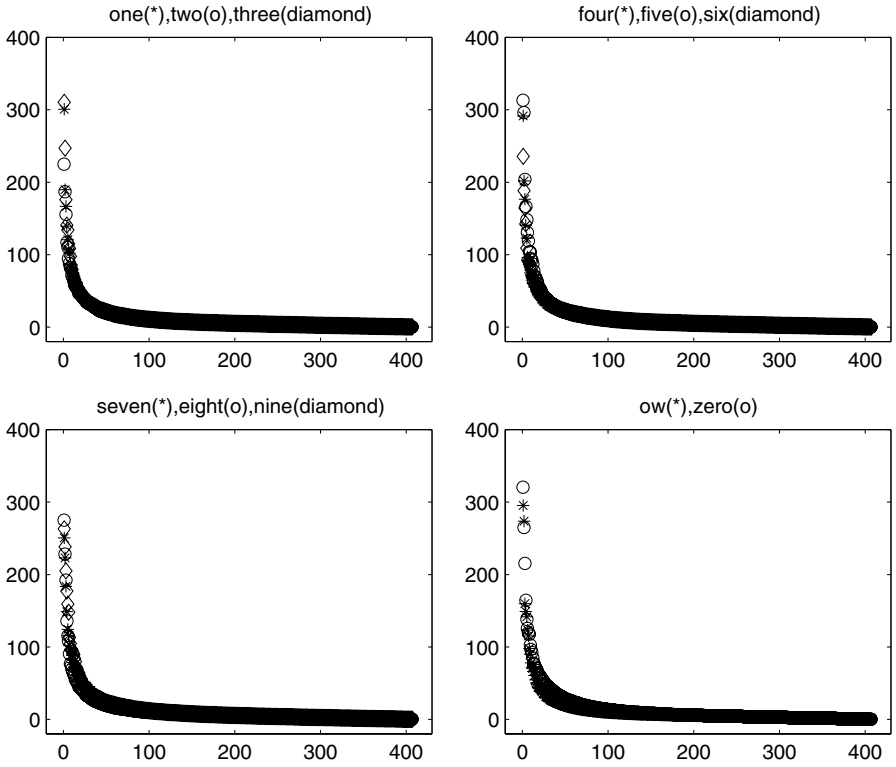

Fig. 3. The variations of the square roots of the eigenvalues of the within-class scatter matrices obtained for all digits.

$$\mathbf{a}_{\text{com}} = \mathbf{P}^{\perp}\mathbf{a}_{\text{com}} = \sum_{j=k}^{n} \langle \mathbf{a}_{\text{com}}, \mathbf{v}_j \rangle \mathbf{v}_j. \tag{17}$$

Under the above assumptions, the metric $\mathbf{F}$ given in (10) can be transformed into

$$\mathbf{F} = \frac{1}{2}\sum_{i=1}^{m} \|\mathbf{a}_i - \mathbf{P}\mathbf{a}_i - \mathbf{P}^{\perp}\mathbf{a}_{\text{com}}\|^2 = \frac{1}{2}\sum_{i=1}^{m} \|\mathbf{P}^{\perp}(\mathbf{a}_i - \mathbf{a}_{\text{com}})\|^2. \tag{18}$$

The minimization of $\mathbf{F}$ with respect to $\mathbf{a}_{\text{com}}$ then gives

$$\mathbf{a}_{\text{com}} = \mathbf{P}^{\perp}\mathbf{a}_{\text{ave}} = \mathbf{P}^{\perp}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbf{a}_i\right). \tag{19}$$

Using the above relation for the common vector $\mathbf{a}_{\text{com}}$, the metric $\mathbf{F}$ can also be written as:

$$\mathbf{F} = \frac{1}{2}\sum_{j=k}^{n} \mathbf{v}_j^{\mathrm{T}} \boldsymbol{\Phi} \mathbf{v}_j. \tag{20}$$

From the minimization of $\mathbf{F}$ with respect to $\mathbf{v}_j$ under the constraints $\|\mathbf{v}_j\| = 1$ for $j = k,\ldots,n$, the basis vectors ($\mathbf{v}_j$) of the difference and indifference subspaces will turn out to be the eigenvectors ($\mathbf{u}_j$) of the within-class scatter matrix $\boldsymbol{\Phi}$:

$$\boldsymbol{\Phi} = \sum_{i=1}^{m} (\mathbf{a}_i - \mathbf{a}_{\text{ave}})(\mathbf{a}_i - \mathbf{a}_{\text{ave}})^{\mathrm{T}}. \tag{21}$$

After minimization, the metric $\mathbf{F}$ will become

$$\mathbf{F}_{\min} = \frac{1}{2}\sum_{j=k}^{n} \mathbf{u}_j^{\mathrm{T}} \boldsymbol{\Phi} \mathbf{u}_j = \frac{1}{2}(\lambda_k + \lambda_{k+1} + \cdots + \lambda_n), \tag{22}$$

where $\lambda_k, \lambda_{k+1}, \ldots, \lambda_n$ are the smallest eigenvalues of the within-class scatter matrix $\boldsymbol{\Phi}$ and $\mathbf{u}_k, \mathbf{u}_{k+1}, \ldots, \mathbf{u}_n$ are the corresponding eigenvectors. These are also the orthonormal basis vector set of the indifference subspace $\mathbf{B}^{\perp}$.

