FEATURE SELECTION THROUGH ORTHOGONAL EXPANSION IN ISOLATED WORD RECOGNITION

Eduardo Lleida, Climent Nadeu, José B. Mariño

Dpto. Teoría de la Señal y Comunicaciones, Universidad Politécnica de Cataluña, Barcelona, Spain

SUMMARY

The use of an orthogonal expansion for feature selection and data compression in isolated word recognition is presented in this paper. Assuming that the spectral evolution, given by a LPC analysis, is a noisy measure in the sense that is a linear combination of a set of "real" features, the orthogonal expansion has as objective to find these "real" features. These new feature vectors are used to perform the recognition process. The recognition results as well as the meaning of the orthogonal expansion are discussed.

1. INTRODUCTION

Signal feature measurement plays an important role in pattern recognition systems. Basically, it is a data reduction technique whereby a large number of speech samples are transformed into a smaller set of features. This set of features is called 'Template' in the IWR environment. Generally, feature measurement methods are block processing models, but speech signal has stationary parts which are represented with several blocks. Therefore, the new representation space still has redundancy. This block processing can be done by several methods [1], and the LPC technique was chosen in this work. Thus, a Template is composed by N vectors of p dimension, being p the order of the LPC analysis. This N vectors are an uniform sampling of the temporal variation of the LPC spectrum.

Many methods have been developed to reduce the size of the Template using, generally, the spectral distance to pick up the most relevant spectrum. 'Trace Segmentation', 'Variable Frame Coding', 'Vector Quantization' are examples of these techniques [2].

In speech recognition, we wish to select information about the words being spoken and to jettison as much irrelevant information as possible. One way is to reduce the variability within-class (repetitions of the same utterance)

and to increase the separability among the different words in the vocabulary [3]. Another way is to assume that there is an underlying set of "real" features all independent, and that each of the measures we are working on is an "impure" feature in the sense that is a linear combination of these "real" features. Then, the objective is to find a transformation which recovers the "real" features [4].

Given an input Template of $N \times p$ dimension, this transformation must obtain a set of parameters, output Template, with $M \times p$ dimension, being $M \ll N$. Different restrictions for computing the transformation matrix give different sets of parameters but, all of them obey the same formulation

$$y_i(n) = \sum_{m=1}^M \alpha_{mi} \phi_m(n) \quad (1)$$

where, $1 \leq n \leq N$; $1 \leq i \leq p$; $y_i(n)$ is the n th LPC vector, $\phi_m(n)$ is the m th transformation function and α_m is the new feature vector.

In this work, a special set of ϕ -functions, known as orthogonal functions, have been used. So, we try to characterize the temporal variation of the LPC spectrum with a set of orthogonal functions. These orthogonal functions are obtained from the input Template and expand a new vector space where the input Template is represented with less coefficients. Furthermore, this technique can be seen as a method to describe and represent the temporal evolution of the LPC parameters [5,6]. So, the sequence $y_i(n)$ can be represented by a superposition of the members of any orthogonal family of functions. The best family of orthogonal functions will be that which fits better the temporal evolution, and then the number of new coefficients will be minimum.

We present here an isolated-word speaker dependent recognition system that makes use of a family of orthogonal functions computed by means of a modified Karhunen-Loève expansion to describe the temporal evolution of the LPC

spectrum. The new coefficients are used to carry out the word recognition.

In section 2 a description of the mathematical formulations is presented. Section 3 will explain the test data, the training procedure and the results of the isolated word recognition.

2. ORTHOGONAL EXPANSION

Given a $N \times p$ matrix Y of spectral parameters $\{y_i(n)\}$ representing N frames, a finite family of orthogonal functions, which represent the time evolution of spectral parameters, can be found in accordance with (1). The purpose of this expansion is to find a new $M \times p$ matrix α with $M \ll N$. The new coefficients $\alpha_i(m)$ are called the spectral coefficients of the LPC spectrum time evolution relative to the new orthogonal basis.

