For the conference on "Digital Processing of Signals on Communications", at Unversity of Loughbrough, England 22-25 April 1985.

CRITERION BASED CLUSTERING TECHNIQUES APPLIED TO SPEAKER INDEPENDENT SPEECH RECOGNITION

A.MOKEDDEM* - H.HUGLI* - F.PELLANDINI*

SUMMARY

This paper addresses speaker independent speech recognition (SISR) performed by an approach which uses multiple template references. Two contributions are made. We present first an analysis of isolated word utterances which gives insight into the nature of speaker variability and justifies the multiple template approach. We then present the new application to SISR of criterion based clustering procedures. After the theoretical description, we give the results of tests performed with an isolated word recognizer showing the very good performance of the proposed procedures.

1. Introduction

The aim of speaker independent speech recognition is to recognize the speech pronounced by any speaker from a large population. The main problem to be solved is the large variation of pronounciation of the same word among different speakers, the so-called inter-speaker variation. To solve it, the multiple template reference approach turned out to be successful /4/,/7/. In this approach, we describe the inter-speaker variation statistically : one tries to describe all different utterances by few and typical pronounciations which, then, represent them during recognition. All such representatives are templates to be matched. They form, once grouped according to the word they represent, multiple template references .

The question arises whether such groups or clusters of utterances really exist. In order to answer it, we study the distribution of the isolated word utterances pronounced by various speakers by means of factor analysis. With this approach we can represent the utterances in a space R^m with m small, typically R^2 This representation permits one to analyse in a visual way the distribution of the words. This is our first contribution.

* Institut de microtechnique de l'université de Neuchâtel 71 rue de la Maladière 2000 Neuchâtel 7 - Switzerland

In selecting the representatives, automatic clustering can be used. The various known methods can be divided into two categories according to the way they build up clusters. We call cluster-parallel such procedures where all clusters are build up simultaneously and cluster-sequential such procedures where clusters are build up sequentially.

Among the clustering procedures previously applied to speaker independent speech recognition (SISR), k-means iteration (or basic Isodata) and Isodata /5/,/7/,/8/ are of the first category, unsupervised learning procedures UWA and UFA /3/,/4/,/6/ of the second.

Here we propose and present the novel application to SISR of other clustering procedures, namely clustering based on a criterion function. There is the Criterion based Exchange procedure (CEx) and the Criterion based Threshold procedure (CTh) described as follows.

The first, the CEx, is a cluster-parallel clustering procedure which improves the partition quality iteratively by transferring elements from cluster to cluster. These element transfers are governed by a criterion function and in that, this method differs from the basic isodata procedure.

The second, the CTh, is a cluster-sequential clustering procedure that iteratively removes from the set of elements to be clustered, the elements forming the best cluster. At each iteration, the cluster chosen is the one, among all clusters found by distance thresholding, which minimizes a criterion function. The fact that each cluster is a possible candidate makes this a general version of the UWA clustering algorithm previously used.

Both clustering procedures will be presented and, based on isolated word recognition tests, performance figures will be given. This is our second contribution.

2. Recognizer

To put further results in their context, we describe here the isolated word recognizer used.

2.1 Preprocessing

The short term energy spectrum is measured by a 14 channel filter bank, covering logarithmically the frequency range from 75 Hz to 4800 Hz and sampled every 10 ms, resulting in a spectrogram x(k,l) where k is the k-th channel and l is the l-th instant of sampling.

Start and end of word detection is achieved by an algorithm using two thresholds adapted to ambient noise, one for low frequency channels and the other for high frequency channels.

Two normalizations are made. The first, the amplitude normalization, compensates the global and local variations of the voice level. It is achieved by dividing each element of the input matrix by a normalizing factor f(k,l). We use a normalization per zone that associates one zone z(k,l) to every element x(k,l) of the input matrix. The factor f(k,l) is then equal to the mean value of x(k,l) over the zone.

One bit quantization was chosen to obtain well compressed data /2/.

The second normalization, time normalization, consists of compressing the time axis linearly to the same fixed length.

Finally we obtain the utterance features as a 280 bit binary matrix.