The projection of any feature vector $\mathbf{a}_i$ onto the difference subspace $\mathbf{B}$ (Eq. (16)) yields $\mathbf{a}_{i,\mathrm{dif}}$, which has components only in the significantly varying directions of the class C. The basis vectors of $\mathbf{B}$ are the eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$ corresponding to the largest eigenvalues of $\boldsymbol{\Phi}$.

The common vector $\mathbf{a}_{\mathrm{com}}$ (Eq. (19)) is obtained for each class separately using the projection of the average vector $\mathbf{a}_{\mathrm{ave}}$ of the class C onto the indifference subspace $\mathbf{B}^{\perp}$, where the variation of the feature vectors is the smallest.

The idea proposed in this section is explained with a sample class in a two-dimensional feature space by the following example.

Example. Let the feature vectors of the class C be:

$$\mathbf{a}_1 = [0 \quad 3]^{\mathrm{T}} \quad \mathbf{a}_2 = [0 \quad 1]^{\mathrm{T}} \quad \mathbf{a}_3 = [3 \quad 2]^{\mathrm{T}}.$$

Then the average vector and the within-class scatter matrix of this class are given below with the eigenvalues and eigenvectors,

$$\mathbf{a}_{\mathrm{ave}} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \boldsymbol{\Phi} = \begin{bmatrix} 6 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\lambda_1 = 6 \quad \lambda_2 = 2$$

$$\mathbf{u}_1 = [1 \quad 0]^{\mathrm{T}} \quad \mathbf{u}_2 = [0 \quad 1]^{\mathrm{T}}.$$

The class C is shown with its typical constant probability density contours in Fig. 4. For class C, the difference in the $\mathbf{u}_1$ direction will be discarded, and only the components in the direction $\mathbf{u}_2$ will be retained for the common vector.

The common vector $\mathbf{a}_{\mathrm{com}}$ from (19), the difference vectors $\mathbf{a}_{1,\mathrm{dif}}$, $\mathbf{a}_{2,\mathrm{dif}}$, $\mathbf{a}_{3,\mathrm{dif}}$ from (16) and the error vectors $\boldsymbol{\epsilon}_1$, $\boldsymbol{\epsilon}_2$, $\boldsymbol{\epsilon}_3$ from (9) can be written as:

$$\mathbf{a}_{\mathrm{com}} = \left(\mathbf{a}_{\mathrm{ave}}^{\mathrm{T}}\mathbf{u}_2\right)\mathbf{u}_2 = 2[0 \quad 1]^{\mathrm{T}} = [0 \quad 2]^{\mathrm{T}}.$$

$$\mathbf{a}_{1,\mathrm{dif}} = \mathbf{u}_1\left(\mathbf{u}_1^{\mathrm{T}}\mathbf{a}_1\right) = [1 \quad 0]^{\mathrm{T}}\left([1 \quad 0][0 \quad 3]^{\mathrm{T}}\right) = [0 \quad 0]^{\mathrm{T}}.$$

$$\mathbf{a}_{2,\mathrm{dif}} = \mathbf{u}_1\left(\mathbf{u}_1^{T}\mathbf{a}_2\right) = [1 \quad 0]^{\mathrm{T}}\left([1 \quad 0][0 \quad 1]^{\mathrm{T}}\right) = [0 \quad 0]^{\mathrm{T}}.$$

$$\mathbf{a}_{3,\mathrm{dif}} = \mathbf{u}_1\left(\mathbf{u}_1^{\mathrm{T}}\mathbf{a}_3\right) = [1 \quad 0]^{\mathrm{T}}\left([1 \quad 0][3 \quad 2]^{\mathrm{T}}\right) = [3 \quad 0]^{\mathrm{T}}.$$

$$\boldsymbol{\epsilon}_1 = \mathbf{a}_1 - \mathbf{a}_{1,\mathrm{dif}} - \mathbf{a}_{\mathrm{com}} = [0 \quad 3]^{\mathrm{T}} - [0 \quad 0]^{\mathrm{T}} - [0 \quad 2]^{\mathrm{T}} = [0 \quad 1]^{\mathrm{T}}.$$

$$\boldsymbol{\epsilon}_2 = \mathbf{a}_2 - \mathbf{a}_{2,\mathrm{dif}} - \mathbf{a}_{\mathrm{com}} = [0 \quad 1]^{\mathrm{T}} - [0 \quad 0]^{\mathrm{T}} - [0 \quad 2]^{\mathrm{T}} = [0 \quad -1]^{\mathrm{T}}.$$

$$\boldsymbol{\epsilon}_3 = \mathbf{a}_3 - \mathbf{a}_{3,\mathrm{dif}} - \mathbf{a}_{\mathrm{com}} = [3 \quad 2]^{\mathrm{T}} - [3 \quad 0]^{\mathrm{T}} - [0 \quad 2]^{\mathrm{T}} = [0 \quad 0]^{\mathrm{T}}.$$

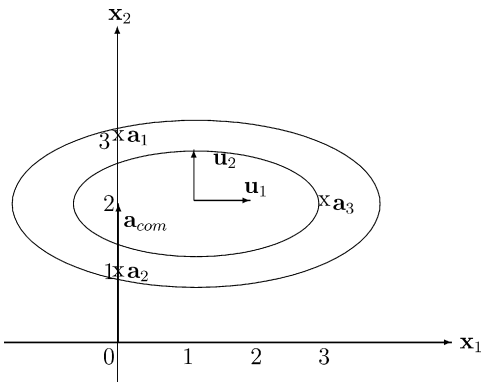

Fig. 4. One class with its equal probability density contours (ellipses) and common vector.

