CAPITULO DE : "SPEECH RECOGNITION ..."
SPRINGER-VERLAG, 1991

Specific Distance for Feature Selection in Speech Recognition

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

Dpto. of Signal Theory and Communications,U.P.C., Apdo. 30.002, 08080 Barcelona, Spain

Abstract: In this paper, the use of a specific metric as a feature selection step is investigated. The feature selection step tries to model the correlation among adjacent feature vectors and the variability of the speech. We propose a new procedure which performs the feature selection in two steps. The first step takes into account the temporal correlation among the N feature vectors of a template in order to obtain a new set of feature vectors which are uncorrelated. This step gives a new template of M feature vectors, with M << N. The second step defines a specific distance among feature vectors to take into account the frequency discrimination features which discriminate each word of the vocabulary from the others or a set of them. Thus, the new feature vectors are uncorrelated in time and discriminant in frequency.

Keywords: Feature selection, Specific distance, Data Compression, Karhunen-Loève transform, Principal component analysis, Discriminant analysis.

I. INTRODUCTION

The problem of feature selection in speech recognition can be studied by different ways. One way is to study the feature selection as a problem of *data compression* to reduce the computational time and memory requierements. From this point of view, a lot of similar techniques were implemented in the last years where a new template is obtained removing those feature vectors which are similars. Trace segmentation and variable frame coding are two classical techniques [1-3]. Another point of view, proposed in this paper, is to assume that there is an underlying set of "real" uncorrelated features, and the features we are working on are "impure" in the sense that they are a linear combination of those "real" features. Then, the objective is to find a transformation which recovers the "real" features [4]. These two points of view select the feature in the time dimension, that is, it is a *temporal selection*. Thus the *temporal selection* obtain a new template where the feature vectors are uncorrelated or without temporal redundancy. However, the *temporal selection* doesn't take into account the variability and separability among words. Thus, a *frequency selection* step which reduces the within-class variability and increases the separability among words is needed. This is the second step proposed in our feature selection procedure which is done by defining a *specific distance* for each feature vector. Thus, we propose a feature selection procedure which makes use of a representation criteria for *temporal selection* and discriminant criteria for *frequency selection.* Figure 1 shows the process of *temporal* and *frequency selection.*

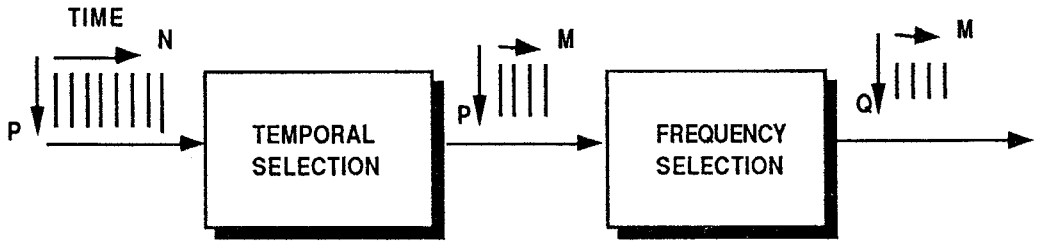

Figure 1. Two step feature selection procedure (M<N; Q<P)

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

II. 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. Basically, the problem is to represent the sequence of spectra by a superposition of the members of an orthogonal family of functions where the input template is representated with less coefficients.

Given a NxP matrix Y of spectral parameters $\{y_i(n)\}$ representing N frames of P "impures" features, the transformation obey the following formulation

$$y_i(n) = \sum_{m=1}^{N} \alpha_m(i)\phi_m(n) \qquad 1 \le i \le P \; ; \; 1 \le n \le N \qquad (1)$$

where $\alpha_m(i)$ is the mth "real" feature vector and $\phi_m(n)$ is the mth transformation function of the transformation matrix {T}.

There are two definitions of transformation functions:

1- *Data independent transformation functions*. The transformation functions are members of a deterministic family of orthogonal functions as the Discret Cosinus Transform (DCT).

2- *Data dependent transformation functions*. The transformation functions are found from the data using a criteria of minimum square error. The error of representing $y_i(n)$ with M "real" feature vectors is defining as follow

$$\varepsilon = E\{(y_i - \hat{y}_i)(y_i - \hat{y}_i)^t\} \qquad (2)$$

where $\hat{y}_i = \sum_{m=1}^{M} \alpha_m(i)\phi_m(n)$ is the estimation of "impure" feature vectors with M "real" feature vectors. With the constraint that the transformation functions are orthogonal, the transformation functions are found solving the eigensystem

