STATISTICAL FEATURE SELECTION FOR ISOLATED WORD RECOGNITION

E. Lleida, C. Nadeu, E. Monte, J.B. Mariño

E.T.S.I. Telecomunicación, T.S.C. Dept.
Apdo. 30.002, 08080 Barcelona, Spain

ABSTRACT

In this paper we present a new procedure for feature selection in isolated word recognition (IWR). The feature selection is performed in two steps. The first step takes into account the temporal correlation among feature vectors in order to obtain a transformation matrix which projects the initial template of N feature vectors to a new space where they are uncorrelated. This step gives a new template of M feature vectors, being M<<N. The second step takes into account the frequency discrimination features which discriminate each word of the vocabulary from the others or a set of them. An important characteristic of this process is that the new templates do not need time-alignment with the references in the comparison step, avoiding the use of the dynamic time-warping process. The speech recognition results show a significant improvement in the recognition performance with a digit data base and the confusable E-set.

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 classic pattern-matching approach used in IWR assumes [1] 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. Thus, speech signal has stationary parts which are represented by several feature vectors, having a great redundancy [2,3]. 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 [4]. 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 nth LPC vector, the transformation obey the following formulation

$$y(n) = \sum_{m=1}^{M} \alpha_m \phi_m(n) \qquad (1)$$

where ϕ_m is the mth transformation function and α_m is the new mth feature vector.

A typical family of functions which performs this transformation is the Karhunen-Loève functions [5]. It minimizes the mean square error between a vector and its estimation by a linear combination. In this work, the Karhunen-Loève transform (KLT) is used to remove the temporal correlation in order to obtain M uncorrelated vectors, being M<<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 (ϕ_m). These functions are found from a training set.

In order to reduce the within-class variability and increase the separability among words, a new transformation is proposed in the frequency dimension. In this case, taking the M uncorrelated vectors, a transformation matrix associated with each vector of a word is sought. It maximizes the distance among this vector and the corresponding vectors of the other words. This correspondence among vectors is lineal because of the variance order of the vectors. Thus, a new vector δ_m, called discriminant feature vector, will be obtained by transforming the uncorrelated vector α_m with a family of discriminant functions as follows

$$\delta_m(q) = \sum_{i=1}^{P} \alpha_m(i)\, \varphi_{m,q}(i) \qquad 1 \le q \le Q \qquad (2)$$

where $\varphi_{m,q}$ is the qth discriminant function of the mth vector which is found from a training set by optimizing a criterion function that uses the between-class and within-class distance.

After these processes, a template of MxQ 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 NxP 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

This work was supported by the PRONTIC grant nº 105/88.

means of the KL-expansion, 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

$$C_{yy}^{w} = \frac{1}{P-1}\sum_{i=1}^{P} (\underline{y}_i - \bar{\underline{y}})(\underline{y}_i - \bar{\underline{y}})^t \qquad (3)$$

where

$$\bar{\underline{y}} = \frac{1}{P}\sum_{i=1}^{P} \underline{y}_i \qquad (4)$$

$$\underline{y}_i = \{y_i(1), y_i(2), \ldots\ldots, y_i(N)\}$$

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

$$C_{yy}\,\phi_m = \lambda_m\,\phi_m \qquad (5)$$

where C_{yy} can be equal to an average of the C_{yy}^{w} of each word or an average of all the covariance matrises 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 α_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. Figure 1 shows the first three eigenvectors computed averaging ten covariance matrix of the word /sɛ/. The speech signal was analized by an LPC processor and 8 Log-Area ratios was extracted in each frame. It can be seen that for this word, which is composed by an unvoiced sound followed by a voiced sound, the first eigenvector carries information of the unvoiced sound /s/, the second one about the voiced sound /ɛ/ and the third one about the transitions, specially the /s/-/ɛ/ transition. This result shows that the temporal evolution of the Log-Area ratios for the unvoiced sound is orthogonal to the voiced sound. When the word has only voiced sounds, the first eigenvector is quite similar to an average of the temporal evolution of the first log-Area coefficients [6]. 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 uncorrelated feature vector obtained in the temporal selection in order to discriminate between words. 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.

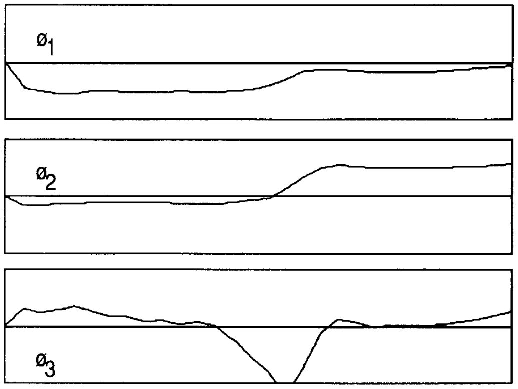

Figure 1. First three eigenvectors of the word /sɛ/.

In this step, a set of Q discriminant functions $\varphi_{m,q}$ is asociated to each vector α_m of a word which increases the separability of this vector from the mth 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 mth feature vector of any utterance of it, forms the correct class and the mth 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

$$W = E\{(\alpha_c - \alpha_c^p)\ (\alpha_c^p - \alpha_c^p)^t\} \qquad (6)$$

and the between-class mean distance matrix as

$$B = E\{(\alpha_i - \alpha_c^p)\ (\alpha_i - \alpha_c^p)^t\} \qquad (7)$$

where α_c is a realization of the correct class, α_c^p is the reference prototype of the correct class and α_i is a realization of the incorrect class, the criterion function to be maximized is defined as [4,5]

$$J = t_r(F_c B F_c^t) - \lambda(tr(F_c W F_c^t) - 1) \qquad (8)$$

where F_c is the discriminant matrix of the correct class vector, $F_c^t = [\varphi_{m,1}, \varphi_{m,2}, \ldots, \varphi_{m,Q}]$.

The solution of this optimization problem is the eigensystem $(W^{-1}B)\varphi_{m,k} = \lambda_k \varphi_{m,k}$. 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 [1], 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.

III. TRAINING PROCESS

Test data base

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

A small E-set Spanish {b,c,d,e,g,p,t} data base consisting of seven repetitions uttered by two male and one female speakers recorded in a laboratory enviroment were also used.

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 and 10 ms for the E-set 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

Two cases can be distinguished in the temporal selection training:

Case 1. A transformation matrix T_g for all the words of the vocabulary. In this case, the covariance matrix C_{yy} is obtained averaging the covariance matrix of each training word. The results of this process are quite similar to the Discrete Cosine transform[6].

Case 2. A transformation matrix T^w for each word of the vocabulary. Then, the covariance matrix C_{yy} is obtained using several repetitions of a word 'W'. In this case, each transform matrix has information about the specific temporal variation of the Log-Area ratios which form the word.

The output of the temporal selection are templates of M feature vectors being M equal for all templates. The frequency selection step 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.

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 can be the same for all the words or specific for each word. The other one is used to select the frequency features and it is specific for each frame obtained in the temporal selection step.

Q \ M	1	2	3
1	18.63	115.11	18.62
2	9.7	8.01	11.41
3	7.23	7.1	3.45
4	4.81	3.9	2.86
5	2.38	2.5	1.37
6	1.97	1.63	1.02
7	0.7	0.8	0.87
8	0.6	0.5	0.55

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

Due to the small data base avalaible, two kinds of experiments were made. The first experiment was made taking six repetitions of the nine speakers digit data base as training. In each recognition experiment, an evidence measure was computed as $Ev=(D2-D1)100/D1$; $0 \le Ev \le 100$; where D2 is the distance to the second candidate and D1 is the distance to the first candidate. Figure 2 shows the recognition results obtained for different values of M and Q using both transformation matrices T_g and T^w.It can be seen that the best results, 0.22 % of error rate with a mean evidence of 85,4 %, are obtained with the transformation matrix T_g when M=3 and Q=2, being M and Q equal for all the words. These results show that the T^w matrices do not have good discrimination properties when a word w' different of w is projected with the matrix corresponding to the word w.

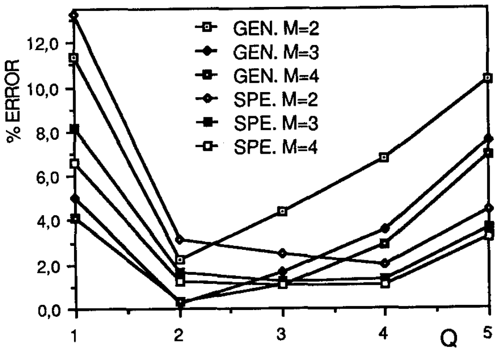

Figure 2. Error rate for several values of M and Q using a general temporal transformation matrix T_g (GEN.) or an specific transformation matrix T^w (SPE.) in the temporal selection step.