Then the minimum value of the CVA metric $\mathbf{F}_{\min}$ will be equal to $\lambda_2 = 2$. The common vector is shown in Fig. 4. It should be noted that the common vector is in the direction of the eigenvector $\mathbf{u}_2$, which belongs to the smallest eigenvalue. The common vector is the projection of the center of the ellipses in Fig. 4 onto the eigenvector $\mathbf{u}_2$.

Taking the projection of the feature vectors $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$ onto the indifference subspace eliminates the larger variations of the equal probability density contours. That is, these projections will be closer to the projection of the class average $\mathbf{a}_{\text{com}}$ when compared with the original whole space.

Since the number of data items exceeds the dimension of the feature vectors, zero eigenvalue cannot be found if the feature vectors are linearly independent. Therefore, the common vector of the sufficient case cannot be calculated using the mathematical derivation of the insufficient case.

4. Theoretical comparison with the insufficient data case

The CVA metric given in Eq. (10) is also valid for the insufficient data case. In this case, since $m \leqslant n$, the number of zero eigenvalues obtained from the within-class scatter matrix is equal to $n - m + 1$. If the eigenvalues are in descending order, then the minimum value of the metric will be zero:

$$\mathbf{F}_{\min} = \frac{1}{2}(\lambda_m + \lambda_{m+1} + \cdots + \lambda_n) = 0. \tag{23}$$

This value of the metric guarantees a 100% recognition rate in the training set. In the sufficient case, the minimum value of the metric is not zero because the within-class scatter matrix has all nonzero eigenvalues. Therefore the metric does not guarantee a 100% recognition rate in the training set. However, the experimental study indicates that a 100% recognition rate in the training set can be obtained when the value of k in (15) is properly selected.

Further important difference is the fact that the common vector in the insufficient data case is obtained by taking the projection of any feature vector onto the indifference subspace, whereas the common vector in the sufficient data case is obtained by taking the projection of the average vector onto the indifference subspace.

5. Theoretical comparison with the other subspace methods

The comparison of the above mentioned nonnegative within-class metrics should be conducted under certain reasonable assumptions. Maximization of the metrics may have a value which is too large and not meaningful for recognition purposes. Since all of the aforementioned metrics will yield a number which is greater than or equal to zero after minimization, one can assume that the metric yielding a smaller number must be better than the other metrics for recognition purposes. In fact, in our previous work, we demonstrated that the metric producing a zero under minimization will yield a 100% recognition rate for the training set (Gülmezoğlu et al., 2001).

5.1. Comparison with the CLAFIC

In this case, the first $(k-1)$ eigenvalues of $\mathbf{\Phi}$ constitute the largest part of the metric $\mathbf{F}_{\text{clafic}}$. This will be called $\mathbf{F}_{\text{clafic}_{\max}}$:

$$\mathbf{F}_{\text{clafic}_{\max}} = \frac{1}{2}\sum_{i=1}^{m} \mathbf{a}_i^{\mathrm{T}} \mathbf{P}_{\text{cor}} \mathbf{a}_i = \frac{1}{2}(\lambda_{1_{\text{cor}}} + \lambda_{2_{\text{cor}}} + \cdots + \lambda_{k-1_{\text{cor}}}), \tag{24}$$

where $\mathbf{P}_{\text{cor}}$ is the projection matrix of the subspace $\mathbf{B}$ and is obtained from the eigenvectors corresponding to the largest eigenvalues of the correlation matrix $\mathbf{Q}$. The choice of k is given in Oja (1983).

Although the within-class distributions will effect the magnitudes of these metrics, a sight can be gained by just comparing the summation of the related eigenvalues for the CLAFIC and CVA metrics.

The minimum of the CLAFIC metric can be defined as:

$$\mathbf{F}_{\text{clafic}_{\min}} = \frac{1}{2}\sum_{j=k}^{n} \mathbf{u}_{j_{\text{cor}}}^{\text{T}} \mathbf{Q}\mathbf{u}_{j_{\text{cor}}} = \frac{1}{2}(\lambda_{k_{\text{cor}}} + \cdots + \lambda_{n_{\text{cor}}}). \tag{25}$$

The minimum of the CVA metric was given in (22). Since the correlation and within-class scatter matrices are related by

$$\mathbf{Q} = \mathbf{\Phi} + \mathbf{Q}_{\text{ave}}, \tag{26}$$

where $\mathbf{Q}_{\text{ave}} = \sum_{i=1}^{m} \mathbf{a}_{\text{ave}}\mathbf{a}_{\text{ave}}^{\text{T}} = m\mathbf{a}_{\text{ave}}\mathbf{a}_{\text{ave}}^{\text{T}}$, and the following can be written as:

$$\text{tr}\mathbf{Q} = \text{tr}(\mathbf{\Phi} + \mathbf{Q}_{\text{ave}}). \tag{27}$$

From here, one can write the following:

$$\lambda_{1_{\text{cor}}} + \cdots + \lambda_{n_{\text{cor}}} = \lambda_1 + \cdots + \lambda_n + m\|\mathbf{a}_{\text{ave}}\|^2. \tag{28}$$

Therefore,

$$\lambda_{1_{\text{cor}}} + \cdots + \lambda_{n_{\text{cor}}} \geqslant \lambda_1 + \cdots + \lambda_n. \tag{29}$$

It must then be expected that after minimization, the metric $\mathbf{F}_{\min}$ of CVA is smaller than the metric $\mathbf{F}_{\text{clafic}_{\min}}$ in general.