We can present the orthogonal transformation in matrix notation as

$$\alpha = \phi^T Y \quad (2)$$

$$Y = \phi^{-1} \alpha \quad (3)$$

the basis matrix ϕ belongs to the class of orthogonal transforms defined by

$$\phi^{-1} = \phi^T \quad (4)$$

the matrix ϕ is the $N \times N$ basis matrix of the transformation, Y is the $N \times p$ input matrix and α is the $N \times p$ transform matrix. If we want to use M functions of the N possible functions,

$$Y' = \phi'^T \alpha' \quad (5)$$

with $\phi' = \{\phi_1 \ \phi_2 \ \phi_3 \ \dots \ \phi_M\}$, the error commits in the approximation of each y_i vector is

$$E_i = \|y_i - y_i'\|^2 = \sum_{j=M+1}^N \phi_j^T y_i y_i^T \phi_j \quad (6)$$

if we define the total error as

$$E = \sum_{i=1}^p E_i \quad (7)$$

it can be shown that

$$E = \sum_{j=M+1}^N \phi_j^T M_{yy} \phi_j \quad (8)$$

where $M_{yy} = Y Y^T$

Minimizing the error function with respect to the unknown functions ϕ_j we can obtain that

$$M_{yy} \phi_j = \beta_j \phi_j \quad (9)$$

and the error is given by

$$E = \sum_{j=M+1}^N \beta_j \quad (10)$$

Thus, the transform matrix ϕ is composed by the M eigenvectors with the M largest eigenvalues, ranking them from the largest to the smallest one. The new coefficients α_j take information of all frames of the input template and they are mutually orthogonal. Then, we are assuming that the coefficients with the largest variances are the "real" features and those with the smallest variances are noise. So, the optimum number M of orthogonal functions must be related to the number of different sounds that make up the utterance.

The matrix M_{yy} can be seen as an estimation of the correlation matrix of the vector y_i , assuming that the vectors y_i , $1 \leq i \leq p$, belong to the same class. Thus, this procedure is a kind of Karhunen-Loève expansion, rotating the feature space of temporal variation in order to uncorrelate the feature vectors and to make null the inner product among feature vectors.

It can be noted that this procedure is applied to all input template in one step and we obtain a description of all the utterance. Applying a criterion of compactness in the search of the ϕ -functions (1) we can obtain the Time Decomposition method developed by Atal [7].

3. RECOGNITION EXPERIMENTS

Test data base

A data base consisting of ten repetitions uttered by six male and three female speakers (900 words) of the Catalan digits (table 1) recorded in a quiet room were used.

1 /u/ 2 /dos/ 3 /tres/ 4 /kuatra/
5 /sink/ 6 /sis/ 7 /set/ 8 /vuit/ /buit/
9 /nou/ 0 /zeru/ /seru/

table 1. Pronunciation of the Catalan digits.

Preprocessing

The speech signal was sampled at 8 KHz, pre-emphasized ($H(z)=1-0.95z^{-1}$) and 8 Parcor coefficients were computed each 15 ms. 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 means

of an algorithm based on the signal energy. After the LPC analysis, templates are normalized to a fixed number N of frames, being N equal for all the words.

Recognition system

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. Two methods of comparison have been developed, one makes use of a single transformation matrix to obtain the new feature vectors and the other one makes use of a transformation matrix for each word of the vocabulary. Then, references are constituted by the new feature vectors and a transformation matrix. In this way, given an input template, the new feature vectors are computed using the transformation matrix of reference, which is the same that was used to compute the reference feature vectors.

In the recognition test, each repetition of the word vocabulary was alternatively taken as reference and the others as test, so obtaining 900 test for each speaker. As a result, a distance and an evidence defined as

$$EV = (D2 - D1) 100 / D2 \quad (11)$$

$$0 \leq EV \leq 100$$

where D2 is the distance to the second candidate and D1 is the distance to the first candidate, is obtained for each recognition test.

Training process

Case 1. A transformation matrix for all the words of the vocabulary. In this case, the matrix M_{yy} is computed using one or more repetitions of each word. This is done averaging the matrix M_{yy} of each word. Fig. 1 shows the effects of these ϕ -functions in the word /sink/. A drawing of the ϕ functions shows that this transformation is quite similar to the Discrete Cosine Transform. Thus, the first new feature vector takes information of the average spectrum and the next ones take information about the sinusoidal variations of the temporal evolution of the spectrum.