$$C_{yy}\phi_m = \lambda_m\phi_m \qquad (3)$$

where C_{yy} is the covariance matrix defined as

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

with

$$\bar{\underline{y}} = \frac{1}{P}\sum_{i=1}^{P} \underline{y}_i \qquad (5)$$

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

From this eigensystem, N eigenvalues and their corresponding eigenvectors are obtained. However, only the M eigenvectors with the *largest* eigenvalues are retained. Thus, the transformation matrix {T} is composed by the M eigenvectors with the M largest eigenvalues, ranking them from the largest to the smallest one. Because of the orthogonal property of the transformation functions, the new "real" feature vectors are computed as a linear combination of the "impure" feature vectors as follows [5]

$$\alpha_m(i) = \sum_{n=1}^{N} y_i(n)\phi_m(n) \qquad 1 \le i \le P \; ; \; 1 \le m \le M \qquad (6.a)$$

which is known as the Karhunen-Loève transform (KLT) or

$$\alpha_m(i) = \sum_{n=1}^{N} (y_i(n) - \bar{y}_i(n))\phi_m(n) \qquad 1 \le i \le P \; ; \; 1 \le m \le M \qquad (6.b)$$

which is Known as Principal Component Analysis (PCA).

The principal properties of the new representation are:

-Coefficients with the largest variances are the "real" features.
-The new "real" features are uncorrelated.
-Feature vectors are arranged in variance order, thus, *no time-alignment* is needed in the comparison step.

The transformation matrix is computed in the training process. We distinguish two cases:

1- *General matrix.* 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.

2- *Specific matrix.* A transformation matrix T^w for each word of the vocabulary. Then, the covariance matricx C_{yy} is obtained using several repetitions of the word 'w'.

Figure 2 shows the evolution of the first three transformation functions for the *General matrix* and for the *Specific matrix* of the word /sɛt/.

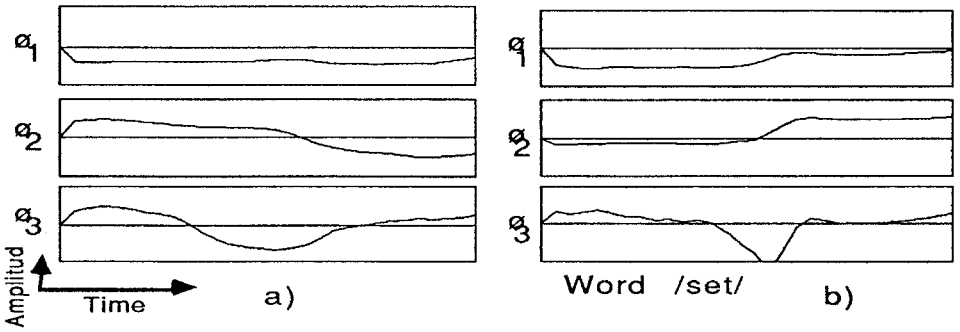

Figure 2. a)First three transformation functions of the *General matrix*. b)First three transformation functions of the word /sɛt/.

III. 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. This step can be seen as a method for finding a *specific distance measure* for each reference vector. This *specific distance* takes into account the discriminant properties of the feature vectors which reduces the within-class variability and increases the separability among feature vectors. Thus, the *frequency selection* step is related with the comparison step. Defining the weighted Euclidean distance between the test vector α_i and the reference vector α_j as

$$d(i,j) = \| F_j (\alpha_i - \alpha_j) \|^2 \qquad (7)$$

a *specific distance* matrix $F_j = \{fd_1, fd_2, ..., fd_Q\}$ of Q weighting vectors for each reference vector of each word has to be computed. Figure 3 shows the relation between *specific distance* and *frequency selection.*

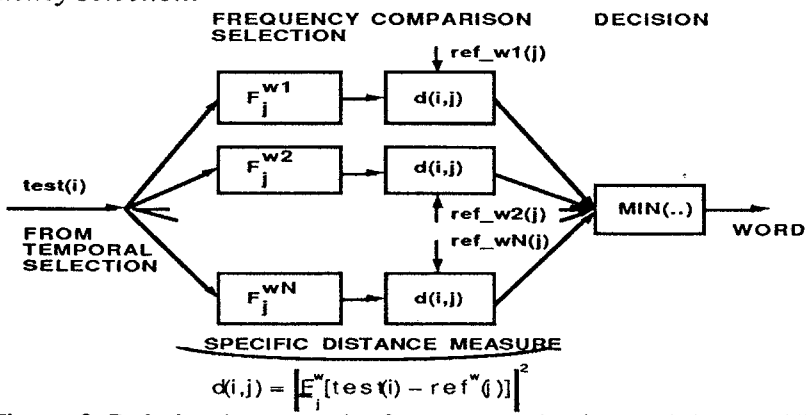

Figure 3. Relation between the frequency selection and the specific distance measure point of view