5.2. Comparison with the SELFIC

The SELFIC method starts with the subtraction of the average vector from each feature vector at the initial step. The correlation matrix $\mathbf{Q}$ of the CLAFIC method will be substituted with the within-class scatter matrix $\mathbf{\Phi}$ in the SELFIC method. The following two mathematical operations should not have been conducted in the SELFIC method, one is the normalization of the feature vectors (radially aligned classes will overlap on the hypersphere of the feature space in this case), and the other is the initial subtraction of the average vector which makes the common vector for any class a zero vector.

Another difference between SELFIC and CVA is that SELFIC maximizes its metric (Watanabe et al., 1967) whereas CVA minimizes it (Gülmezoğlu et al., 2001). The maximization of the metrics does not give us any hint to recognition rates whereas minimization of the metrics gives us a certain hint as it is stated in Gülmezoğlu et al. (1999).

5.3. Comparison with the SIMCA

Comparison between the minimums of SIMCA and CVA can also be made from (8):

$$\mathbf{F}_{\text{simca}_{\min}} = \frac{1}{2}\sum_{i=1}^{m} \mathbf{a}_i^{\text{T}}\mathbf{P}^{\perp}\mathbf{a}_i - 2m\mathbf{a}_{\text{ave}}^{\text{T}}\mathbf{P}^{\perp}\mathbf{a}_{\text{ave}} + m\mathbf{a}_{\text{ave}}^{\text{T}}\mathbf{a}_{\text{ave}} \tag{30}$$

and from (18)

$$\mathbf{F}_{\min} = \frac{1}{2}\sum_{i=1}^{m} \mathbf{a}_i^{\text{T}}\mathbf{P}^{\perp}\mathbf{a}_i - 2m\mathbf{a}_{\text{ave}}^{\text{T}}\mathbf{P}^{\perp}\mathbf{a}_{\text{ave}} + m\mathbf{a}_{\text{ave}}^{\text{T}}\mathbf{P}^{\perp}\mathbf{a}_{\text{ave}}. \tag{31}$$

Obviously, only the third terms differ and the third term of the metric $\mathbf{F}_{\min}$ of CVA is smaller than the third term of the metric $\mathbf{F}_{\text{simca}_{\min}}$. Therefore, $\mathbf{F}_{\min} \leqslant \mathbf{F}_{\text{simca}_{\min}}$.

5.4. Comparison with the FLDA

Theoretical comparison of CVA with previous subspace methods was possible since in all of these subspace methods the metrics contain only the within-class scatter. None of these previous subspace methods contains the between- or total-class scatters. Since FLDA employs the between-class scatter in its metric, it cannot be compared with the CVA method theoretically. If the comparison between FLDA and CVA is desired, one has

to make priori assumptions about the between- or total-class scatters. No such priori assumptions have been encountered in the literature by the authors.

6. Decision criterion and experimental results

6.1. Decision criterion for CVA

CVA developed for the sufficient data in this paper, CLAFIC and SELFIC are applied to the TI-digit database for speech recognition purposes. The decision criteria for the CVA method is given below.

For an unknown feature vector $\mathbf{a}_x$, a vector, called the remaining vector $\left(\mathbf{a}^l_{x,\mathrm{rem}}\right)$, can be defined as:

$$\mathbf{a}^l_{x,\mathrm{rem}} = \sum_{j=k}^{n} \left\langle \mathbf{a}_x, \mathbf{u}^l_j \right\rangle \mathbf{u}^l_j = (\mathbf{P}^{\perp})^l \mathbf{a}_x, \tag{32}$$

where l denotes the index of the classes. The Euclidean distance between the common vectors and the remaining vectors is employed as the decision criterion in CVA, and is given with the following formula:

$$\mathbf{C}^* = \underset{l}{\operatorname{argmin}} \left\| \mathbf{a}^l_{x,\mathrm{rem}} - \mathbf{a}^l_{\mathrm{com}} \right\| = \underset{l}{\operatorname{argmin}} \left\| (\mathbf{P}^{\perp})^l (\mathbf{a}_x - \mathbf{a}^l_{\mathrm{ave}}) \right\|, \tag{33}$$

that is, if the feature vector $\mathbf{a}_x$ belongs to the class C^l, the distance between $\mathbf{a}^l_{x,\mathrm{rem}}$ and $\mathbf{a}^l_{\mathrm{com}}$ should be a minimum.

6.2. Experimental results

All of the previous methods have been applied to the TI-digit database consisting of 11 isolated digits. In our TI-digit database, there are 112 speakers repeating each digit twice in the TI-training set and 111 speakers repeating each digit twice in the TI-test set. Since we need to have $m > n$, our training set is constructed by taking 224 repetitions from the TI-training set and 202 repetitions from the TI-test set. The remaining 20 repetitions in the TI-test set are used in our test set for each digit. Therefore, the repetitions, i.e. the feature vectors, in the test set and the feature vectors in the training set are completely disjoint in the experiments. An equal number of male and female speakers is used in the training set (213 male and 213 female) and in the test set (10 male and 10 female).

