ELSEVIER

Pattern Recognition Letters 17 (1996) 803-809

Pattern Recognition Letters

Feature selection for pattern classification with Gaussian mixture models: A new objective criterion

S. Krishnan, K. Samudravijaya, P.V.S. Rao *

Computer Systems and Communications Group, Tata Institute of Fundamental Research, Homi Bhabha Road, Bombay 400 005, India

Received 18 July 1995; revised 1 February 1996

Abstract

The selection of a feature set is an important aspect of the pattern classification process. The Fisher ratio is commonly used to rank features with respect to their effectiveness for a given classification task. The procedure used implicitly assumes a symmetric and unimodal probability density for each class. In this paper, we propose a generalized definition of the Fisher ratio as applicable to Gaussian mixture densities, which can represent multi-modal or skewed distributions. The validity and usefulness of the proposed definition is tested by a Monte Carlo simulation experiment. The correlation between the classification results and the proposed objective criterion is found to be better than that attained with the conventional uni-modal measure.

Keywords: Feature selection; Pattern classification; Gaussian mixture models

1. Introduction

Pattern classification is the process of assigning a given test pattern of unknown affiliation to one of M classes. In order to achieve this, a set of features is extracted from the test pattern and is used to compute the a posteriori probability of each of the M classes conditioned on the test pattern. The test pattern is assigned to the class having the highest a posteriori probability. The correctness of such classification depends on the extent of confusion of the classes in the given feature space. The choice of an appropriate feature set is an important aspect of the pattern classification process. A larger feature set may yield better performance, but could incur higher computation cost. It is therefore desirable to select a subset of the feature set without unduly sacrificing classification accuracy.

The quality of a feature can be evaluated based on two considerations: (a) pattern generation knowledge, and (b) statistics. From the viewpoint of knowledge, a feature is considered to be good if it succeeds in characterising the pattern classes. From the statistical viewpoint, each feature can be evaluated by its ability to provide (i) large inter-class separation, and (ii) small intra-class spread (Duda and Hart, 1973). A judicious combination of these two criteria enables us to prune a large set of features into an effective subset of a smaller number of features. This has the following advantages. Firstly, a smaller set of features implies fewer free parameters of the classifier and, hence, a more reliable estimation is possible with limited training data. Secondly, dimensionality reduction also reduces the computational load during classification: a

* Corresponding author. E-mail: rao@tifrvax.tifr.res.in.

PII S0167-8655(96)00047-5

major advantage for real-time applications.

The primary objective in the feature selection problem is to select a set of features to form an N-dimensional feature space that yields the best classification in the sense of being as close to Bayes decision as possible. For a given set of patterns belonging to a particular class, a maximum likelihood decision criterion gives the best classification score, assuming model correctness. Any measure that shows a reasonably good correlation with this classification score is suitable for ranking and pruning the feature set.

Fisher proposed a criterion for ranking a set of features in terms of their discriminative power in the case of a two-way classification problem (Fisher, 1936). However, he assumes that each feature has a uni-modal, symmetric distribution for each class. In this paper, we generalise the Fisher criterion to cases where the distribution is multi-modal or skewed and can be modeled as a linear combination of Gaussian distributions. We consider a set of one-dimensional independent components as a feature set. The number of such independent components that constitutes this set is the dimension of the feature space. The validity and usefulness of the proposed definition is tested by a Monte Carlo experiment.

The Fisher ratio as an objective measure for feature selection is introduced in Section 2. The proposed definition of the Fisher ratio as applicable to a Gaussian mixture distribution is also included in the same section. The details of the Monte-Carlo experiment conducted for testing the validity and usefulness of the proposed definition are given in Section 4. Experimental results are discussed in Section 5, and the conclusions of this study are stated in Section 6.

2. Fisher criterion for feature selection

In order to prune a feature set, it is essential to define a suitable objective measure. This measure should reflect the ability of a feature to provide (i) large inter-class separation, and (ii) small intra-class spread between patterns. One such measure, defined in statistical literature, is the pairwise Fisher ratio between any two classes (Duda and Hart, 1973). It is defined as

$$F^{\mathrm{u}}_{ijk} = \frac{(\mu_{ik} - \mu_{jk})^2}{(\sigma^2_{ik} + \sigma^2_{jk})}. \tag{1}$$

Here, μ_{ik} and μ_{jk} denote the cluster means for the kth vector component of classes i and j, respectively; σ^2_{ik} and σ^2_{jk} are the corresponding cluster variances. It is easy to see that this measure is maximum when the inter-class separation is maximized and the intra-class spread is minimized. Also, the measure is independent of scaling. This facilitates a fair comparison of the features using this measure.

