Principal Component Analysis for HMM-Based Pathological Voice Detection

M. Sarria-Paja, G. Castellanos-Dominguez, N. Gaviria-Gómez

1Intelligent Machines and Pattern Recognition Group, Instituto Tecnológico Metropolitano, Medellín-Colombia
2Control and Digital Signal Processing Group, Universidad National de Colombia, Manizales-Colombia.
3Facultad de Ingeniería, Universidad de Antioquia, Medellín-Colombia.

Abstract: This paper presents a methodology for feature selection in dynamic problems based on the analysis of the variation of linear components in acoustic features combined with an estimation of the ratio between a compactness measure to the separation measure. The methodology is applied to the automatic detection of voice disorders by means of stochastic dynamic models; results showed a significant reduction in the number of features, 96.6% of accuracy, and a 62.2% of computational cost reduction.

Keywords: Dynamic features, HMM, PCA, feature selection, pathological voice, clustering.

I. INTRODUCTION

During phonation of sustained vowels, the normal voice is a regular and periodic signal; however changes in its waveform can be appreciated if some disorders arise. Moreover, the classical distortion measures based on fluctuations of acoustic measures may be complemented with dynamic features obtained from its contours, as pointed out by other studies [1]. One of the key properties that make dynamic features useful is that they consider changes in the temporal structure of the excitation signal. Short term features combined with dynamic classifiers (e.g. Hidden Markov Models - HMM), have been used in the classification of pathological voices [2]. But, it is no clear whether gathering of dynamic features should lead to an improved representation capability, and hence to higher performance of the dynamic classifier. Namely, in voice recognition where the training data are labeled, a projection is often required to emphasize the discrimination between the clusters. Therefore, a more detailed study should be conducted to assess the relevance of dynamic features that describe pathologies, which could be used in data analysis and evaluation to support diagnose by automatic dynamic classifiers.

Performance in training of pattern recognition systems to detect pathologies can be increased, if proper feature extraction is done. Training procedures usually deal with a high number of features, nevertheless a high dimension input space means significant processing time, higher cost of the collected biosignal records since more observations are needed, and the well known curse of dimensionality phenomena [3].

The aim of this paper is to assessment an approach combining PCA with HMM for pathological voice diagnosis based on a concrete cluster validity measure. For this purpose, the feature selection methodology presented in [2], is adopted, and by incorporating both measures of cluster separability and cluster compactness, it is showed that one can provide analysis of clustering scatter for groups with varying populations. The method is tested on the voice disorders database developed by The Massachusetts Eye and Ear Infirmary Voice Laboratory (MEEIVL).

This paper makes a contribution to the effort to make an automatic discrimination between pathological and normal voice.

II. METHODS

A. Hidden Markov Models

Hidden Markov models are double-layer stochastic processes, composed of a hidden layer that controls the time evolution of spectral characteristics of an observable layer. A hidden Markov model has N distinct states and each state is uniquely defined by an observation (or output) probability density, usually a mixture Gaussian density (continuous case) or a discrete density (Discrete case), that provides a likelihood for a given vector having been generated by the state. The transition from a state is governed by the state transition probabilities and influenced by the current observation vector. The state observation and transition probabilities provide a probabilistic mechanism for association of a time sequence of vectors with a given HMM model [4].

B. Principal Component Analysis in dynamic features

Widely known approaches, like PCA [5,6,7] and sequential search methods, have been customized as feature selection methods for the use with a HMM classifier. Assuming that the input contour data are highly correlated, linear transformation methods such as PCA tries to exploit the correlation present in the data by projecting the data onto a new space where the axes are orthogonal to each other.