After end-point detection, the speech frames consisting of 256 samples are pre-emphasized and analyzed to calculate 11 root-melcep parameters. These parameters are then stacked in order to construct the feature vector for each repetition of each digit. After this process, the dimension of the feature vector for each digit varied from 110 to 407 ($110 \leqslant n \leqslant 407$). Therefore, when the dimension of each feature vector is less than 407, it is extended to 407 by padding random values only at the end of the vector. The digits "four", "three" and "ow" require the most padding. The average numbers of padded values for these digits are 242.15, 239.74 and 229.28, respectively. The overall percentage of padding in the database is 50.7%. The within-class scatter matrix $\mathbf{\Phi}$ with a size of 407×407, its eigenvalues and eigenvectors are calculated using $m = 426$ feature vectors in each class. Since 20 feature vectors in the test set are too few to determine the recognition accuracy, the leave-twenty-out method is applied instead of the leave-one-out method. Thus, the testing process is repeated 11 times to cover all the repetitions in the TI test set. The average recognition rates obtained from these iterations are given for the training and test sets. Since all the eigenvalues of the within-class scatter matrix are found to be nonzero, the eigenvectors corresponding to the different numbers of the smallest eigenvalues $(n - k + 1)$ in the CVA and CLAFIC method are used in the recognition process. Whereas the eigenvectors corresponding to the different numbers of the largest eigenvalues $(k - 1)$ in the SELFIC method are taken in the recognition stage.

As mentioned previously in Section 3, the choice of the number k, which determines the dimensions of the indifference subspaces, can be conducted using Eq. (13) with $L = 5\%$. The value of k can also be determined approximately by specifying the point where the eigenvalues for all the digit classes start to vary slowly, as seen in Fig. 3. For the TI-digit database, the value of k is determined as 48 $(= 407 - 360 + 1)$ when the above

criteria are applied. The recognition rates of CVA and CLAFIC are given in Table 1 for this value of k. The recognition rates of the SELFIC method in Table 1 are given for $k = 360$ $(407 - 48 + 1)$ since the eigenvectors corresponding to the largest eigenvalues are used in the SELFIC method. Since the digit "four" overlaps with the digits "five"' and "ow" in the SELFIC method, the digit "four" is recognized as "five" or "ow". Therefore, the recognition rate is zero for this digit as seen from Table 1. In order to see the effect of different selections of k for the CVA and CLAFIC methods, the value of k is changed between 8 $(= 407 - 400 + 1)$ and 58 $(= 407 - 350 + 1)$. The results are given in Figs. 5 and 6 for the training and test sets, respectively. As can be seen from these figures, although the recognition rates for the training set decrease when k decreases, the recognition rate for the test set reaches a maximum value of 98.85% in CVA when k reduces to 28.

When CVA and CLAFIC are used in the classification of feature vectors in the training set, 100% recognition rates are obtained for the first four smallest eigenvalues. On the other hand, SELFIC gives the maximum recognition rate of 92.38% with the 350 largest eigenvalues.

If the pairwise Fisher criterion is used in the experimental study, 96.54% and 92.27% recognition rates are obtained for the training and test sets, respectively. When Mahanalobis based aPAC ($\gamma = 0.5$) is used in the experimental study, recognition rates reduce to 91.72% for the training set and to 87.27% for the test set.

In this study, we have also compared the subspace method CVA and the whole space method HMM only with respect to their classification performances and processing times. HMM with continuous densities (CDHMM) is used to model each class in the TI-digit database. We have selected a left-to-right model without skipping, and employed a diagonal within-class scatter matrix for multivariate Gaussian density functions. HMMs were trained using twenty iterations of the Baum–Welch method. The same number of frames (37) and, same training and test sets were used as in the CVA. No random values are padded to the end of the utterances when applying the HMM method; that is, words with short duration remained short and words with long duration remained long. HMM models are trained using a different number of states (Q) and mixtures (M). In order to find the best parameter setting, the development test set (devtest) is constructed by taking 20% of the training data. The highest recognition rates are obtained from the devtest for $Q = 6$ and $M = 8$.

The leave-twenty-out method is also applied to HMM. The average recognition rates obtained from 11 different training and test sets are given in Table 1. These rates are slightly lower than the rates obtained from the CVA method. Therefore, CVA outperforms HMM slightly.

CVA requires much shorter processing time in the training and test phases. We have implemented CVA and HMM in Matlab and measured the processing time of the training process for CVA and HMM in a personal computer, Pentium IV with 3 GHz and 1 Gbyte RAM. The processing time of CVA and HMM in the training phase per class is measured as 4.1 s and 224 s, respectively. The processing time in the test phase is measured as 0.015 s for CVA and 0.18 s for HMM.