2.2 Comparison

Time alignment is done by dynamic programming (DTW) /1/. In the particular form used here, the path range is extended at both the begining and the end of the two words to be compared in such a way that word limit detection errors can be compensated.

3. Analysis of inter-speaker variation

With the multiple template approach to SISR one tries to describe all different pronounciations of a word with a limited number of clusters. This approach admits implicitly the following hypothesis : among a given population, there exist different typical pronounciations of each word. We are going to verify the validity of this hypothesis by analysing utterances of isolated words by different speakers.

Let n be the number of utterances pronounced by the different speakers. We can then compute the distances d(i,j) (i=1,...,n , j=1,...,n) between the various utterances. Under certain conditions /10/, we can represent the n pronounciations in a space R^{n-1}, where n-1 is the dimension of that space, in such a way that the distances are exactly equal to the original ones. Obviously, a representation in R^{n-1} cannot be used for a visual analysis. We seek therefore a representation of the utterances in a sub-space R^m with m small, for exemple m=2,

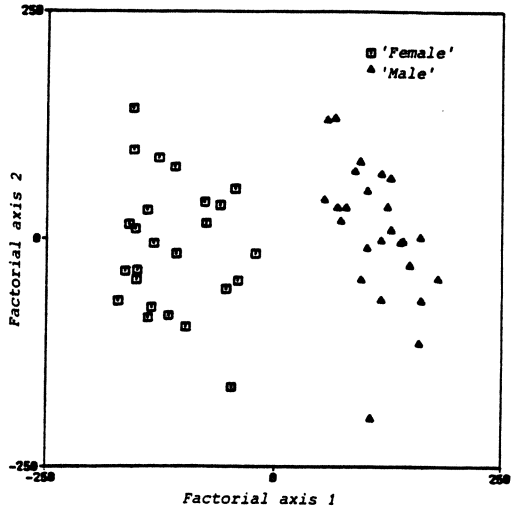

Fig.1 Projection of the french word "terminer" pronounced by 25 female and 25 male speakers onto the plane spanned by the first and the second factorial axes

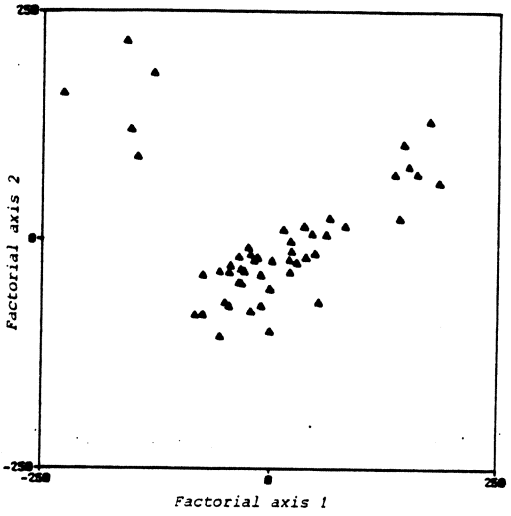

Fig.2 Projection of the french word "huit" pronounced by 50 male speakers onto the plane spanned by the first and the second factorial axes

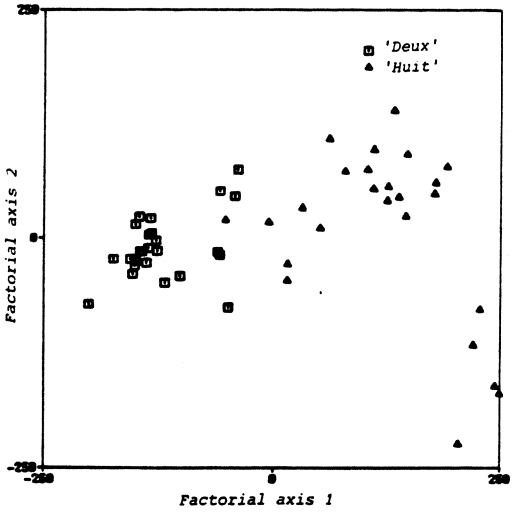

Fig.3 Projection of the french words "deux" and "huit" pronounced by 25 male speakers onto the plane spanned by the first and the second factorial axes

such that the distance between the samples i and j in the sub-space R^m are as near as possible to the original d(i,j).

