EVALUATION OF CRITERION BASED CLUSTERING PROCEDURES FOR GENERATING MULTIPLE TEMPLATE REFERENCES IN SPEAKER INDEPENDENT SPEECH RECOGNITION

A. Mokeddem - H. Hugli - F. Pellendini

Reprinted from IEEE SEVENTH INTERNATIONAL
CONFERENCE ON PATTERN RECOGNITION Proceedings
July 30 - August 2, 1984 Montreal, Canada

EVALUATION OF CRITERION BASED CLUSTERING PROCEDURES FOR GENERATING MULTIPLE TEMPLATE REFERENCES IN SPEAKER INDEPENDENT SPEECH RECOGNITION

A. Mokeddem - H. Hügli - F. Pellandini

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

Abstract - Speaker independent speech recognition can be done by creating multiple template references. In this paper, we present the new application to this field of two criterion based clustering procedures, namely : the Criterion based Exchange algorithm (CEx) and the Criterion based Threshold algorithm (CTh). After the theoretical description, we give the results of a series of evaluation tests performed with an isolated word recognizer :
1) analysis of criterion functions in the case of both CEx and CTh,
2) comparison of methods for choosing the cluster representative,
3) comparison of CEx and CTh with the UWA algorithm previously used in other works.

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 a same utterance among different speakers, the so-called inter-speaker variation. To solve it, the multiple template reference approach turned out to be succesful [4],[7]. In this approach, one tries to describe all different pronounciations 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 utterance they represent, multiple template references .

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

In this paper we propose and study the new 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 extremizes a criterion function. The fact that each cluster is a possible candidate makes this a general version of the UWA clustering algorithm previously used.

2. RECOGNIZER

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

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 applied. 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 420 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. CLUSTERING PROCEDURES

3.1 Principle

Given C, the set of all the elements Xi , $i=1,\ldots,I$ (different pronounciations of the same utterance), find the few representatives Rk , $k=1,\ldots,m$ (multiple templates) which describe it. The solution has two distinct steps : 1) the clustering itself which divides C in disjoint clusters Ck, $k=1,\ldots,m$ in such a way that a given criterion be fulfilled, 2) the choice of the representative Rk for each cluster Ck.

3.2 Criterion based Exchange procedure (CEx)

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

1. Choose any 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 :
   $\Delta F(Xi, Ck \longrightarrow Cl) < 0$

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

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

4. Go to 2.

3.3 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 all not exceed an a priori fixed threshold T.

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

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

the best is selected, i.e. the one that extremizes 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)
   $A(Xi) = \{ Xj \in C' \ / \ d(Xi,Xj) < T \}$

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

4. C' = C' - Ck

5. If $C' \neq \emptyset$: k = k+1 and Go to 2.
   Else : Stop

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

3.5 Criterion function

3.5.1 Definitions

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

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

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 are derived which measure the homogeneity of a cluster. These metrics will be used to define criterion functions for CEx and CTh procedures.

$$M_{1q}(Ck) = \underset{Xi \in Ck}{\text{Min}} \ L_q(Xi,Ck)$$

$$M_{2q}(Ck) = \frac{1}{n_k} \sum_{Xi \in Ck} L_q(Xi,Ck)$$

$$M_{3q}(Ck) = \underset{Xi \in Ck}{\text{Max}} \ L_q(Xi,Ck)$$

3.5.2 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. Various criterion functions will be considered here. Criterion functions are now defined as follows :

$$FA_{pq} = \sum_k M_{pq}(Ck)$$

$$FB_{pq} = \sum_k M_{pq}(Ck).\ (n_k - 1)$$

$$FC_{pq} = \sum_k M_{pq}(Ck).\ n_k .(n_k - 1)$$

where n_k is the number of elements in Ck.

3.5.3 Criterion function for CTh

The following criterion functions are defined for one cluster based on homogeneity functions M_{11}, $M_{1\infty}$, M_{21} (3.5.1.) and the number of elements n_A in the candidate-cluster A :

$H1 = M_{11}(A) + M_{1\infty}(A)$
$H2 = (M_{11}(A) + M_{1\infty}(A)) \ / \ n_A$
$H3 = 1 \ / \ n_A$
$H4 = M_{11}(A)$
$H5 = M_{1\infty}(A)$
$H6 = M_{21}(A)$

H1, H2 and H3 were chosen to find out the separate or combined effects of the homogeneity functions (which should be minimized) on one hand, and the number of elements in A (which should be maximized), on the other hand.

H1, H4, H5, H6 were chosen for comparing various criterion functions that are identical to the homogeneity functions (3.5.1.).