For longer isolated words in the vocabulary, one should normally use the insufficient data case of CVA. Another suggestion could be to use only a few initial phonemes of those words. In order to realize this case, the experimental study is continued by taking, not all, but only the first 120 elements of the feature vectors

Table 1
Average recognition rates obtained by using the CVA, CLAFIC and SELFIC methods as percentages for $n = 407$ and $k = 48$

Words	Training set				Test set			
	CVA	CLAF	SELF	HMM	CVA	CLAF	SELF	HMM
One	100.00	99.89	90.63	99.60	98.18	97.73	77.27	98.64
Two	99.53	99.68	96.07	99.34	99.10	99.10	87.73	97.27
Three	99.74	100.00	99.10	99.72	98.18	96.82	88.18	99.55
Four	99.79	99.79	0.00	99.21	96.82	96.82	0.00	97.73
Five	99.36	99.89	96.86	98.91	96.82	96.82	90.00	93.64
Six	99.23	99.76	93.77	99.96	98.18	97.73	87.73	100.00
Seven	99.77	100.00	97.18	99.79	97.73	98.18	89.55	99.09
Eight	100.00	99.98	90.31	99.75	99.55	98.64	80.45	99.09
Nine	98.80	99.70	98.61	98.93	95.91	96.82	91.36	98.18
Ow	100.00	100.00	97.95	97.85	100.00	98.64	95.91	96.36
Zero	99.34	100.00	88.28	99.73	99.10	99.10	80.45	98.18
Average	99.60	99.87	86.25	99.34	98.14	97.85	78.97	97.97

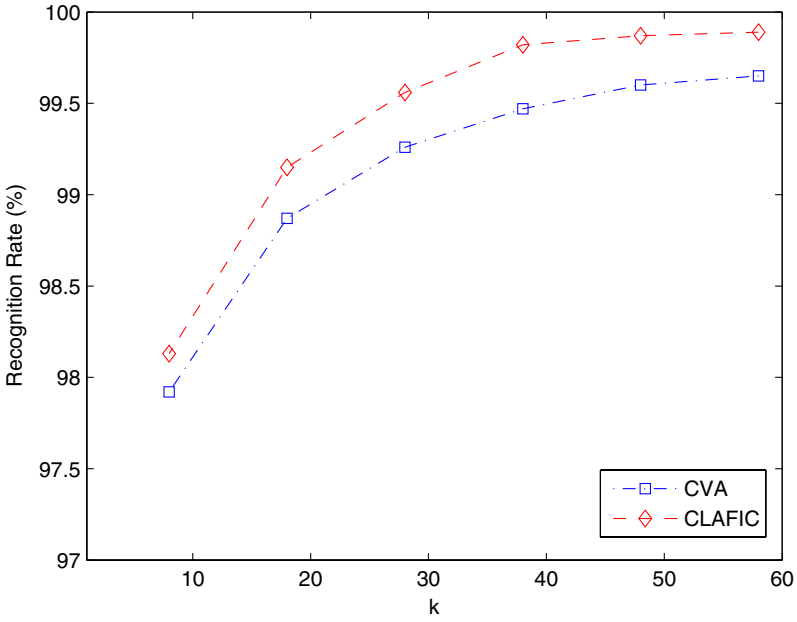

Fig. 5. Recognition rates for the CVA and CLAFIC methods for the training set.

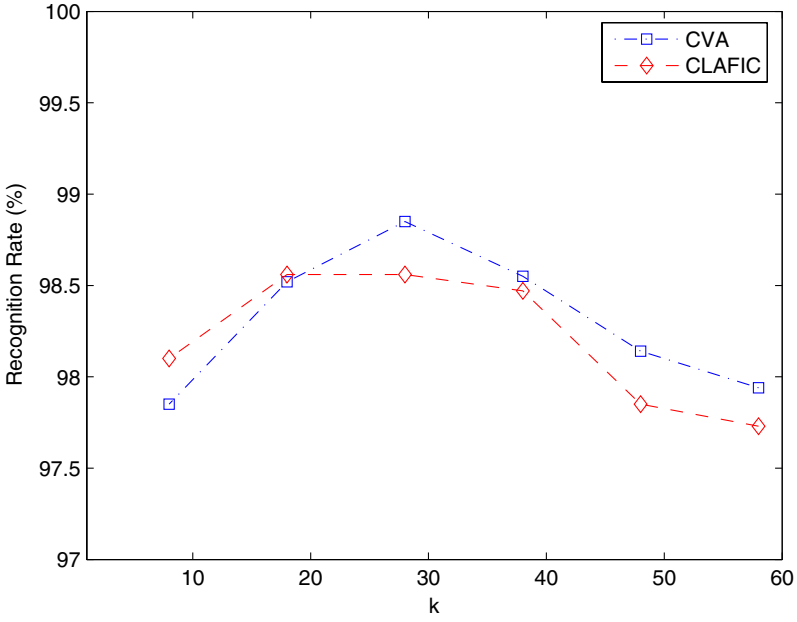

Fig. 6. Recognition rates for the CVA and CLAFIC methods for the test set.

($n = 120$). If the dimension of the feature vectors is less than 120, some random values are padded at the end of these feature vectors. Generally, the feature vectors with the dimension of 120 correspond to the first two or three phonemes of the digits. The number of the feature vectors in each class in the training set was taken as $m = 224$. As mentioned above, the value of k is determined approximately by specifying the point where the eigenvalues for all the digit classes start to vary slowly. The recognition rates of the CVA and CLAFIC methods for $k = 16$ $(120 - 105 + 1)$ and the recognition rates of the SELFIC method for $k = 105$ $(120 - 16 + 1)$ are given in Table 2 for the training and test sets. Similar recognition rates are obtained for this case, with the exception of a small decrease in the test set. The results are still satisfactory to build a CVA classifier.