Thus, this data analysis turns out to be a factor analysis problem /11/. Basic steps are as follows : from the distance matrix, we compute the variance-covariance matrix and find its eigenvalues. The eigenvectors corresponding to the m greatest eingenvalues define the factorial axes. We call t_m, the percentage of the total data variance, associated to a given factorial axis. More details concerning the method can be found in /10/, /11/.

Figure 1 shows the projection of the utterances of the french word "terminer" pronounced by 25 male and 25 female speakers onto R^2 (the plane spanned by the first and the second factorial axis). Male speakers are represented by a triangle and female speakers by a square. t_1 and t_2 are 29 and 11% respectively. The figure shows the clear existence of male and female clusters. It suggests the use of at least two templates to represent this particular word.

Figure 2 shows the projection onto R^2 of the utterances of the french word "huit" pronounced by 50 male speakers. t_1 and t_2 are 18 and 13% respectively. This example clearly speaks for the existence of clusters due to pronounciation differences.

Figure 3 shows the projection onto R^2 of the utterances obtained by pronounciations of the french words "deux" and "huit" pronounced by 25 male speakers. The word "deux" is represented by a square and the word "huit" by a triangle. t_1 and t_2 are 33 and 9% respectively. The purpose of this figure is to visualise the consequence on classication of the existence of clusters. It clearly shows that using the property of utterances being clustered will help to discriminate the words.

In summary, using factor analysis to visualize utterances of isolated words by different speakers, we showed the existence of clusters. Most obvious are the existence of man and women clusters which reflect anatomy differences. Next, other clusters were showed which reflect pronounciation differences. Finally, we gave one example where using the existence of clusters will clearly improve word discrimination. Herewith we gave some insight into the nature of inter-speaker variation and justify the hypothesis made above.

4. Clustering procedures

4.1 Principle

Given C, the set of all the elements Xi, i=1,...,I (different pronounciations of the same word), find the few representatives Rk, k=1,...,m (multiple templates) which describe it.

The solution has two distinct steps : 1) the clustering itself which divides C in disjoint clusters Ck, k=1,...,m in such a way that a given criterion be fulfilled, 2) the choice of the representative Rk for each cluster Ck.

4.2 Criterion based Exchange procedure (CEx)

The CEx is a cluster-parallel iterative clustering procedure producing a fixed number m of clusters. It minimizes a criterion function F by iteratively transferring elements from cluster to cluster in such a way that F decreases.

1. Choose initial partition C1, C2, ... , Cm

2. a) Find Xi of C for which there exists a cluster Cl such that the transfer of Xi from its cluster Ck to cluster Cl decreases F :
   $\Delta F(Xi, Ck \rightarrow Cl) < 0$

   b) Stop if no such element Xi of C can be found
   $\forall\, Xi \in C,\ l=1,\ldots,m : \Delta F(Xi, Ck \rightarrow Cl) \geqslant 0$

3. Transfer Xi to the cluster Ck' which minimizes ΔF, i.e.
   $\Delta F(Xi, Ck \rightarrow Ck') = \underset{l}{\text{Min}}\ \Delta F(Xi, Ck \rightarrow Cl)$

4. Go to 2.

4.3 Criterion based Threshold procedure (CTh)

The CTh is a cluster-sequential clustering procedure that, as clusters are created, gradually removes the elements from C', the set of elements still to be clustered, until C' is empty. At each iteration a cluster Ck is created which fulfills the threshold condition, i.e., a cluster of elements Xj around a center element Xi whose distances d(Xi,Xj) do not exceed an a priori fixed threshold T.

The important point is that now, at each iteration, among all possible clusters A(Xi) fulfilling the threshold condition :

$$A(Xi) = \left\{ Xj \in C' \ /\ d(Xi,Xj) < T \right\}$$

the best is selected, i.e. the one that minimizes the criterion function H(A(Xi)) measuring the homogeneity in A(Xi). Note that, in this case, H applies to a sole cluster.