The definition of the Fisher ratio can be extended to multi-class problems in many ways (Duda and Hart, 1973). It has been defined as the ratio of between class scatter and within class scatter. We have used the arithmetic average of the pairwise Fisher ratios corresponding to all distinct pairs of classes:

$$F^{\mathrm{u}}_{k} = \frac{1}{J(J-1)} \frac{\sum_{i=1}^{J}\sum_{j=1}^{J} P_i P_j F^{\mathrm{u}}_{ijk}}{\sum_{i=1}^{J}\sum_{j=1}^{J} P_i P_j}, \quad i \neq j. \tag{2}$$

Here, J denotes the number of classes; P_i and P_j are the a priori probabilities of the classes i and j, respectively. If the training data set is a fair representation of the test set, it is reasonable to expect the a priori probability of a class to be proportional to the number of data points belonging to that class in the training set. As in the case of the two-class problem, the best N features (in descending order of the Fisher ratios) are selected for the classification task. This criterion has been widely used to rank the features for various classification tasks (Pruzansky, 1964; Das and Mohn, 1971; Krishnan and Samudravijaya, 1993; Paliwal, 1992; Phillips, 1988). Here, the distribution of each feature for each class is assumed to be Gaussian. In addition, the features are implicitly assumed to be uncorrelated.

3. Generalization to Gaussian mixture distributions

The utility of the existing definition of the Fisher ratio is, however, limited to situations where the assumption of a uni-modal symmetric distribution holds. This assumption is not strictly valid in many real-life situations. There are instances where the distribution is multi-modal or skewed. Such distributions can be reasonably approximated as linear combinations of Gaussians. A powerful feature of a Gaussian mixture model is its ability to provide smooth approximations to arbitrarily-shaped densities (Hartman et al., 1990; McLachlan and Basford, 1987). Moreover, a linear combination of diagonal covariance Gaussians can be used to model correlations between features (Reynolds and Rose, 1995).

The Gaussian mixture model has been employed for modeling heterogeneous data in diverse fields. Some examples are genetic counselling (Basford and McLachlan, 1985), describing teaching behaviour (Aitkin et al., 1981) and modeling speech sounds for speech recognition (Juang and Rabiner, 1985; Krishnan, 1994; Lee, 1988) and speaker identification (Reynolds and Rose, 1995). Now we will formulate the generalization of the Fisher ratio to Gaussian mixture models.

Let x denote a K-dimensional random vector. Under the finite mixture model, each observation, x_j can be viewed as arising from a superpopulation G which is a mixture of a finite number, say L, of populations $G_1, \ldots, G_L$ in some mixing proportions $\pi_1, \ldots \pi_L$, respectively, where

$$\sum_{l=1}^{L} \pi_l = 1 \quad \text{and} \quad \pi_l \geqslant 0 \quad (l = 1, \ldots, L). \tag{3}$$

The probability density function of an observation x in G can be represented in the finite mixture form

$$p(x \mid \lambda) = \sum_{l=1}^{L} \pi_l b_l(x), \tag{4}$$

where $b_l(x)$ is a K-variate Gaussian function corresponding to G_i and is of the form

$$b_l(x) = \frac{1}{(2\pi)^{K/2}|\Sigma_l|^{1/2}} \times \exp\{-\tfrac{1}{2}(x - \boldsymbol{\mu}_l)^{\mathrm{T}} \Sigma_l^{-1} (x - \boldsymbol{\mu}_l)\}, \tag{5}$$

where $\boldsymbol{\mu}_l$ and Σ_l are the mean vector and covariance matrix of the lth component Gaussian. The union of the parameters of the individual distributions characterizes the superpopulation and is denoted by

$$\lambda = \{\pi_l, \boldsymbol{\mu}_l, \Sigma_l\}. \tag{6}$$

Let us now consider two classes C_L and C_N having L and N mixture components, respectively. Let S denote the set of pairs of mixture components (l, n) where $1 \leqslant l \leqslant L$ and $1 \leqslant n \leqslant N$. It is possible to define conventional Fisher ratios (Eq. (1)) for each of these pairs. We now define the generalized Fisher ratio between the distributions for C_L and C_N as follows:

$$F_{ijk}^{\mathrm{m}} = \frac{\sum_{l=1}^{L} \sum_{n=1}^{N} \pi_{il} \pi_{jn} F_{lnk}^{\mathrm{u}}}{\sum_{l=1}^{L} \sum_{n=1}^{N} \pi_{il} \pi_{jn}}. \tag{7}$$