3.6 Representative of a cluster

Different methods are considered for choosing the representative R(Ck) of a cluster Ck :

a) mean: R(Ck) is the mean of the elements in the cluster Ck.

b) minimean : R(Ck) is the element Xi* of Ck that minimizes the metric $L_1(Xi,Ck)$ in the cluster Ck :

$$R(Ck) = Xi^* \in Ck \text{ such that } M_{11}(Ck) = L_1(Xi^*,Ck)$$

c) minimax : R(Ck) is the element Xi* of Ck that minimizes the metric $L_\infty(Xi,Ck)$ in the cluster Ck :

$$R(Ck) = Xi^* \in Ck \text{ such that } M_{1\infty}(Ck) = L_\infty(Xi^*,Ck)$$

4. EXPERIMENTS AND RESULTS

4.1 Criterion functions for CEx

For studying the behaviour of various criterion functions in speech clustering, the following 'correct clustering' experiment is conducted :

EXP1 : 10 speakers (5 men and 5 women) pronounce 6 times the chosen vocabulary, providing 60 elements for each word. Now, starting from 10 initial clusters chosen at random, we determine for each of the criterion functions the partition given by the CEx procedure. If the criterion function is well adapted and if the intra-speaker distances are smaller than the inter-speaker distances, one obtains, in the same cluster, all six elements belonging to the same speaker.

Firstly we conducted experiments permitting to confirm that the final partition obtained is only very weakly sensitive to the choice of the initial partition.

4.1.2 Correct clustering rate

In order to compare the different criterion functions, we define, for the experiment EXP1, a correct clustering rate Tc which measures the partition correctness.

Let C1, C2, ... , Cm be the partition given by CEx and Vmax(Ck) the maximum number of elements that belong to a same speaker in the cluster Ck. One can say that the cluster Ck belongs to the this speaker. The correct clustering rate is defined as follows :

$$Tc = \left(\sum_k Vmax(Ck) \right) / I$$

where I is the total number of elements in C.

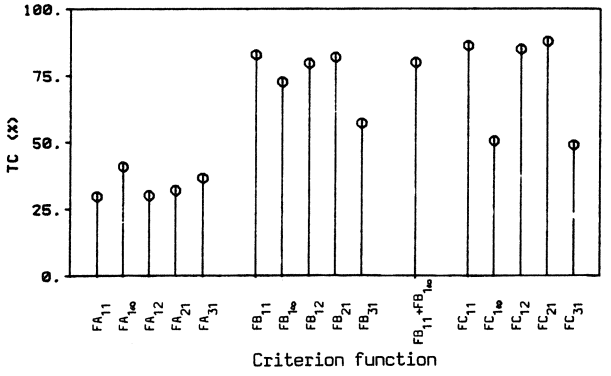

Fig. 1. Tc for the various criterion functions (CEx procedure)

Fig. 1. gives the measured correct clustering rates Tc, averaged over the vocabulary for the different criterion functions in the case of CEx.

The results show clearly that the choice of the criterion function is very important. Again, we find the bad types as FA_{pq}. Good types are FB_{pq} and FC_{pq}.

Of the bad type are all criteria (as FA_{pq}) based on metrics which do not somehow increase with an increased number of elements per cluster. For these criteria, the algorithm tends to put very few elements in certain clusters, and the rest in one big cluster.

4.2 Criterion functions for CTh

EXP2 : 10 speakers (5 men and 5 women) pronounce 6 times the chosen vocabulary, providing 60 elements for each word. Now, we determine for each of the criterion functions (H1,...,H6) the partition given by the CTh procedure.

The number of clusters found is not necessarily equal to the number of speakers. The clusters obtained are arranged in the following manner : $n_1 > n_2 > \dots > n_m$, where n_k is the number of elements in Ck. The rate Tc is calculated as follows :

$$Tc = \sum_{k=1}^{k'} Vmax(Ck) / I \qquad \text{with : } k' = Min(L,m)$$

where L is the number of speakers and I the total number of elements in C.

The optimal partition i.e., the one corresponding to Tcmax is found by varying the threshold by small steps. The table 1 gives for every criterion function the measured mean and the standard deviation of Tcmax.

Based on table 1, we can conclude that functions H1, H4, H5 and H6 based uniquely on homogeneity functions give similar result. They are slightly better than H2 and H3 which show practically the same results due the prevailing effect of n_A in H2 and H3. In the last column, we give the results obtained for UWA procedures.

	H1	H2	H3	H4	H5	H6	-	UWA
Tcmax	79.5	73.3	73.3	79.2	78.6	79.1	-	73.5
σ	8.8	9.9	9.9	8.2	9.8	8.8	-	9.9

Table 3. : Mean and standard deviation of Tcmax for various CTh and the UWA procedures.

4.3 Recognition results

Two kind of recognition tests were performed with a 13-word vocabulary.