Case 2. A transformation matrix for each word of the vocabulary. Then, the matrix M_{yy} is computed for each word using one or more repetitions of them. In this case, each transformation matrix has information about the specific temporal variation of the spectra which make up the word. Fig. 2 shows the effects of transforming the word /sink/ with its specific transformation matrix. It can be noted that the first ϕ -function takes information about the voiced sound /n/, the second one about the unvoiced sound /s/ and the third one about the voiced sound /i/. The next ϕ -functions take information about the transition

intervals and small variations in the spectral evolution

Recognition results

Recognition experiments were carried out with the two training methods. Table 2 shows the recognition errors obtained for several values of M when a single transformation matrix is used. It can be noted how the best results are obtained when M, the number of ϕ -functions used to transform, takes the value of 3. A recognition error of 2,14 % is achieved using only 3 new feature vectors. Table 3 shows the confusion matrix when M is 3 and Fig. 3 plots the distribution of the evidence. It can be seen that the values of the evidence are quite high, taking the maximum value for evidences between 80 and 90 %. This shows that the discrimination among different words provided by this new set of features is quite good. In the table 3 can be seen that the worst results are obtained by the words /vuit/ and /set/ because of the difficulty for detecting the final sound /t/ that sometimes were not uttered by the speakers.

Table 4 shows the recognition errors when a transformation matrix for each word is used. An recognition error of 2,95 % was obtained when M is 3, being worse than the preceding case. In any case, the evidence distribution and confusion matrix are quite similar. Also, some experiments were carried out recognizing in the original space by using the template Y' reconstructed by M ϕ -functions of the single transformation matrix. Table 5 shows the recognition errors. It can be noted that the results for $M < 10$ are worse than the obtained in the transform space, showing that the new feature space has better discrimination properties.

4. CONCLUSION

In this paper, the use of an orthogonal expansion for feature selection in IWR is presented. We obtain a new set of feature vectors which describes the input template in an orthogonal base. These new feature vectors are used to perform the recognition process, given good results for a few vectors. Also, the new features increase the discrimination among different words, being the optimum number of ϕ -functions related to the number of different sounds that make up each word. The computation time is reduced a lot because of the compression rate is about 10. Further restriction for computing the transformation matrix will be studied to describe better the temporal evolution of the spectrum.

REFERENCES

- [1] W. A. Lea, "Electronic Speech Recognition: The Elements of Speech Recognition", MacGraw- Hill, 1986.
- [2] R. Pieraccini, R. Billi, "Experimental Comparison among Data Compression Techniques in Isolated Word Recognition", ICASSP-83, Boston.
- [3] W. Ahmed, G. Bekey, "A New Algorithm for Isolated Word Recognition with large Within-Class Variability", ICASSP-85, pp. 878-881, 1985.
- [4] T. Parsons, *Voice and Speech Processing*, McGraw-Hill, 1987.
- [5] M.G. Hall, A. Openheim, A.S. Willsky, "Time Varying Parametric Modeling of Speech", *Signal Processing*, 1983.
- [6] A. Torices-Arguelles, J. Rubio-Ayuso, "On Isolated Words Recognition Systems using Time Dependent Linear Prediction", *European Conference on Speech Technology*, pp. 280-283, Edinburgh 1987.
- [7] B.S. Atal, "Efficient Coding of LPC parameters by Temporal Decomposition", *Proc. ICASSP-83*, pp. 81-84, 1983.

M	1	2	3	4	10	20
%	12.81	2.80	2.14	2.42	2.88	2.99

Table 2. Recognition errors for a single transformation matrix (Case 1).

		word recognized									
		0	1	2	3	4	5	6	7	8	9
word to be recognized	0	797					6		1		6
	1		792			9				4	15
	2			796	14						
	3				810						
	4					810					
	5	5					791	1	13		
	6							3	804	2	
	7	15					17	5	773		
	8			13		25				766	6
	9	6	5							1	799

Table 3. Confusion matrix for a single transformation matrix.

M	1	2	3	4	10	20
%	9.50	3.36	2.95	3.23	3.10	3.03

Table 4. Recognition errors for specific transformation matrix (Case 2).

M	1	2	3	4	5	10	20
%	12.92	3.16	2.80	2.97	2.94	2.11	1.75

Table 5. Recognition errors in the Y space.

Fig.3. Distribution of the evidence.

Fig.1. a) Spectral sequence for the word /sink/. b)c)d) spectral sequence obtained using the first, the second and the third ϕ -function respectively. e)d) id. using the first three ϕ -functions. f) id. using the first eight ϕ -functions.

Fig.2. a)b)c) Spectral sequence obtained using the first, the second and the third ϕ -function respectively for the word /sink/. d) id. using the first three ϕ -functions.