Table 2
Average recognition rates obtained by using the CVA, CLAFIC and SELFIC methods as percentages for $n = 120$ and $k = 16$

Words	Training set			Test set		
	CVA	CLAF	SELF	CVA	CLAF	SELF
One	99.55	99.55	99.11	98.65	98.65	74.77
Two	98.66	98.66	79.46	99.10	97.30	49.55
Three	98.66	100.00	99.55	96.85	95.95	69.37
Four	99.55	100.00	7.14	96.40	96.85	2.70
Five	100.00	100.00	99.55	97.30	98.20	74.32
Six	94.20	93.30	68.30	83.33	79.28	33.33
Seven	97.77	98.66	96.88	94.14	89.64	51.35
Eight	100.00	100.00	90.18	98.20	97.30	59.91
Nine	98.21	98.21	94.20	97.75	96.85	66.22
Ow	100.00	99.55	66.07	98.20	97.30	31.98
Zero	99.11	99.55	97.32	98.20	96.85	67.57
Average	98.70	98.86	81.62	96.19	94.92	52.83

Table 3
The eigenvalues (λ) of the digit "one" in the training set for the CVA and CLAFIC methods

λ	CVA	CLAFIC
1	2.0263833	2.0968891
5	4.3241740	4.3251429
10	6.9388887	7.0029827
20	13.840821	14.211669
30	26.455506	26.970437
40	45.454204	48.171274
50	84.644623	88.054337
60	135.77382	137.72936
70	236.11065	236.17084
80	420.64866	421.14791
90	875.92080	897.92442
100	1872.4006	3107.7840
105	3451.5778	3481.2727
110	6364.2895	6547.9868
115	11954.141	15585.626
117	29801.539	32413.348
118	32795.068	38886.823
119	40678.009	109903.71
120	110541.02	676866.59

Thus, the experimental results verify the theoretical comparison of CVA with other subspace methods. The eigenvalues of the digit "one" in CVA, given in Table 3, are smaller than those of the CLAFIC method. Therefore, the summation of the eigenvalues in CVA always gives smaller numbers than the summation of the eigenvalues in the CLAFIC method. This verifies that CVA is better than CLAFIC for recognition purposes.

7. Conclusion and discussion

Both the theoretical development of CVA and its comparison with other subspace methods are given in this paper for the sufficient number of data items in the training set. The common vector formulated in this paper is in accordance with previous derivations for the insufficient data. Therefore, the common vector is in the direction of the eigenvectors that belong to the smallest (including zero) eigenvalues of the within-class scatter matrix of a class.

Table 1 shows that a recognition rate of 99.6% is obtained for $k = 48$ in the training set. It is also obvious from the table that CVA is able to achieve the recognition rate of 98.85% for $k = 28$ in the test set. These results are obtained when the digits are taken in full length, and when the short digits with less than 407 features are padded with small random numbers.

The main advantages of CVA compared to HMM are that CVA is easy to implement and does not require such complex operations as HMM. If one wants to avoid the burden of calculation of probability density functions while building a classifier, CVA is a reasonable choice. But in the CVA method, a whole utterance is treated as a single vector instead of a sequence of independent vectors. Therefore, the number of parameters in CVA is larger than that in HMM. This is a disadvantage of CVA because the memory requirement of CVA is higher than that of HMM. Since the state-of-the-art HMM requires feature vectors augmented with delta and delta-squared parameters, we were not able to compare it with the sufficient case of CVA. If CVA uses higher dimensional vectors then the sufficient case reverts to the insufficient case. The recognition rates of CVA for the insufficient case are lower than the results given in this paper for HMM for the TI-digit database (Gülmezoğlu et al., 2001).

CVA does not encounter scatters in the other classes at all, in similar with HMM, besides its own class. The CVA method uses an indifference subspace for each of the classes separately, whereas FLDA in general calculates only one single subspace for all the classes in the training set. This may reduce recognition rates compared to the CVA method. For the class scatters given in Fig. 1, considering only within-class scatter as in CVA, gives better classification rates than that of considering both within- and between-class scatters as in FLDA. It is obvious that the scatter of the classes in any database will affect the performance of classifiers, that is, FLDA can give better results than CVA for different class scatters, as shown in Fig. 2.

The higher performance of CVA can be obtained for different selections of k. This suggestion can be verified from increasing recognition rates given in Figs. 5 and 6. One can see that increasing the dimension improves the recognition rates up to $k = 28$. When the number of dimensions of the indifference subspace becomes closer to the number of dimensions of the whole feature space, recognition rates start to fall appreciably.

Another important point is that normalization of the feature vectors in the SELFIC and CLAFIC methods is not necessary, and in fact, should not be performed because the radially aligned classes will overlap on the hypersphere upon normalization procedure. From the experimental work we have seen that the digit "four" is radially aligned with the digits "five" and "ow", and this can be seen from Table 1, since recognition rates of the digit "four" drop to zero in the SELFIC method after normalization. It is clear that CVA is also resistant to damage caused by normalization for the case of the insufficient number of data items, if classes are radially aligned (Gülmezoğlu et al., 2001).

The authors are aware of the fact that CVA extracts temporal variations in the utterance, as well as all the other differences. The least varying directions are more important in pattern classifiers than the largely varying directions (Landgrebe, 2002) and a separate indifference subspace must be calculated for each class.

Acknowledgements

The authors thank H.H. Erkaya and Karin Marsden for extensive English editing of the manuscript, and two anonymous reviewers and Editor for their critical comments which greatly helped to improve the presentation of the paper. This work was supported by the Research Fund of Eskişehir Osmangazi University.