Here π_{il} is the mixing proportion of the lth population corresponding to ith class, π_{jn} is the mixing proportion of the nth population for jth class, F_{lnk}^{u} is the unimodal Fisher ratio and is computed using Eq. (1). It is easy to see from Eq. (3) that the denominator sums up to 1.

The rationale behind the above definition is as follows. The product $\pi_{il}\pi_{jn}$ can be thought of as the probability of the occurrence of the pair (l, n) in S. Multiplying the conventional Fisher ratio by this product, ensures that those pairs of mixtures that have higher probability of occurrence will be given more weight. In fact, the above definition is nothing but the expected value of the conventional Fisher ratio, the expectation being taken over the set S. Also, in the limiting case, when $L = N = 1$ the generalized Fisher ratio reduces to the conventional Fisher ratio Eq. (1).

The generalized F-ratio between the distributions for a two-class problem (Eq. (7)) can be extended to a multi-class problem as

$$F_k^{\mathrm{m}} = \frac{1}{J(J-1)} \frac{\sum_{i=1}^{J} \sum_{j=1}^{J} P_i P_j F_{ijk}^{\mathrm{m}}}{\sum_{i=1}^{J} \sum_{j=1}^{J} P_i P_j}, \quad i \neq j, \tag{8}$$

where the notations have the same meaning as in Eq. (2).

4. Simulation experiment

4.1. Data generation

The proposed objective criterion can be said to be valid if the correlation between F_k^{m} (Eq. (8)) and the classification performance is positive. The criterion would be useful if the classification performance correlates better with F_k^{m} than with F_k^{u} (Eq. (2)). Both these issues have been studied through Monte Carlo simulation experiments.

A uniform deviate was used to initialize the parameter set λ (Eq. (6)). The means of Gaussians were initialized by the random numbers after appropriate shifting and scaling. The mixing proportions, π_{il}, were initialized by random numbers and normalized to satisfy Eq. (3). The covariance matrix of each of the component Gaussians was assumed to be diagonal and the variance of each normal deviate was set to unity.

For each class, 10000 feature vectors were generated using Eqs. (4) and (5). Since each class has the same number of feature vectors, the a priori probabilities of all the classes are assumed to be equal. Care was taken to choose a random number generator, by using the Bays-Durham shuffling procedure (Press et al., 1992), so that correlations among features were minimum. During the generation of feature vectors, their class labels were noted.

4.2. Parameter estimation

Parameter estimation was based on an iterative procedure using the K-means clustering algorithm. Each iteration of this algorithm consists in binning all data points to their nearest-neighbour partition based on an appropriately chosen distance metric (with respect to the partition mean), followed by the reestimation of the means from the binned data. Typically the Euclidean distance metric suffices for running the K-mean algorithm (Lee, 1988). Convergence is based on the total distortion in all the partitions. The nearest neighbour rule is used to label the Gaussian mode of a data vector $\boldsymbol{x}$ belonging to a given class as

$$l^{\text{opt}} = \text{ArgMin}_l \, \|(\boldsymbol{x} - \boldsymbol{\mu}_l)\|, \tag{9}$$

where l^{opt} is the nearest cluster index and $\boldsymbol{\mu}_l$ is the cluster centre.

Below, we define some notations which are used in the estimation procedure.

l = cluster index,
i = class index,
$\boldsymbol{\mu}$ = cluster centre (multivariate Gaussian mean),
$\boldsymbol{x}$ = data vector,
σ = cluster variance (multivariate Gaussian variance),
π = mixture strength of Gaussian,
N_{il} = number of data points in lth cluster of ith class,
N_i = number of data points in ith class.

