DYNAMIC TIME WARPING AND VECTOR QUANTIZATION IN ISOLATED AND CONNECTED WORD RECOGNITION.
A. Boyer*, J.-P. Haton*, J. di Martino *

ABSTRACT

The first part of this paper describes an algorithm using dynamic programming which allows endpoint relaxation and then achieves an implicit segmentation of the pattern to be compared.

In the second part, we introduce vector quantization in order to reduce the memory size occupied by data (one or several patterns for each word of the vocabulary). We propose several recognition methods using dynamic time warping and we compare their performances.

In the last part, we extend the Bridle and Nakagawa algorithm by using endpoint relaxation, syntactic constraints, vector quantization and we propose a method which takes into account liaisons and coarticulation effects at the boundaries of connected words.

INTRODUCTION

Devices for isolated and connected word recognition with restricted vocabulary use largely dynamic programming since that makes it possible to compensate for the essentially non linear distortions due to variations in the speaker’s utterance.

However, dynamic programming requires to memorize at least an acoustic image of each word of the dictionary. Vector quantization enables to reduce the memory size needed for storage. Some algorithms using dynamic time warping together with vector quantization are proposed.

DYNAMIC PROGRAMMING AND ENDPOINT RELAXATION

Let $T=t(1)\ldots t(l)$ and $R=r(1)\ldots r(J)$ the two patterns where :
$t(i)$ and $r(i)$ are two vectors coding the speech signal at the $i$th unit; they are given by classical parametrization such as LPC, filter bank, etc...
$I$ and $J$ are the sample lengths of pattern $T$ and $R$ respectively.

The aim of dynamic programming is to find the best warping path between both patterns. It can be obtained through iterative assessment of the following relations:

$$D(i,j) = \min \{ D(i-1,j-1) + 2 \cdot d(i,j),
D(i,j-1) + d(i,j),
D(i-1,j) + d(i,j) \}$$

(1)

where $D(i,j)$ is the cumulated distance related to the optimum partial path to the point $(i,j)$, and $d(i,j)$ is the local distance between both vectors $i$ and $j$. $D(I,J)$ is the dissimilarity rate between both patterns.

In order to compensate for the errors due to speech- non speech detection, it is worthwhile to release boundary constraints that is to i.e. not to have to

* CRIN/INRIA LORRAINE BP 239 54506 VANDOEUVRE LES NANCY FRANCE
start the warping path from point (1,1) and to end at (I,J), but in the neighbourhood of those points.

Relation (1) is no longer valid since the assumption of constant warping path length is not fulfilled.

New relations of dynamic programming must be defined, which normalizes in length the cumulated distances:

\[ d_1 = D(i-1,j) + d(i,j) + l(i-1,j) + 1 \]
\[ d_2 = D(i-1,j-1) + 2 \cdot d(i,j) + l(i-1,j-1) + 2 \]
\[ d_3 = D(i,j-1) + d(i,j) + l(i,j-1) + 1 \]
\[ d = \min(d_1, d_2, d_3) \]

if \( d = d_1 \) then
\[ D(i,j) = D(i-1,j) + d(i,j) \]
\[ L(i,j) = L(i-1,j) + 1 \]

if \( d = d_2 \) then
\[ D(i,j) = D(i-1,j-1) + 2 \cdot d(i,j) \]
\[ L(i,j) = L(i-1,j-1) + 2 \]

if \( d = d_3 \) then
\[ D(i,j) = D(i,j-1) + d(i,j) \]
\[ L(i,j) = L(i,j-1) + 1 \]

where \( l(i,j) \) is the length of the optimum warping path leading to the point \( (i,j) \).

**VECTOR QUANTIZATION AND DYNAMIC PROGRAMMING**

The vocabulary that we use consists of the ten French digits. The training corpus is composed of two repetitions of each digit by twenty speakers (ten males, ten females). Most of them are untrained speakers. To test the algorithms, we use a corpus which includes two other repetitions of each digit by the same speakers together with four repetitions by fifteen new speakers. The speech signal is 16 kHz sampled and with a linear 32 filter banks analyzed (0-8000 Hz frequency range). All algorithms described in the paper are written in C and implemented on a Masscomp mini-computer.

We build up a codebook for each word of the vocabulary by using the threshold algorithm (ref 1). Our problem is to recognize multi-speaker isolated words after their coding by vector quantization.

We have tested the method proposed by Burton (ref 2) and obtain 85.1% as recognition rate for the speakers who took part in the training (first group) and 81.5% for the others (second group). This rather poor rate can be explained by the fact that no temporal information has been taken into account. We notice that the correct answer always fits one of the three codebooks which give the smallest distortion.

We therefore use the following method:

1. Selection by minimizing coding distortion of the three best candidates to recognition,
2. Comparison by dynamic programming with those three reference patterns,
3. Decision by minimizing the dissimilarity distance DTW(i).

The recognition rate thus obtained is 86.4% for the first group of speakers and 82.8% for the second one. This still unsatisfactory result is due to the fact that coding distortion has been neglected. The method was then improved by modifying the decision criteria in minimizing the sum of the coding distortion and the dissimilarity rate.

We introduce in the criteria a coefficient \( \beta \) which multiplies DTW(i) and \( (1-\beta) \) multiplies the coding distortion. We tried several values of \( \beta \) between 0 and 1. The optimum for recognition rate is achieved for \( \beta = 0.5 \). Then we obtain 89.2% of good recognition for the
first group and 84.1% for the second one, which is a real improvement of the previous method.

If we authorize endpoint relaxation at the boundaries of the word, the recognition rate increases a bit to reach 89.7% for the first group and 85.2% for the second one.

If we consider the coding reference patterns, we notice that the code-vectors can appear only at particular places in the word. This remark lead Burton (ref 6) to introduce multi-section codebooks and Rabiner (ref 8) to take into account the probability of occurrence of a vector in a particular place in the word. This latter method necessitates to normalize the pattern in length before recognition. We propose to divide the word in n parts and to determine during the training the frequency of occurrence of a code-vector in a given part of the word. Then we do not need a length normalization before recognition. We determine the coding distorsion by summing the distance between vectors and the code-vector and the inverse of its frequency of occurrence. The optimum of the recognition rate was obtained for n=15: 91.9% for the first group and 88.4% for the second one.

The gain in memory size (75% in our application) does not affect the recognition rate which is still satisfactory, particularly if some improvements are introduced in the dynamic time warping algorithm such as to take into account both dissimilarity distance between two forms and the coding distorsion. The compuation time is about the same for all algorithms we have presented.

CONNECTED WORD RECOGNITION

We will now describe an experimentation in single speaker connected word recognition using dynamic programming. The vocabulary consists of the ten french digits. Five speakers have pronounced each digit once as reference patterns. During the recognition, the speakers pronounced sequences of one, two, three and four connected digits. We also introduced a reference for silence in order to allow speakers to introduce pauses between words.

The Bridle and Nakagawa algorithm (ref 3, ref 4) is a very powerful algorithm because it is a one pass algorithm. It uses a generalization of relations (1) and searches the warping path between the input signal and all the sequences of reference patterns in a three dimensional space.

We have tested it on the corpus described previously and we obtain about 82% of recognition. This poor recognition rate is partly explained by the existence of paths with long horizontal and vertical parts. So we introduced in the algorithm weighting coefficients to penalize the too long horizontal and vertical paths. Furthermore, we take into account another coefficient which increases the distance if we compare a vector issued from the silence reference with any other vector and if the local distance between both is greater than an arbitrary threshold. If these three coefficients are equal to the average distance between two vectors of reference patterns, the average recognition rate is 90.7%.

We have extended the Bridle and Nakagawa algorithm by allowing endpoint relaxation at the boundaries between words. If we use the formalism introduced by Ney (ref 5) who defines two types of local neighbourhood (the within template which appears in a word and the between templates which occurs at the boundary of two words), it is only an increase of the between templates neighbourhood which includes the last points of all reference patterns.
and not only the last one. The recognition rate reaches then 91.5%. The increase is more significant than in isolated word recognition because the endpoint relaxation can compensate for some coarticulation effects.

The use of syntactic constraints makes it possible to significantly decrease the number of comparisons during the recognition process. The introduction of such constraints leads also to a modification of the between templates neighbourhood that has to be restricted for each word to the words which can appear before (ref7).

We have developed an algorithm with syntactic constraints which can be used to take into account liaisons and coarticulation effects at the boundaries of connected words if the grammatical rules are phonological or liaisons rules. For a given vocabulary, the syntactic network can be built automatically (ref8).

The problem of the memory size occupied by data is the same as in isolated word recognition. Therefore we tried to introduce vector quantization in the dynamic time warping algorithm. We consider as local distance in the algorithm the sum of the distance between the vector and the code vector and the probability of occurrence of the code vector. The algorithm has been tested on the same corpus as before and we obtained 86% of recognition.

Thus, the introduction of vector quantization decreases significantly the recognition rate and some improvements still remain to be done.

CONCLUSION

To use vector quantization as a first step in the recognition process is a good means to reduce the memory size occupied by data. In isolated word recognition, it does not decrease the recognition rate if we introduce some modifications in the criteria decision and it allows for multi speaker recognition.

Syntactic constraints are introduced in the Bridle and Nagagawa algorithm to decrease the number of comparisons during the recognition process. The same algorithm can be used to account for liaisons and coarticulation effects at the boundaries of connected words. The endpoint relaxation is also a good means to take into account some coarticulation effects and the experimental results show a real improvement of the recognition rate.

ref4: S. Nakagawa, Proc. ICCASP, pp 296-299, Boston, 1983
ref6 : A. Boyer, thèse de l'université de Nancy 1 en informatique, 1987