1. Initialization :
   k = 1 (k-th cluster)
   C' = C

2. For each $Xi \in C'$ find the candidate-cluster A(Xi) :
   $A(Xi) = \left\{ Xj \in C' \ /\ d(Xi,Xj) < T \right\}$

3. Find the candidate-cluster minimizing the criterion function :
   $Ck = A(Xi^*) \ /\ H(A(Xi^*)) \leqslant H(A(Xi)) \quad \forall\ Xi \in C'$

4. C' = C' - Ck

5. If C' ≠ Ø then : k = k+1 and Go to 2.
   Else : Stop

With CTh, the number of clusters created is variable and depends on T

4.4 Criterion function

4.4.1 Definitions

To an element Xi of cluster Ck we associate the following general metrics :

$$L_q(Xi,Ck) = \left(\frac{1}{n_k - 1} \sum_{Xj \in Ck} d(Xi,Xj)^q \right)^{1/q}$$

Note that for q=1, we obtain the mean of distances between all Xj and Xi. For q=2, we obtain the rms value of these distances. For $q = \infty$, we obtain the maximum distance.

From $L_q(Xi,Ck)$ several metrics may be derived which measure the homogeneity of a cluster. These metrics will be used to define criterion functions for CEx and CTh procedures. Let us define the following homogeneity function :

$$M_q(Ck) = \min_{Xi \in Ck} L_q(Xi,Ck)$$

4.4.2 Criterion function for CEx

Clustering procedures defined above minimize a criterion function which is supposed to measure the quality of a partition. The real world problem is to find which criterion function really measures the partition quality in SISR. Several criterion functions were defined and tested. The results published elsewhere /9/, have shown the existence of different classes of criterion functions with different behaviour and recognition performance. The following class of criterion function behaved well :

$$F_q = \sum_k M_q(Ck).(n_k-1)$$

where n_k is the number of elements in Ck

3.5.3 Criterion function for CTh

The criterion functions for CTh procedures is defined for only one cluster. Here also, various criterion functions based on the homogeneity functions and the number of elements in the candidate-cluster A were defined and tested /9/. The results have shown that the criterion functions based only on the homogeneity functions behaved better. We use the following criterion function :

$$H = M_1(A)$$

4.5 Representative of a cluster (template)

The representative R(Ck) of a cluster Ck is chosen as follows : R(Ck) is the element Xi* of Ck that minimizes the metric $L_1(Xi,Ck)$ in the cluster Ck :

$R(Ck) = Xi^* \in Ck$ such that $M_1(Ck) = L_1(Xi^*,Ck)$

4.6 Outlier elimination

Outliers, i.e., isolated elements, may disturb the clustering. Therefore, before clustering is applied, outliers must be detected and discarded. The detection is based on the distance of the element to its nearest neighbour.
For each of the N elements of a given word, we compute the minimal distance to the N-1 remaining elements :

$$Dmin(i) = \min_{\substack{i=1,\ldots,N \\ i \neq j}} d(i,j)$$

An element is discarded if

$$Dmin(i) > \overline{Dmin} + s$$

where $\overline{Dmin}$ is the mean value and s the standard deviation of the distances Dmin(i).

4.7 Variable number of templates per word

In a given vocabulary, there exist words which are easy to recognize and others which are difficult to recognize. If the total number of reference templates for that vocabulary is fixed, it seems reasonable, a priori, to assign a larger number of templates for words which are difficult to recognize than the number of templates assigned for words which are easy to recognize.

This is illustrated by the word confusion matrix given in the figure 4. This matrix lists the result of recognition tests where the utterances of each speaker are tested against those of all other speakers.

Indeed, it may be observed that the number of confusion errors is highly variable. It is minimum for the word 'six' and maximum for the word 'cinq'. The number of clusters per word is then set to :

$$c(m) = (e(m)/e).(c-1) + 1$$

where e(m) is the error rate of the m-th word, e the global error rate and c the averaged number of reference templates per word.