TST1 : Recognition test where test speakers did not participate to the references creation. For creating the reference templates we used a set of 4 repetitions per word by 15 speakers (total : 60 elements). For recognition tests we used 9 repetitions per word by 5 different speakers (total : 585 tests)

TST2 : Recognition test where the same speakers but different repetitions were used for training and test. For creating the reference templates we used a set of 3 repetitions per word by 15 men and 5 women (total : 60 elements). For recognition tests we used 6 different repetitions per word by the same speakers (total:1560 tests).

Figures 2a) and 2b) illustrate recognition results as a function of the number of clusters, obtained by the CEx algorithm for the criterion functions FB_{11}, $FB_{1\infty}$, $FB_{11}+FB_{1\infty}$ and FC_{21}. The representative of each cluster is determined by the minimax metric. The strong decrease of error rate for a small number of clusters (small compared to the number of speakers in the training set which is 15 resp. 20) proves the success of these methods applied to SISR.

Figure 3 shows the recognition error rate obtained by CEx for three manners of choosing the representative of clusters. The results show the general superiority of mean over minimean and minimax and also, the superiority of minimean over minimax when the number of clusters is small. As a conclusion, mean is the most favourable method. Sometimes however, minimean may be preferred because it requires much less computation.

Figure 4 compares the recognition error rate for the three algorithms CEx, CTh and UWA. CEx was used with the $FB_{11}+FB_1$ criterion; CTh with H1 ; UWA, as described under 3.4. In the case of CTh and UWA algorithms, the threshold, for each word, is chosen in such a manner that we obtain a number of clusters not exceeding m (m=5,4,...,1). We can observe the superiority of CEx and CTh with respect to UWA algorithm in our tests.

In summary, our tests showed the good performances of both criterion based clustering procedures. Although these results cannot be generalized without some care, there are hopes that they will also be useful in speech recognizers different than the one used.

5. CONCLUSION

We have studied two criterion based clustering procedures for creating multiple template references for speaker independent speech recognition (SISR). The tests conducted with an isolated word recognizer have shown :

1. The importance of choosing adequate criterion. The criterion functions FB_{pq} and FC_{pq} behaved well for CEx.
2. The superiority of mean over minimean and superiority of minimean over minimax for the choice of cluster representatives
3. The superiority of CEx and CTh algorithms over UWA algorithm.

These performances clearly show the advantages of criterion based clustering algorithms and lead to further applications in SISR.

ACKNOWLEDGEMENTS

This work was supported by the the "Commission pour l'Encouragement des Recherches Scientifiques" (CERS nO 1158, Bern, Switzerland) and the following companies : ASULAB S.A., AUTOPHON A.G., CEH S.A., CIR S.A., HASLER A.G. and METTLER A.G..

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 Tech niques 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.

a)

b)

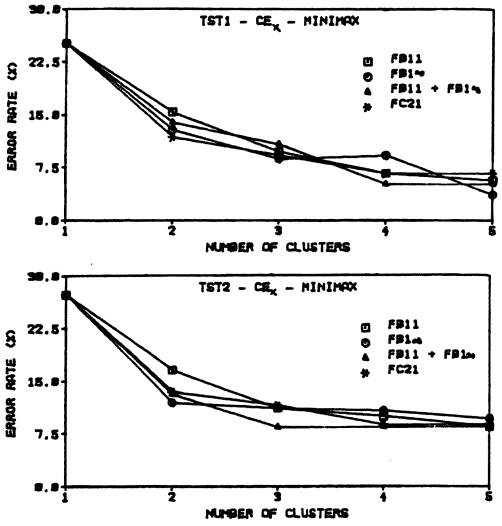

Fig. 2. Error rate as function of the number of clusters. (Algorithm : CEx , cluster representative : minimax) a) Test : TST1 , b) Test : TST2

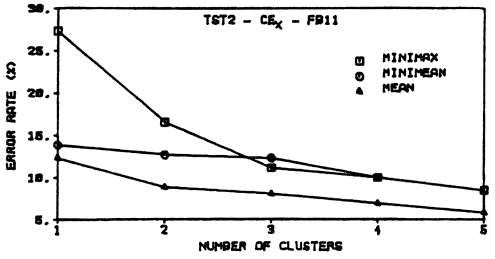

Fig. 3. Error rate for different methods of choosing the cluster representative. (Test : TST2 , algorithm : CEx with FB_{11})

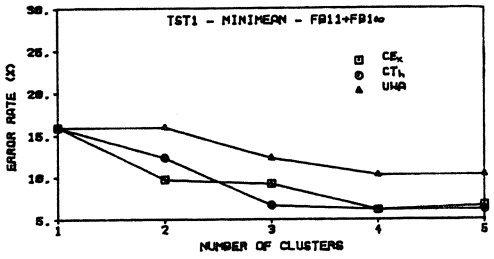

Fig. 4. Error rate for the three algorithms CEx with $FB_{11}+FB_1$, CTh with H1 and UWA. (Test : TST1 , cluster representative : minimean)