In order to have significant results, ten recognition experiments were made taking in each recognition experiment a training set with (six) different repetitions (multispeaker experiment). In this way, all the data base was used as test. In this experiment M and Q were fixed and equal to 3 and 2 respectively and the T_g matrix was used for temporal selection. Table 2 shows the confusion matrix of this experiment. The worst results are obtained by the word /trɛs/,/u/ and /sink/. Nevertheless, the error rate is small, 1.19 % for utterances out of the training set and 0.26 % for utterances inside the training set. An experiment with only the temporal selection step shows a deterioration of the recognition rate to 7 % showing the need of improving this step.

Recognized Word

	0	1	2	3	4	5	6	7	8	9
0	893					4				3
1		888						2		10
2			900							
3			10	885			5			
4					900					
5	8					884		8		
6							900			
7						3	3	894		
8			1						889	
9										900

Total error: 57 (14 (0.22%) within and 43 (1.19%) out of the training set)
mean evidence: 85.4 %

Table 2. Confusion matrix for the multispeaker experiment.

The second experiment was 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. The same experiment was performed using a classical speaker independent system as in [7] where the best results were obtained using two candidates per word. The results are shown in table 3. In our system, each word has only one candidate in the reference set,i.e. the mean vector of the training set, using a transformation matrix T_g for all the words of the vocabulary. The number of temporal features M were equal to 3 and the frequency features Q were selected for each word in orden to minimize the error rate. With these conditions, the error rate is 1,66 % with a mean evidence of 77 % in our system and an error rate of 2 % with an evidence of 45 % for the clasical system. It can be noted the high evidence mean obtained in our approach.

	% error	evidence
Clustering System	2.00	45 %
Feature Selection	1.66	77 %

Table 3. Results for the speaker independent experiments.

With the E-set data base, the results are quite different. In this case, the optimal number M of temporal features is 9 and Q equal to 4 with the T_g matrix. These results show the difficulty of this data base where the most significant diferences are in the transitions and this information needs several eigenvectors to be retained. With these values of M and Q, the recognition rate is 12.53 %, averaging seven experiments where six different repetitions were used as training in each experiment. Table 4 shows the confusion matrix.

Recognized Word

	B	C	D	E	G	P	T
B	129	1	7				
C	1	137		8			1
D	7	1	141				
E				136		4	7
G		3			144		
P	1	9		15		109	13
T				32		9	106

Total error: 129 (12.53 %)
mean evidence: 53 %

Table 4. Confusion matrix for the E-set experiment.

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_g with M vectors, Q discriminant vectors and V vocabulary words, the number of multiplications is (NxPxM) for the temporal selection step VxMxPxQ for the frequency selection step and VxMxQ for the comparison step. Thus, in our experiments with the digit data base where N=30,P=8,M=3,Q=2 and V=10 the number of multiplications is 1308, when in a classical system with dynamic time warping and one template per word reference is $(N^2/3)xVxP = 24000$.

IV. 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, giving a new frame with Q discriminat features. The transformation matrices are obtained in a training process. Although the test data base was small this approach shows a potential improvement on an IWR system giving a small error rate and a very small computational load. Further studies will be made in order to improve the recognition rate in the temporal selection step.

REFERENCES

[1]E.L Bocchieri, G.R. Doddington, "Frame-specific statistical features for speaker independent speech recognition". IEEE trans. on ASSP, Vol 34, Ag. 1986.

[2]E. Lleida, C. Nadeu, J.B. Mariño, "Speech parametrization and recognition using block and recursive linear prediction with data compression", European Conference on Speech Technology, pp. 300-303, Edinburgh- 1987.

[3]R. Pieraccini, R. Billi, "Experimental comparison among data compression techniques in IWR", ICASSP-83, Boston.

[4]E. Lleida, C. Nadeu, J.B. Mariño, "Feature selection through orthogonal expansion in IWR", MELECON-89, Lisboa, 1989.

[5]K. Fukunaga, *Introduction to statistical pattern recognition*, Academic Press, 1972.

[6]E. Lleida, "Feature compression and selection in speech recognition", Ph. D. thesis (in Spanish), Universidad Politécnica de Cataluña, 1989.

[7]L.R. Rabiner et al. "Speaker-Independent recognition of isolated words using clustering techniques", Trans. on ASSP-27, Ag. 1979.