References

Barkana, A., Gülmezoğlu, M.B., Edizkan, R., 1995. Work done in Osmangazi University on speech recognition and building a database. In: Proceedings of the Workshop on Speech Processing. Ankara, Turkey, pp. 37–47.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition, first ed. Clarendon Press, Oxford.

Deller, J.R., Proakis, J.G., Hansen, J.H.L., 1993. Discrete-Time Processing of Speech Signals. Macmillan Publishing Company, New York.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press, San Diego.

Gülmezoğlu, M.B., Barkana, A., 1998. Text-dependent speaker recognition by using Gram-Schmidt Orthogonalization method. In: Proceedings of the IASTED International Conference on Signal Processing and Communications. Canary Islands, Spain, pp. 438–440.

Gülmezoğlu, M.B., Dzhafarov, V., Keskin, M., Barkana, A., 1999. A novel approach to isolated word recognition. IEEE Transactions on Speech and Audio Processing 7, 620–628.

Gülmezoğlu, M.B., Dzhafarov, V., Barkana, A., 2001. The common vector approach and its relation to the principal component analysis. IEEE Transactions on Speech and Audio Processing 9, 655–662.

Haeb-Umbach, R., Ney, H., 1992. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In: Proceedings of the IEEE International Conference on Acoustic, Speech, Signal Processing, pp. 13–16.

Keskin, M., Gülmezoğlu, M.B., Barkana, A., 1995. Removal of various differences in isolated word recognition by using Gram-Schmidt method. In: Proceedings of the Third National Conference on Signal Processing and Applications. Cappodocia, Turkey, pp. 25–30.

Keskin, M., Gülmezoğlu, M.B., Parlaktuna, O., Barkana, A., 1995. Isolated word recognition by extracting personal differences. In: Proceedings of the Sixth International Conference on Signal Processing Applications and Technology. Boston, USA, pp. 1989–1992.

Kohonen, T., Nemeth, G., Bry, K.-J., Jalanko, M., Riittinen, H., 1979. Spectral classification of phonemes by learning subspaces. In: Proceedings of the IEEE International Conference on Acoustic, Speech, Signal Processing, pp. 97–100.

Kohonen, T., Riittinen, H., Jalanko, M., Reuhkala, E., Haltsonen, S., 1980. A thousand-word recognition system based on the learning subspace method and redundant hash addressing. In: Proceedings of the Fifth International Conference on Pattern Recognition, pp. 158–165.

Kohonen, T., Riittinen, H., Reuhkala, E., Haltsonen, S., 1984. On-line recognition of spoken words from a large vocabulary. Information Sciences 33, 3–30.

Kuhn, R., Junqua, J.C., Nguyen, P., Niedzielski, N., 2000. Rapid speaker adaptation in eigenvoice space. IEEE Transactions on Speech and Audio Processing 8, 695–707.

Landgrebe, D.A., 2002. Hyperspectral image data analysis. IEEE Signal Processing Magazine 19, 17–28.

Lee, C., Landgrebe, D.A., 1993. Feature extraction based on decision boundaries. IEEE Transactions on Pattern Analysis and Machine Intelligence 15, 388–400.

Loog, M., Haeb-Umbach, R., 2000. Multi-class linear dimension reduction by generalized Fisher criteria. In: Proceedings of the International Conference on Spoken Language Processing, pp. 1069–1072.

Loog, M., Duin, R.P.W., Haeb-Umbach, R., 2001. Multi-class linear dimension reduction by weighted pairwise Fisher criteria. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 762–766.

Marrison, D.F., 1967. Multivariate Statistical Methods. McGraw Hill, New York.

Oja, E., 1983. Subspace Methods of Pattern Recognition. John Wiley & Sons Inc., New York.

Parsons, T.W., 1986. Voice and Speech Processing, first ed. McGraw Hill, New York.

Rabiner, L., Juang, B-H., 1993. Fundamentals of Speech Recognition. Prentice-Hall Inc., New Jersey.

Saon, G., Padmanabhan, M., Gopinath, R., Chen, S., 2000. Maximum likelihood discriminant feature spaces. In: Proceedings of the IEEE International Conference on Acoustic, Speech, Signal Processing, pp. 1747–1750.

Schukat-Talamazzini, E.G., Hornegger, J., Niemann, H., 1995. Optimal linear feature transformations for semi-continuous hidden Markov models. In: Proceedings of the IEEE International Conference on Acoustic, Speech, Signal Processing, pp. 369–372.

Swets, D.L., Weng, J., 1996. Using discriminant eigenfeatures for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 18, 831–836.

Tou, J.T., Gonzales, R.C., 1974. Pattern Recognition Principles, second ed. Addison-Wesley, Massachusetts.

Watanabe, S., Lambert, P.F., Kulikowski, C.A., Buxton, J.L., Walker, R., 1967. Evaluation and selection of variables in pattern recognition. In: Tou, J.T. (Ed.), Computer and Information Sciences II. Academic Press, New York, pp. 91–122.

Wold, S., 1976. Pattern recognition by means of disjoint principal component models. Pattern Recognition 8, 127–139.

Yang, J., Yang, J.Y., Zhang, D., 2002. What's wrong with Fisher criterion. The Journal of the Pattern Recognition Society 35, 2665–2668.