PRINCIPAL AND DISCRIMINANT COMPONENT ANALYSIS FOR FEATURE SELECTION IN ISOLATED WORD RECOGNITION

E. Lleida, C. Nadeu

Dpto. of Signal Theory and Communications, ETSIT-UPC.
P.O. Box 30.002, 08080 Barcelona, Spain

ABSTRACT

In this paper we propose the use of the Principal Component Analysis and Discriminant Analysis as feature selection step in a word recognition system. The classic pattern-matching approach used in isolated word recognition assumes that the adjacent feature vectors are uncorrelated and that the variability of speech can be accounted for the same distance measure for all words. However, these assumptions are not true, and a feature selection process is needed to deal with these problems. This work proposes the use of two steps as feature selection process. One is related with the best representation for temporal selection and the other one is related with the discriminant properties for frequency selection.

I. INTRODUCTION

The first step in any speech recognition system is the signal feature measurement. Typically, the speech signal is modeled by a sequence of feature vectors called "Template" in the IWR environment. Generally, feature measurement methods are block processing models giving \( N \) vectors of \( P \) features. In this work, the LPC technique has been chosen as a feature measurement method.

The speech signal has stationary parts which are represented by several feature vectors, having a great redundancy [1,2]. Therefore, we can look for a new model in which the correlation among feature vectors is removed. For this purpose we assume that there is an underlying set of "real" uncorrelated features, and the features we are working on are "impure" in the sense that is a linear combination of those "real" features. Then, the objective is to find a transformation which recovers the "real" features [3]. Basically, the problem is to represent the sequence of spectra by a superposition of the members of any orthogonal family of functions where the input template is represented with less coefficients. If \( y(n) \) is the \( n \)th LPC vector, the transformation obeys the following formulation

\[
y(n) = \sum_{m=1}^{M} \alpha_m \phi_m(n)
\]  

(1)

where \( \phi_m \) is the \( m \)th transformation function and \( \alpha_m \) is the new \( m \)th feature vector.

This work was supported by the PRONTIC grant 105/88.

A typical family of functions which performs this transformation is obtained by means for the Principal Component Analysis (PCA). It minimizes the mean square error between a vector and its estimation by a linear combination. In this work, the Principal Component Analysis is used to remove the temporal correlation in order to obtain \( M \) uncorrelated vectors, being \( M \ll N \). Thus, a new template is obtained with \( M \) uncorrelated vectors which are arranged in variance, not being required time-alignment to compare two templates. The temporal information is retained in the transformation functions (\( \phi_m \)). These functions are found from a training set.

In order to increase the separability among words, a new transformation is proposed in the frequency dimension. In this case, taking the \( M \) vectors obtained in the temporal selection, a transformation matrix associated with each vector of a word is sought. Two approaches are proposed; one maximizes the projected variance with the constraint that these functions must be orthogonal to a set of given functions which represent to the others words or a set of words of the vocabulary. The purpose of this step is to obtain a new set of functions which projects a word over a new axis which are orthogonal to the axis of maximum variance of the other words. The second approach maximizes the distance among this vector and the corresponding vectors of the others words. Thus, a new vector \( \delta_m \), called discriminant feature vector, will be obtained by transforming the vector \( \alpha_m \) as follows

\[
\delta_m(q) = \sum_{l=1}^{P} \alpha_m(i) \phi_m,q(l)
\]  

(2)

1\leq q \leq Q
where \(\phi_{m,q}\) is the \(q\)th discriminant function of the \(m\)th vector. After these processes, a template of \(M \times Q\) dimension is obtained where \(M < P\) and \(Q < P\), and the classification is done in this new space by comparing the discriminant vectors by means of the Euclidean distance and without time alignment.

In section 2 a description of the feature selection process is presented. Section 3 explains the training process and the test data base. The recognition experiments are reported in section 4.

II. FEATURE SELECTION

The feature selection process is performed in two steps called Temporal Selection and Frequency Selection.

TEMPORAL SELECTION

Temporal selection is the first step in our feature selection process. Its purpose is to obtain a time compression by removing the correlation of the temporal evolution of the spectrum. Given a \(N \times P\) matrix \(Y\) of spectral parameters \([y_i(n)]\) representing \(N\) frames of \(P\) features, a finite family of orthogonal functions can be found in accordance with (1) by means of the Principal Component Analysis (PCA), reducing a large set of correlated features into a smaller number of uncorrelated features.