		Reference									
		0	1	2	3	4	5	6	7	8	9
T	0	918	36	206	18	0	12	1	33	1	0
	1	31	744	8	385	11	38	1	4	3	0
e	2	170	32	940	3	1	12	2	28	32	5
	3	27	110	6	1057	8	4	0	12	0	1
s	4	0	13	0	30	896	65	52	117	3	49
	5	15	41	7	23	28	585	177	341	3	5
t	6	0	0	0	0	0	45	1151	22	6	1
	7	13	0	4	2	11	152	413	625	4	1
	8	11	0	49	0	15	23	274	23	665	165
	9	2	0	15	0	75	12	80	17	119	905

Fig.4 Word confusion matrix for the french digits. The speaker reference and the speaker test are different (25 male and 25 female speakers)

5. Recognition results

Open tests were performed with an isolated word recognizer using a 13-word french vocabulary (three control words : en-avant, en-arrière, terminer, and the ten digits : zéro, un, ..., neuf). The training data set consisted of one repetition of each word by 25 male and 25 female speakers, while another set of three repetitions of each word by 5 male and 5 female speakers was used for recognition tests.

Figure 5 gives the recognition performance using multiple templates per word. On one hand, the templates are chosen arbitrarly, while on other hand, the templates are selected by the CEx algorithm using different criterion functions. The superiority of the clustering algorithm CEx with respect to a random choice of template references may be observed clearly. The strong decrease of error rate for a small number of clusters (small compared to the number of speakers in the training set) proves the success of the CEx clustering procedure applied to SISR.

Figure 6 compares CEx and CTh, the two clustering algorithms presented here with the UWA, the clustering algorithm previously used in SISR. CEx was used with the F_1 criterion, CTh with H as defined in 4.5.3, UWA as described in /4/. In the case of CTh and UWA algorithms, the threshold, for each word, is chosen iteratively in such a manner that we obtain a number of clusters not exceeding m (m=6,5,4,...,1). We can observe the superiority of CEx and CTh with respect to UWA algorithm in our tests.

Figure 7 compares the recognition error rate obtained by the CTh algorithm in the following three cases : 1) CTh algorithm without rejection of outliers and with the same number

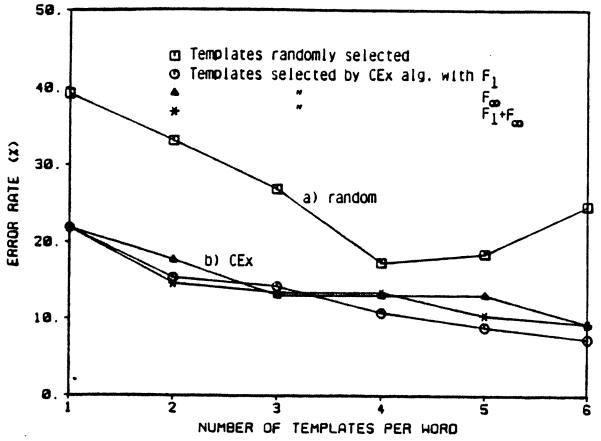

Fig.5 Recognition performance of the multiple templates approach using clustering (CEx) and not using clustering (random)

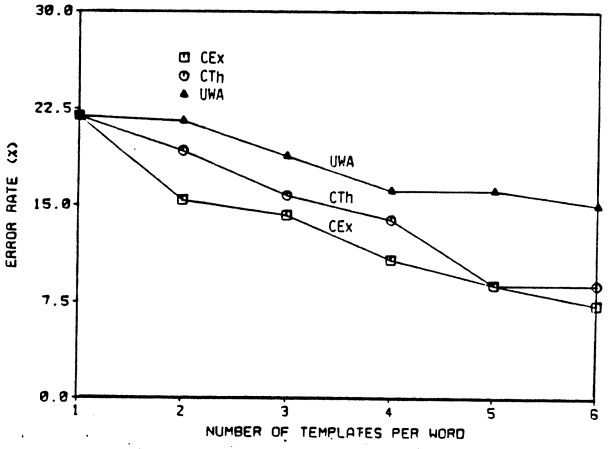

Fig.6 Comparison of the three clustering algorithms

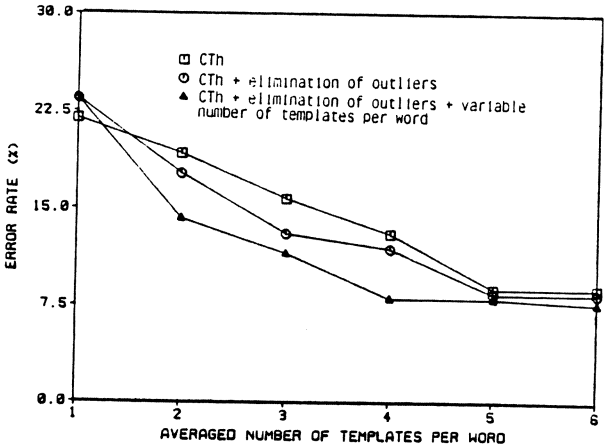

Fig.7 Effect of the outlier elimination and the variable number of templates per word