The reestimation formulae for Gaussian mixture parameters (mean, variance and mixture strength) of partitioned data are given by

$$\boldsymbol{\mu}_{il} = \frac{1}{N_{il}} \sum_{n \in l} \boldsymbol{x}_{in}, \tag{10}$$

$$\sigma_{il}^2 = \frac{1}{N_{il}} \sum_{n \in l} (\boldsymbol{x}_{in} - \boldsymbol{\mu}_{il})(\boldsymbol{x}_{in} - \boldsymbol{\mu}_{il})^{\text{T}}, \tag{11}$$

$$\pi_{il} = \frac{N_{il}}{N_i}. \tag{12}$$

In the estimation algorithm, the data is binned into the clusters using the nearest-neighbour criterion and an Euclidean distance measure. The convergence criterion used a maximum of 100 iterations and the convergence threshold was fractional distortion (based on the difference in distortions between successive iterations) of 5%. The parameters (mean and variance of each cluster, as well as the mixing proportions for each of the Gaussian mixtures) set for a given class of data were computed from data belonging to that class by using Eqs. (9)-(12). No tying of the parameters was made across the classes.

4.3. Maximum likelihood based classification

The discriminative power of a feature component k was estimated as follows. Class labels were assigned to all datapoints based on the maximum likelihood criterion under the assumption that the superpopulation consists of a mixture of univariate Gaussians. A datapoint, x_k, was assigned the class i iff

$$p_i(x_k \mid \lambda_i) > p_j(x_k \mid \lambda_j) \quad \forall j. \tag{13}$$

Here, λ_i is the parameter set (Eq. (6)) corresponding to the class i. $p_i(x_k \mid \lambda_i)$ is computed using Eq. (4). The class labels of data, thus assigned, were compared with their true class affiliations to arrive at the classification scores, C_k, of the feature k. Here, the classification score is defined as the ratio of correctly classified patterns to the total number of patterns.

4.4. Correlation computation

Let $\{F^{\text{u}}\}$ denote the series $F_1^{\text{u}}, \ldots, F_K^{\text{u}}$ of unimodal Fisher ratios where K is the dimension of the feature vector. Likewise, $\{F^{\text{m}}\}$ represents the corresponding series for the Gaussian mixture case. Also, let $\{C\}$ denote the series $C_1, \ldots, C_K$ of classification scores for various features. Let r^{u} and r^{m} denote the linear corre-

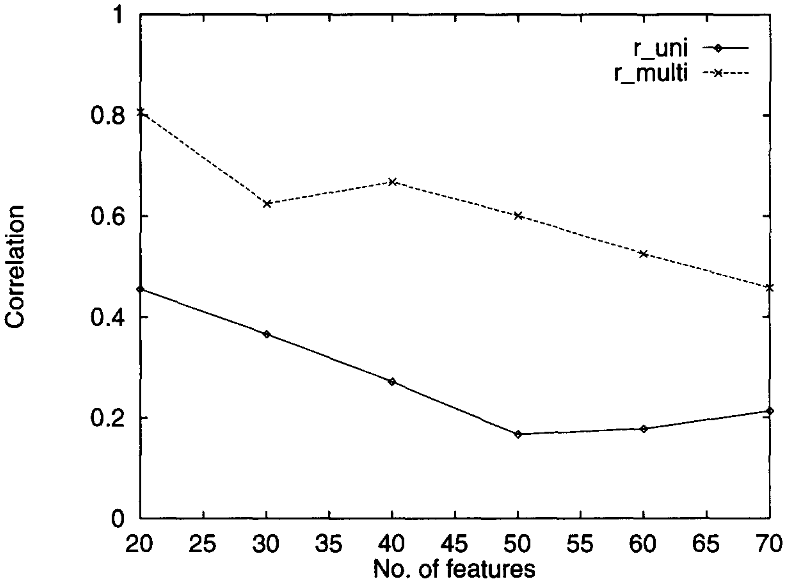

Fig. 1. Linear correlation coefficients between classification results and F-ratios are plotted against the number of features.

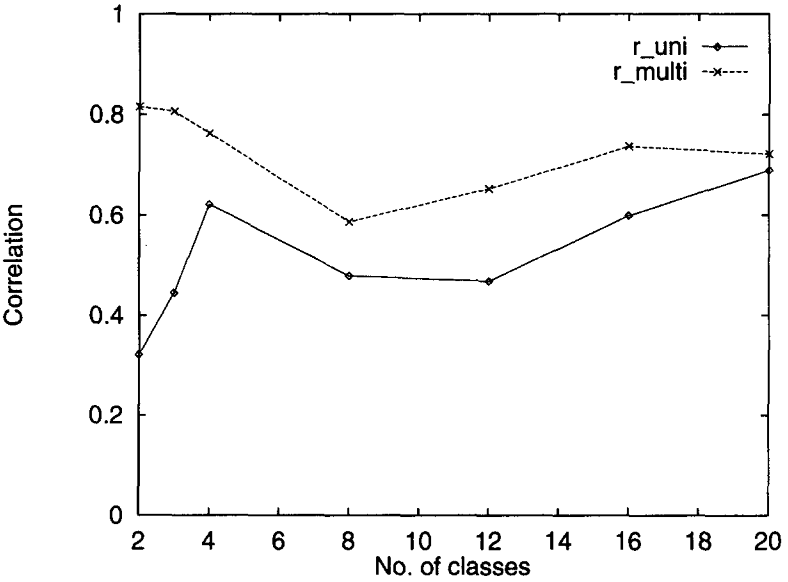

Fig. 2. Linear correlation coefficients between classification results and F-ratios are plotted against the number of classes.

lation coefficients of the classification performance series $\{C\}$ with $\{F^{\mathrm{u}}\}$ and $\{F^{\mathrm{m}}\}$, respectively. For pairs of quantities (x_i, y_i), $i = 1, \ldots, N$, the linear correlation coefficient r is given by the formula

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}, \qquad (14)$$

where $\bar{x}$ is the mean of the x_i's, $\bar{y}$ is the mean of the y_i's.

The experiment was conducted with 100 random realizations of the parameter set λ, Eq. (6). The average linear correlation coefficients for unimodal ($\overline{r^{\mathrm{u}}}$) and mixture ($\overline{r^{\mathrm{m}}}$) cases were computed for a range of features, classes and mixtures.

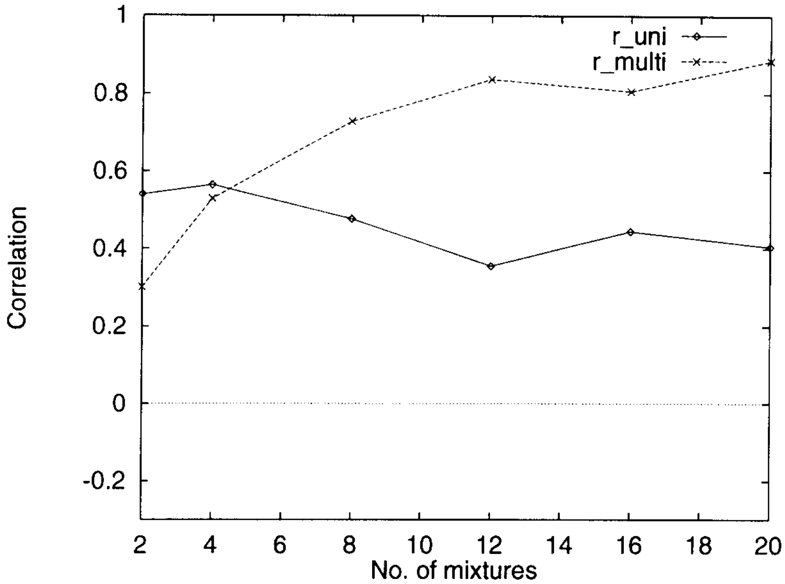

Fig. 3. Linear correlation coefficients between classification results and F-ratios are plotted against the number of component Gaussians.

5. Results and discussion

Positive correlation was observed between the classification performance C_k and the proposed discrimination criterion, F_k^m, for all the features and for all realizations of the parameter set λ (refer to Eq. (6)). The observed correlation values ranged between 0.5 and 0.8. This observation validates the generalization of the Fisher ratio (Eq. (8)) proposed in this work.

The average linear correlation coefficients of the classification performance with the conventional multi-class Fisher ratio and the generalized criterion are plotted in Fig. 1 against the number of features. Here, the number of classes is 3 and the number of component densities is 16. It is clear from the figure that the proposed definition has stronger correlation with the classification results, than the conventional definition. The difference between the two correlation coefficients is positive for a wide range of dimensions of the feature vector.

The variation of the correlation coefficients with the number of classes is shown in Fig. 2. Here, the number of features is 20 and the number of mixture components is 16. The performance of the proposed criterion decreases with increasing number of classes. This is due to the increase in number of parameters that are to be estimated from the given (limited) set of data points. It can also be noted from Fig. 3 that correlation with classification scores increases with increase in the number of component densities for the proposed criterion while it decreases for the conventional definition.