If the covariance matrix of a template \(Y\) corresponding to a word \(w\) is defined as

\[
C_{yy}^w = \frac{1}{P} \sum_{i=1}^{P} (Y_i - \bar{Y})(Y_i - \bar{Y})^T
\]

(3)

where

\[
\bar{Y} = \frac{1}{P} \sum_{i=1}^{P} Y_i
\]

(4)

\[Y_i = \{y_i(1), y_i(2), \ldots, y_i(N)\}\]

then the orthogonal functions are obtained in the training step from the eigensystem

\[
C_{yy} \phi_m = \lambda_m \phi_m
\]

(5)

where \(C_{yy}\) can be equal to an average of the \(C_{yy}^w\) of each word or an average of all the covariance matrices of all vocabulary words. From this eigensystem, \(N\) eigenvalues and their corresponding eigenvectors are obtained. However, only the \(M\) eigenvector with the largest eigenvalues are retained. Thus, the transformation matrix is composed by the \(M\) eigenvectors with the \(M\) largest eigenvalues, ranking them from the largest to the smallest one. Then, the new coefficients \(\alpha_m\) have information about the interdependency among the feature vectors. It must be noticed that each orthogonal function is computed using the \(P\) features of each frame, thus, these functions carry information of the correlation of the \(P\) features. The first eigenvector represents the temporal trajectory of the spectrum with the largest variance, the second one represents the best temporal trajectory which can be obtained if the first eigenvector information is removed from the covariance matrix. As the eigenvalue decreases, the eigenvector associated carries information of the small variation of the temporal trajectory of the spectrum. The new feature vectors are obtained projecting the initial template with the transformation matrix resulting a sequence of uncorrelated feature vectors. This sequence is the best representation of the initial template in the mean square error sense with the least number of frames. The new feature vectors are ranked from the largest to the smallest variance so the new template needs no time-alignment to be compared with another template.

FREQUENCY SELECTION

The second step of the feature selection process is to compute a transformation matrix for each new feature vector obtained in the temporal selection in order to select the frequency parameters. In the previous step a representation criterion was used in order to obtain a subset of \(M\) uncorrelated vector which retain as much information as possible of the initial template. However, this transformation does not take into account the discriminant properties of the feature vectors. Thus, after the temporal selection, a frequency selection step is proposed to obtain a set of discriminant features.

Two methods are studied. The first one is the classical two classes linear discrimination analysis. In this step, a set of \(Q\) discriminant functions \(\phi_{m,q}\) is associated to each vector \(\alpha_m\) of a word which increases the separability of this vector from the \(M\) mths vectors of the other words. The new feature vector is obtained by means of eq. (2) in order to find the discriminant functions, two classes of vectors are defined. For a word \(w\), the \(m\)th feature vector of any utterance of it forms the correct class and the \(m\)th feature vector of the other words forms the incorrect class. Thus, the problem is to maximize the mean of the between-class distance minimizing at the same time the mean of the within-class distance.

Defining the within-class mean distance matrix as

\[
W = E((\alpha_c - \alpha_c^P)(\alpha_c - \alpha_c^P)^T)
\]

(6)

and the between-class mean distance matrix as

\[
B = E((\alpha_i - \alpha_c^P)(\alpha_i - \alpha_c^P)^T)
\]

(7)

where \(\alpha_c\) is a realization of the correct class, \(\alpha_c^P\) is the reference prototype of the correct class and \(\alpha_i\) is a realization of the incorrect class, the criterion function to be maximized is defined as [3,5]

\[
J = \text{tr}(F_CBF_C^T) - \lambda \text{tr}(F_CWF_C^T) - 1
\]

(8)

where \(F_C\) is the discriminant matrix of the correct class vector, \(F_C^T = [\phi_m,1,\phi_m,2,\ldots,\phi_m,Q]\).

The solution of this optimization problem is the eigensystem \((W^{-1}B)\phi_{m,k} = \lambda_k \phi_{m,k}\). Therefore, the discriminant matrix is formed by the \(Q\) eigenvectors with the \(Q\) largest eigenvalue of \(W^{-1}B\), whenever their eigenvalues were greater than 1. If an eigenvalue is smaller than 1 the within-class mean distance is greater
than the between-class mean distance. Thus, only those
eigenvectors whose eigenvalues are greater than 1 can be
used as discriminant functions. As in [4], this process
can be seen as a method for finding an specific-frame
distance, rotating the frequency dimension in order to
better characterize each uncorrelated feature vector of
each word.