In order to find the discriminant matrix F_j, two classes of vectors are defined. For a word 'w', the mth feature vector of any utterance of it forms the *correct class* (α_{ce}) and the mth feature vector of the other words forms the *incorrect class* (α_i). Thus, defining the *mean interclass distance* as the mean distance between the incorrect class vectors and the mean correct class (α_c) feature vector

$$E(D_{inter}) = E(\|F_c\ (\alpha_i - \alpha_c)\|^2) \tag{8}$$

where α_c is mean vector of the correct class, and taking into account the matrix relation

$$\|a\|^2 = tr(a\ a^t) \tag{9}$$

we can rewrite (.) as

$$E(D_{inter}) = tr(F_c\ E((\alpha_i-\alpha_c)(\alpha_i-\alpha_c)^t)F_c^t)) \tag{10}$$

defining the *between-class mean distance* matrix as

$$B = E((\alpha_i-\alpha_c)(\alpha_i-\alpha_c)^t) \tag{11}$$

the mean interclass distance is as follow

$$E(D_{inter}) = tr\ (F_c\ B\ F_c^t) \tag{12}$$

In order to take into account the within-class variability a *mean intraclass distance* is defined as

$$E(D_{intra}) = tr\ (F_c\ W\ F_c^t) \tag{13}$$

where

$$W = E((\alpha_{ce} - \alpha_c)(\alpha_{ce} - \alpha_c)^t) \tag{14}$$

In this way, the criterion function to be maximized is [4,6]

$$J = tr(fd_k\ B\ fd_k^t) - \lambda(tr(fd_k\ W\ fd_k^t) - 1) \tag{15}$$

The solution of this optimization problem is the eigensystem

$$(W^{-1}B)fd_k = \lambda_k fd_k \tag{16}$$

therefore, the specific distance matrix is formed by the Q eigenvectors with the Q *largest* eigenvalues of $W^{-1}B$, whenever

$$E(D_{inter}) = \sum_{k=1}^{Q} \lambda_k \gg Q = E(D_{intra}) \tag{17}$$

The discriminant properties can be found in the matrix $W^{-1}B$ because

$$tr(W^{-1}B) = \sum_{k=1}^{P} \lambda_k \tag{18}$$

This process is made for each reference vector of each word of the vocabulary. In the training process, a mean vector (α_c) for each feature vector is computed and used as reference to compute de within-class and between-class mean distance matrices as well as reference vector for the recognition process. Figure 4 shows of the projection of the first feature vector of several words using the first two discriminant eigenvectors of the word /dos/.

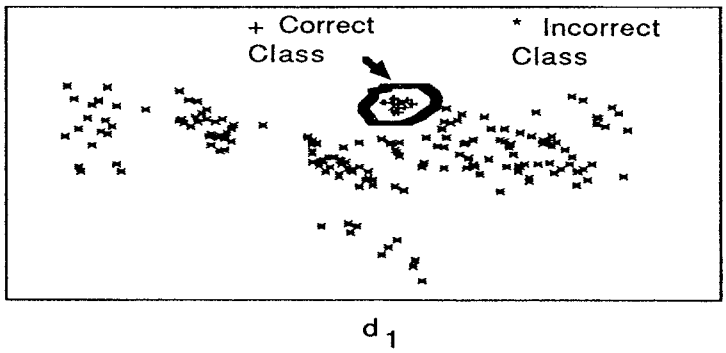

Figure 4. Projection of the first feature vector of several words using the first two discriminant eignevectors of the word /dos/ (correct class).

IV. RESULTS AND CONCLUSIONS

The recognition experiments were made using two data base. A data base consists of ten repetitions of the Catalan digits {u,dos,tres,kuatra,sink,sis,set,vuit,nou,zeru} uttered by six male and three female speakers (900 words) and recorded in a quiet room. The second data base is the Spanish E-set {b,c,d,e,g,p,t} uttered by two male and one female speakers, seven repetitions, recorded in a laboratory enviroment.

For feature extraction, the speech signal was sampled at 8 KHz, preemphasized ($H(z)=1-0.95z^{-1}$) 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. A typical Hamming smoothing window was applied to the data. After the LPC analysis, template were normalized to a fixed number N of frames, with N equal to 30, to apply the temporal selection step.

A classical pattern recognition system which compares an input template with a set of reference templates by means of a linear frame to frame comparison was used. The references, obtained in the trainning process, are constituted by the feature vectors obtained in the feature selection process and two transformation matrices. One is used to temporal selection and the other is to frequency selection which is specific for each frame.