Let \( \xi_j[k], k=1,\ldots,m \) be the \( j \)-th dynamic feature belonging to \( i \)-th observation, where \( j=1,\ldots,p \), \( i=1,\ldots,n \); being \( n \) the number of observations and \( p \) the number of features, which change over time \( k \). Each
vector observation \( \xi_i \) can be represented by a supervector of size \( mp \times 1 \):

\[
\xi_i = [\xi_i[1], \xi_i[2], \ldots, \xi_i[m], \xi_i[1], \xi_i[2], \ldots, \xi_i[m]]^T
\]

The respective covariance matrix, after centering each one of the observation supervectors is computed as:

\[
S = \frac{1}{n} \sum_{i=1}^{n} \xi_i \xi_i^T = \frac{1}{n} GG^T
\]  

(1)

Where \( G \) stands for matrix \( G = [\xi_0^0, \xi_0^1, \ldots, \xi_0^n] \). In most cases, we are far away from computing the eigenvectors \( \mathbf{v} \) and eigenvalues \( \lambda \) of such a huge matrix. Nevertheless, the rank properties of \( G \) can be used, in special, the one that state that \( GG^T \) has the same non-null eigenvalues than \( G^T G \) and the advantage of \( n \ll pm \), as given in [8]:

\[
G^T G \hat{v}_i = \lambda \hat{v}_i
\]  

(2)

being \( \hat{v}_i \) the eigenvectors of \( G^T G \), so that, \( \mathbf{v}_j = G \hat{v}_j \). Therefore, the eigenvectors corresponding to non-zero eigenvalues of \( S \) are \( \mathbf{v}_j = G \hat{v}_j \). The eigenvalues associated with the \( r \) largest eigenvalues of \( S \) are selected as Principal Directions [9], which span an orthonormal basis for a subspace containing most of the information given by observations. Trying to reproduce the observation in the original space as a linear combination of the \( r \) principal directions,

\[
\hat{\xi}_i^0 = \sum_{k=1}^{r} w_k \mathbf{v}_k^T
\]  

(3)

so, from (3) the reconstruction weights \( w_k = \mathbf{v}_k^T \xi_i^0 \) can be thought as the new set of features, and taking advantage of the orthonormality property of the basis, observations can be recognized using geometric criteria to partition the subspace off.

On the other hand, this method allows identify and choose those dynamic features that influence the most. The magnitudes of the entries of the eigenvectors that span the representation basis, tell us the variables to be choose. Let \( \mathbf{p} \) be the vector expressed as;

\[
\mathbf{p} = \sum_{i=1}^{r} \lambda_i |\mathbf{v}_i|
\]

so, that its larger values are the most significant windows from the dynamic features, this sum of absolute values is an approximation due to the equivalence of norms in finite subspaces (\( L^2 \) and \( L^1 \)). Rearranging \( \mathbf{p} \) in the following manner:

\[
\mathbf{p} = [\rho_1, \rho_2, \ldots, \rho_m, \rho_1, \rho_2, \ldots, \rho_m]^T
\]

\[
\Rightarrow \mathbf{P} = \begin{bmatrix}
\rho_1 & \rho_2 & \cdots & \rho_m \\
\rho_1 & \rho_2 & \cdots & \rho_m \\
\vdots & \vdots & \ddots & \vdots \\
\rho_1 & \rho_2 & \cdots & \rho_m
\end{bmatrix}
\]  

(4)

it is possible to obtain the scalar \( \hat{\rho}_j = \sum_{i=1}^{n} \rho_{jk} \), \( j = 1, \ldots, p \) which is the sum of the elements of each column \( j \) from \( \mathbf{P} \) matrix. In consequence, the main assumption is that the largest values of \( \hat{\rho}_j \) point out to the best input attributes since they exhibit higher overall correlations with principal components.

C. Definition of a clustering validity measure

Th Given a set of \( n \) observations in a \( p \)-dimensional input training space, \( \mathbf{x} \in \{\mathbf{x}_i \in \mathbb{R}^p : i = 1, \ldots, n\} \), the main goal of a partitioned clustering is to determine an assignment \( \mathbf{b}_i = \{b_i \in \{0,1\} : j = 1, \ldots, K\}, i = 1, \ldots, n \), such that a given cost function is minimized, where \( b_i = 1 \) if observation \( \mathbf{x}_i \) is assigned to the \( j \)-th partition, and \( b_i = 0 \) otherwise; \( K \) is the number of clusters (in this case, specified by the user). Mainly, the cost function is defined \( C(\mathbf{b}; K) \) as a weighted average, i.e.,

\[
C(\mathbf{b}; K) = \sum_{i=1}^{n} \sum_{j=1}^{K} b_{ij} d(\mathbf{x}_i, \mathbf{m}_j)
\]

where \( \mathbf{m}_j = \sum_{i=1}^{n} b_{ij} \mathbf{x}_i / \sum_{i=1}^{n} b_{ij} \) (named the center of the \( j \)-th cluster). Notation \( d(\mathbf{x}_i, \mathbf{x}_j) \) stands for a distance metric between two observation vectors \( \mathbf{x}_i \) and \( \mathbf{x}_j \). Because it is a reliable metric for early stages of training, the most commonly used distance is the Euclidean metric,

\[
d_e(\mathbf{x}_i, \mathbf{m}_j) = (\mathbf{x}_i - \mathbf{m}_j)^T (\mathbf{x}_i - \mathbf{m}_j)
\]

Although other distance measures, such as Mahalanobis, can also be used in the clustering criterion to take care of hyperellipsoidal-shaped clusters, and which is defined as

\[
d_m(\mathbf{x}_i, \mathbf{m}_j) = (\mathbf{x}_i - \mathbf{m}_j)^T \sum_{j=1}^{K} (\mathbf{x}_i - \mathbf{m}_j)
\]

where \( \sum_{j=1}^{K} \) is the inverse of the \( p \times p \) covariance matrix of the observation set belonging to the \( j \)-th cluster.

The intraclass distance, denoted as \( \delta_i \), is defined as overall statistical distance between the data points inside clusters

\[
\forall j, k = 1, \ldots, n, i = 1, \ldots, K
\]

The overall clustering compactness measure is defined as
The interclass distance, denoted as \( \delta = \max_{i,j} \{ d(b_i, b_j) \} \) (4)

The interclass distance, denoted as \( d(b_i, b_j) \), is the distance between the elements in cluster \( b_i \), and those in cluster \( b_j \):

\[
d(b_i, b_j) = \min_{x_i, x_j \in b_i, x_i, x_j \in b_j} \{ d(x_i, x_j) \}.
\]

Then, overall clustering separation measure is defined as

\[
\Delta = \min_{i,j} \{ d(b_i, b_j) \},
\]

(5)

Assuming \( d(b_i, b_j) \) as Euclidean metric, value \( \Delta \) represents the minimum Euclidean distance between clusters. Inspired by previous definitions, the following goodness-of-clustering measure \( J \) relating the cluster separability and cluster compactness can be used. Namely, the validity measure is defined as the ratio of the compactness measure \( \delta \) to the separation measure \( \Delta \), i.e.,

\[
J = \Delta / \delta
\]

(6)

It is expected that clusters should be as dense as possible and the distance between clusters should be as large as possible. Therefore, the compactness measure \( \delta \) is expected to be small and the separation measure \( \Delta \) to be large. A bigger \( J \) means a more compact and separate cluster configuration and, hence, a better validity measure for the clustering.

C.1 Estimate of clustering validity measure

The sample estimate of the covariance matrix for the \( i \)-th class is given by

\[
\hat{S}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_j - m_i)(x_j - m_i)^T
\]

where \( x_j \) denotes the \( j \)-th recording, \( m_i \) the mean estimate, and \( n_i \) the number of patterns belonging to the \( i \)-th class.

The pooled estimate from all classes of the within matrix can be obtained as,

\[
\hat{S}_w = \frac{1}{C} \sum_{i=1}^{C} \hat{S}_i
\]

being \( n \) the size of the whole sample, and \( C \) the number of classes. The between covariance matrix is computed as

\[
\hat{S}_b = \frac{1}{C} \sum_{j=1}^{C} \sum_{i=1}^{n_j} (m_j - m)(m_j - m)^T
\]

with \( m \) as the overall mean of the entire sample. Finally, the separability measure is derived from the following expression:

\[
J = \frac{\hat{S}_w + \hat{S}_b}{\hat{S}_w}
\]

(7)

It is important to note that this criteria is similar to the multivariate fisher score, which in the two class case is given by the largest eigenvalue of \( \hat{S}_w \hat{S}_b \).

III. EXPERIMENTAL SETUP

A. Database

The Massachusetts Eye and Ear Infirmary Voice Laboratory (MEEIVL). 48 features were computed. These features correspond to 16 measures and its first and second derivatives. These measures are: 12 Mel Frequency Cepstrum Coefficients (MFCC) the Harmonics to Noise Ratio (HNR), the Glottal to Noise Excitation Ratio (GNE), the Normalized Noise Energy (NNE), and the Energy of the frame. A total of 48 features were taken in account [2].

B. Feature selection strategy

For each observation are taken \( j \) \( (j = 1, 2,..., 48) \) dynamic features. These features were selected using the relevance measure presented in section II, then we use the Eq. (7) to estimate the cluster separability, this process generates a curve and each point in the curve is obtained by a incremental representations; that is, a larger set is obtained by adding features to the preceding one.

For feature selection we use the first \( k \) features, such that \( J_k \leq \varepsilon \cdot \max(J) \), where \( 1 \leq k \leq 48 \) and \( 0 < \varepsilon \leq 1 \). In this case \( \varepsilon = 0.99 \).

The accuracy was measured using a \( k \)-folds cross validation strategy. In particular, 11 folds have been used, splitting the 70% of the files for training the classifier, and the remaining 30% for validating. These sets were randomly chosen.

III. RESULTS

As a result of the relevance analysis carried out above, a set of weights for the features was obtained. Fig. 1 shows the weights for each one of the 48 features.

The estimated values for relevance weight of dynamic features shown in Fig 1 are the starting point for selection and subsequent reduction of features; this analysis showed that the most significant features are the instantaneous measures, without the first and second derivatives, since the most weighted are the first 16.
1) reflects on increasing separation between clusters. Nonetheless, this behavior holds up to certain number of features (the most relevant features).

Fig. 2 Cluster separability vs. number of features.

The accuracy results are calculated employing continuous HMMs and the complete set of features. Several values of mixtures (NG=2, NG=3, and NG=4) and states (S=2, S=5, and S=10) were tested. The best results were obtained with NS=2, and NG=3. The results obtained are shown in the Table 1.

<table>
<thead>
<tr>
<th>NG</th>
<th>Number of states (NS)</th>
<th>Accuracy</th>
<th>Accuracy</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>94.1±1.1</td>
<td>94.2±2.6</td>
<td>84.7±3.3</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>94.6±1.8</td>
<td>91.1±1.8</td>
<td>82.3±2.3</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>91.5±1.8</td>
<td>90.4±2.8</td>
<td>81.0±3.0</td>
</tr>
</tbody>
</table>

Table 1. Accuracy results using continuous HMMs

The accuracy results are recalculated employing NS=2, NG=3 and applying the feature selection strategy. The proposed methodology for reducing the number of dynamic features in the identification of pathological voices proved to be useful for the experiments carried out. As a result was obtained an adequate performance while employing a considerably reduced feature set. The presented way of training shows that for the automatic detection of pathological voices is better to use a good set of features than a complex stochastic dynamic training model, because the later may have lower generalization capabilities.

ACKNOWLEDGEMENTS
This work was carried out under grants: Tecnologías de Información y Comunicaciones para la investigación y enseñanza de astronomía en Colombia funded by Instituto tecnológico Metropolitano, Institución Universitaria- Medellín Colombia.

REFERENCES