The second method tries to maximize the projected
variance with the constraint that this projection must be
orthogonal to the others words or a set of them. So, it is a
problem of PCA

\[
\text{maximize } \phi_m^t M_{am} \phi_m^l \text{ with } \phi_m^t \phi_m = 1 \tag{9}
\]

where \( M_{am} \) is the covariance matrix of the m-th vector
with the constraint that these functions \( \phi_m \) must be
orthogonal to a set of given functions \( U = \{u_1, u_2, \ldots, u_N\} \), that is

\[
\phi_m^t u_j = 0 \quad 1 \leq j \leq N \tag{10}
\]

with this constraint, the function \( \phi_m \) is obtained by
means of the Lagrange multipliers maximizing the function

\[
J = \phi_m^t M_{am} \phi_m - \lambda (\phi_m^t \phi_m - 1) - \sum_{n=1}^{N} \mu_n \phi_m^t u_n
\]

The solution are the eigenvectors of the greatest
eigenvalues of the matrix

\[
(I - U^t (UU)^{-1} U) M_{am} \tag{11}
\]

Thus, the function \( \phi_m \) is the best function in the least
square error sense that is orthogonal to a set of given
functions. Therefore, the discriminant matrix is formed
by the Q eigenvectors with the Q largest eigenvalues.

### III. TRAINING PROCESS

**Test data base**

A data base consists of ten repetitions of the
Catalan digits \{u, dos, trcs, kuatra, sink, sis, set, vuit, nou, zer\} uttered by six male and three female speakers
(900 words) and recorded in a quiet room.

**Feature measurement**

The speech signal was sampled at 8 KHz, pre-
emphasized \((H(z)=1-0.95z^{-1})\) and 8 Log-Area ratios
were computed each 15 ms for the digit data base using the
LPC analysis of 30 ms of the speech signal. A typical
Hamming smoothing window was applied to the data. The
beginning and end of every utterance were automatically
detected by mean of an algorithm based on the signal
energy. After the LPC analysis, templates were
normalized to a fixed number \( N \) of frames, being \( N \) equal
to 30 for all the words. The Log-Area ratios were chosen
as feature because of their stability properties since any
kind of transformation gives an stable system.

**Feature selection training**

In this work, we use a transformation matrix \( T \)
for all the words of the vocabulary in the temporal
selection step. In this case, the covariance matrix \( Cyy \) is
obtained averaging the covariance matrix of each training
word. The results of this process are quite similar to the
Discrete Cosine transform[6].

The output of the temporal selection are templates
of M feature vectors being M equal for all templates. The
frequency selection step by linear discriminant analysis
computes a discriminant matrix for each feature vector.
For this purpose, a mean vector of the mth feature vector
is computed and used later as reference. This mean vector
is the reference prototype for the mth vector and it is
used to compute the within-class and between-class mean
distance matrix. In order to take the best discriminant
functions, the number Q can be adapted to each word or
can be fixed and equal to each word. The discrimination
information are in the eigenvalues of \( W^{-1}B \). A big
eigenvalue indicates a good discrimination property for
this feature. Table 1 shows the eigenvalues when \( M \) equal
to 3 for the word /dos/. It can be seen how the first three
eigenvectors have good discrimination properties.
Therefore the word /dos/ can be represented by three
vectors of three features.

The frequency selection step by PCA requires to
define the orthogonalization vectors \( u_j \). For this purpose,
we define two methods. The first method solves a problem
of two classes and the vectors \( u_j \) are the mean vector of
the incorrect class. The second method solves a problem
of multiple classes and the vectors \( u_j \) are the mean
vectors of a set of confusable words of the incorrect class.