of cluster per word, 2) CTh algorithm with rejection of outliers and with the same number of cluster per word, 3) CTh algorithm with elimination of outliers and with a variable number of templates per word. It may be observed that the error rate is improved in both the second and the third case.

In summary, three performance figures quantitatively measure :
a) the general advantage of the clustering technique.
b) the advantage of CEx and CTh algorithms against UWA algorithm.
c) the advantage of using outlier elimination and a variable number of templates per word.

6. Conclusion

In this paper, we have shown by means of factor analysis that the distribution of isolated word utterances pronounced by a large number of speakers can be described by a limited number of clusters. With these results we fully justify the use of the multiple template reference approach to SISR. Then, we presented the two criterion based clustering procedure CEx and CTh and tested their SISR performance. Open tests conducted with an isolated word recognizer have shown the general advantage of the clustering technique, the advantage of CEx and CTh algorithms over the UWA algorithm often used in SISR and also, the advantage of both outlier elimination and the use of a variable number of templates per word.

Acknowledgements

This work was supported by the the 'Commission pour l'Encouragement des Recherches Scientifiques' (CERS n0 1158, Bern, Switzerland) and the following companies : ASULAB S.A., CEH S.A., METTLER S.A., HASLER S.A., AUTOPHON S.A. and CIR S.A..
We wish to thank the speakers who participated in generating our speech data base.

References

/1/ H. Sakoe, S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans on ASSP, Vol. ASSP-26 NO. 1, pp. 43-49, Feb. 1978

/2/ Ngoc C. Bui, Jean J. Monbaron, and Jean G. Michel, "An Integrad Voice Recognition System', IEEE Trans. on ASSP, Vol.ASSP-31, NO. 1, February 1983.

/3/ L.R. Rabiner, "On Creating References for Speaker Independent Recognition of Isolated Words", IEEE Trans on ASSP, Vol. ASSP, No.3, pp.34-42, Feb. 1978.

/4/ L.R. Rabiner, J.G Wilpson, "Considerations in Applying Techniques to Speaker-independent Word Recognition", J. Acoust. Soc. Am., Vol. 66, NO. 3, September 1979.

/5/ Niles, Les, Harvey F. Silverman., N. Rex Dixon, "A Comparison of Three Feature Vector Clustering Procedures in a Speech Recognition Paradigm ". Proc. ICASSP 83, pp.765-768, 1983.

/6/ B. Flocon and P. Lockwood, "A Speaker Independent Isolated Word Recognition System", EUSIPCO-83, pp. 407-410.

/7/ S.E. Levinson, L.R. Rabiner, A.E. Rosenberg, and J.G. Wilpson, "Interactive Clustering Techniques for Selecting Speaker-Independent Reference Templates for Isolated Word Recognition", IEEE Trans on ASSP, Vol. ASSP-27, No.2,April 1979.

/8/ G.H. Ball and D.J. Hall, "Isodata-An Iterative Method of Multivariate Analysis and Pattern Classification,"in Proc.IFIPS Congr.,1965.

/9/ A.Mokeddem, H.Hugli, F.Pellandini, "Evaluation of criterion based clustering procedures for generating multiple template references in speaker independent speech regonition", 7th ICPR August 1984, Montreal

/10/ A.Mokeddem, "Analyse factorielle appliquée aux échantillons multilocuteurs de la parole". Rapport interne IMT Neuchâtel.

/11/ Diday E., Lemaire J., Pouget J., Testu F., "Eléments d'analyse de données", Dunod, Paris, 1982

/12/ J.P. Haton, "Automatic Speech Analysis and Recognition", 1982, D.Reidel Publishing Compagny.