The proposed measure, being an extension of the unimodal F-ratio measure (i.e., conventional), is also constrained by some of the common limitations that exist in using the conventional measure for a feature selection problem. As this method uses a measure based on the separation between means, it is not suitable for data that is poorly separated in means.

Moreover, the method of component-wise evaluation of a feature may not sum up to the best set of elements of the feature set (or space). In a multidimensional case of feature evaluation, it is usual to extend the variance (used in the definition of F-ratio in Eq. (1)) to the trace or determinant of the covariance matrices. It would be interesting to study the performance of the proposed measure in comparison with the measures based on covariance matrices, such as $J_1 = \mathrm{tr}(\Sigma_B \Sigma_W^{-1})$, $J_2 = \det(\Sigma_T)/\det(\Sigma_W)$, etc., where subscripts B, W and T refer to between class, within class and total scatters, respectively (Duda and Hart, 1973). Also, a comparative study of the proposed measure with information-theoretic measures, such as Kullback-Leibler divergence, Bhattacharya distance, etc., would be interesting.

6. Summary and conclusions

The use of the Fisher ratio for feature selection in a pattern classification task has been generalized to cover Gaussian mixture models. The proposed discrimination criterion is based on the mixing proportions of the component densities. The validity and usefulness of the proposed definition were tested using a Monte Carlo simulation. The correlation between the classification results and the proposed objective criterion has been found to be better than the correlation obtained with the conventional uni-modal measure. So, the proposed measure is expected to be suitable for feature selection spanning a wider range of underlying distributions.

Acknowledgements

We thank the referees for their valuable comments which led to a redesign of the simulation work, for confirming the suitability of the proposed method in practical applications, and for suggestions which helped to improve our presentation.

References

Aitkin, M., D. Anderson and J. Hinde (1981). Statistical modelling of data on teaching styles (with discussion). *J. Roy. Statist. Soc. A* 144, 419-461.

Basford, K. and G. McLachlan (1985). Likelihood estimation with normal mixture models. *Appl. Statist.* 34, 282-289.

Das, S. and W. Mohn (1971). A scheme for speech processing in automatic speaker verification. *IEEE Trans. Audio Electroacoust.* 19, 32-43.

Duda, R. and P. Hart (1973). *Pattern Classification and Scene Analysis.* Wiley, New York.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics* 7 (Part II), 179-188.

Hartman, E., J. Keeler and J. Kowalski (1990). Layered neural networks with gaussian hidden units as universal approximations. *Neural Computation* 2, 210-215.

Juang, B.H. and L.R. Rabiner (1985). Mixture autoregressive hidden Markov models for speech signals. *IEEE Trans. Acoust. Speech Signal Process.* 33 (6), 1404-1413.

Krishnan, S. (1994). Speech Recognition by Computer: Spectral Temporal Redundancy and Stochastic Segmental Models. PhD thesis, University of Bombay.

Krishnan, S. and K. Samudravijaya (1993). Optimal feature selection for automatic speech recognition using statistical methods. Technical Report TIFR-CSC-2-1993, C.S.C. Group, Tata Institute of Fundamental Research, Bombay, India.

Lee, K. (1988). Large vocabulary speaker-independent continuous speech recognition: The SPHINX system. Technical Report CMU-CS-88-148, Computer Science Dept., Carnegie-Mellon University, Pittsburgh, PA.

McLachlan, G. and K. Basford (1987). *Mixture Models: Inference and Applications to Clustering,* Series on Statistics: Textbooks and Monographs. Marcel Dekker, New York, Chapter 2, 37-70.

Paliwal, K. (1992). Dimensionality reduction of the enhanced feature set of the HMM-based speech recognizer. *Digital Signal Processing* 2, 155-171.

Phillips, M. (1988). Automatic discovery of acoustic measurements for phonetic classification. *J. Acoust. Soc. Amer.* 84 (S216).

Press, W., A. Teukolsky, W. Vellerling and B. Flannery (1992). *NUMERICAL RECIPES in C, The Art of Scientific Computing.* Cambridge Univ. Press, Cambridge, Chapter 7, 280.

Pruzansky, S. (1964). Talker-recognition proedure based on analysis of variance. *J. Acoust. Soc. Amer.* 36, 2041-2047.

Reynolds, D. and R. Rose (1995). Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.* 3, 72-83.