<table>
<thead>
<tr>
<th>Q</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18.63</td>
<td>115.11</td>
<td>18.62</td>
</tr>
<tr>
<td>2</td>
<td>9.7</td>
<td>8.01</td>
<td>11.41</td>
</tr>
<tr>
<td>3</td>
<td>7.23</td>
<td>7.1</td>
<td>3.45</td>
</tr>
<tr>
<td>4</td>
<td>4.81</td>
<td>3.9</td>
<td>2.86</td>
</tr>
<tr>
<td>5</td>
<td>2.36</td>
<td>2.5</td>
<td>1.37</td>
</tr>
<tr>
<td>6</td>
<td>1.97</td>
<td>1.63</td>
<td>1.02</td>
</tr>
<tr>
<td>7</td>
<td>0.7</td>
<td>0.8</td>
<td>0.87</td>
</tr>
<tr>
<td>8</td>
<td>0.6</td>
<td>0.5</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Table 1. Eigenvalues of the matrix \( W^{-1}B \) for the
first three uncorrelated vectors of the word /dos/.

**IV. RECOGNITION EXPERIMENTS**

A classical pattern recognition system which
compares an input template with a set of reference
templates by means of the Euclidean distance between
frames was used. The system makes use of a linear frame
to frame comparison. The references, obtained in the
training process, are constituted by the new feature
vector obtained in the feature select process and two
transformation matrices. One of them is used to select the
temporal feature vectors and it is the same for all the
The experiments were made with a speaker independent approach. In this case, the training set was made up by ten repetitions of six speakers and three speakers were used as test. In each recognition experiment, an evidence measure was computed as E = (D2 - D1)/100; 0 ≤ E ≤ 100; where D2 is the distance to the second candidate and D1 is the distance to the first candidate.

Four experiments were performed using: a) classical speaker independent system as in [6], b) only temporal selection, c) Discriminant Analysis and d) Principal Component Analysis with constraints. Table 2 shows a resume of the results of the four experiments. In the classical system, the best results where obtained using two candidates per word. The second experiment was made using only the temporal selection step with M = 3. A mean vector was used as reference for each word. In the discriminant analysis [3], each word has only one candidate in the reference set, i.e. the mean vector of the training set. The number of temporal features M were equal to 3 and the frequency features Q were selected for each word in order to minimize the error rate (the mean value of Q was 3). With these conditions, the error rate is 1.66 % with a mean evidence of 77 %. It can be noted the high evidence mean obtained in this approach. Finally, the fourth experiment was made using the PCA with orthogonality constraints. The first step in this experiment was to find the confusion word set for each digit by means of some experiments with and without frequency selection. This confusion word set was used to form the set of orthogonalization vectors U. Thus, the transformation matrix for each vector aj was obtained using vectors of its confusion words. The best results was obtained using all the frequency features, that is, Q = 8. With this conditions, this method gives the best performance, decreasing the recognition error to 0.66 % with a mean evidence of 50 %.

<table>
<thead>
<tr>
<th>Experiment Type</th>
<th>% Error</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Clustering System</td>
<td>2.00</td>
<td>45 %</td>
</tr>
<tr>
<td>b) Temporal Selection</td>
<td>3.00</td>
<td>54 %</td>
</tr>
<tr>
<td>c) Discriminant Analysis</td>
<td>1.66</td>
<td>77 %</td>
</tr>
<tr>
<td>d) PCA with constraint</td>
<td>0.66</td>
<td>50 %</td>
</tr>
</tbody>
</table>

Table 2: Results for the speaker independent experiments.

With regard to the computational load, the number of multiplications needed for recognizing a word is very low. Using templates of NxP dimension with a transformation matrix T with M vectors, Q frequency features and V vocabulary words, the number of multiplications is (NxP x M) for the temporal selection step, VxM x P x Q for the frequency selection step and VxM x Q for the comparison step. Thus, in our experiments with the digit data base where N = 30, P = 8, M = 3 and V = 10 the number of multiplications is 960 for only the temporal selection, 1680 for the Discriminant Analysis with Q = 3, 2880 for the PCA method with Q = 8 and in the classical system with dynamic time warping and two templates per word reference is \((N^2/3) \times V \times P \times 2 = 48000\).

V. CONCLUSION

A two step feature selection process is introduced for isolated word recognition. The first step takes into account the correlation among the N frames of a template giving a new subset of uncorrelated frames. The second step takes into account the discrimination properties of the P features of each uncorrelated frame by means of two methods: Discriminant Analysis and PCA with constraints. Both methods increase the recognition performance of an independent speaker recognition system. The PCA with constraints gives the best results but its training process is more elaborated that the training process of the Discriminant Analysis. Further experiments must be made to test the feature selection process in a more difficult vocabulary.

REFERENCES