We present two experiments. The first experiment is a multispeaker experiment. Six repetitions of the nine speakers digit data base are used as training. In each recognition experiment, an evidence measure was computed as Ev=(D2-D1)100/D1, where D2 is the distance to the second candidate and D1 is the distance to the first candidate. Figure 5 shows the recognition results obtained for different values of M and Q using both temporal transformation matrices T_g (General matrix) and T^w (Specific matrix). For the best result, M=3 and Q=2 with T_g, the mean evidence is 85,4 % showing the good discrimination properties of this feature selection process.

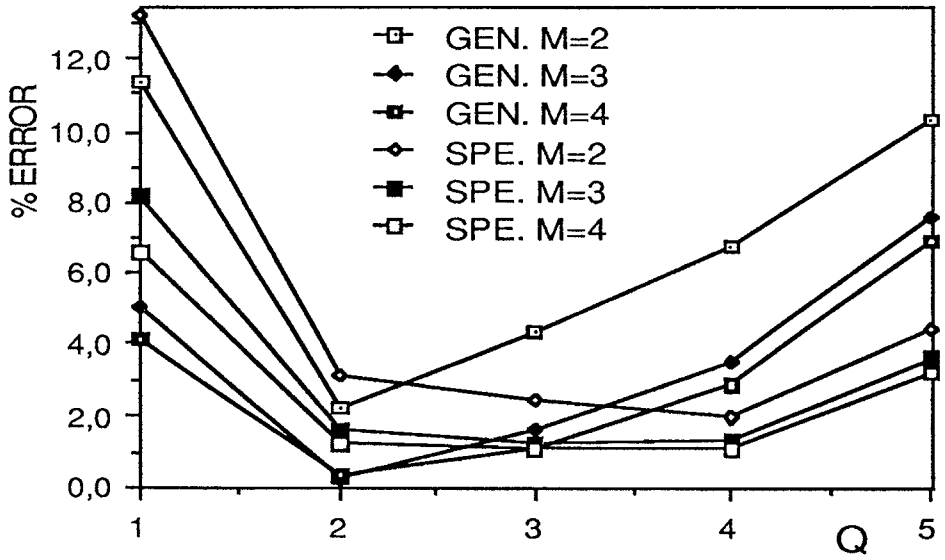

Figure 5. Error rate for several values of M and Q.

With the E-set data base, the results are quite different. In this case, the optimal number M of temporal features is 9 (General matrix) and the discriminant features is 4. 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. Figure 6 shows the confusion matrix.

	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

Figure 6. Confusion matrix for the E-set experiment

The second experiment is a speaker independent experiment. In this case, the training set is made up by ten repetitions of six speakers of the digit data base, using three speakers are as test. Figure 7 shows the results for three different versions: a) classical independent system with a clustering process, b) only temporal selection step and c) temporal and frequency selection steps. This figure shows also the computational load in number of multiplications per recognized word. It can be noted that the best result is obtained using the two steps feature selection process which increase the recognition rate decreasing the number of features and the computational load.

		% error	evidence	# references per word	# features per word	computational load
a)	Classical system	2.00	45 %	2	480	48000
b)	Temporal selection	3.00	54 %	1	24	1000
c)	Two steps selection	1.66	77 %	1	9	1500

Figure 7. Results for the speaker independent experiments

As conclusion, a two step feature selection procedure has been introduced. This feature selection procedure transforms an input template of features to a new representation where the correlation among vectors is removed by using the idea of *Principal Component Analysis* or *Karhunen-Loéve transform* and the variability and separability among words is taking into account using the idea of *Discriminant Analysis* and *Specific distance measure*. As a result, the new template has the feature vectors arranged in variance order, therefore, no time-alignment is needed in the comparison step and each reference vector has associated a specific distance measure which select the discriminant features. The result is an important improvement in the error rate and a very small computational load.

V. REFERENCES

1. Pieraccini R., Billi R.: Experimental comparison among data compression techniques in IWR. ICASSP-83, Boston.
2. Kuhn M.H., Tomaschewski H.: Improvements in IWR. IEEE. trans. on ASSP, vol-31, pp 157-167, Feb. 1983.
3. Lleida E., Nadeu C., Mariño J.B: Speech parametrization and recognition using block and recursive linear prediction with data compresion. European Conference on Speech Technology, pp. 300-303, Edinburgh-1987.
4. Lleida E.:Feature compression and selection in speech recognition. Ph.D. thesis (in spanish), Universidad Politecnica de Cataluña, 1990.
5. Gerbrands Jan J.: On the relationships between SVD,KLT and PCA. Pattern Recognition, vol 14, pp. 375-381, 1981.
6. Fukunaga K.:Introduction to statistical pattern recognition. Academic Press, 1972.

List of the words to be included in the volume index: