SIGNAL PROCESSING III: Theories and Applications
I.T. Young et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

FRAME REDUCTION IN LPC ISOLATED WORD RECOGNITION

Climent Nadeu, Eduardo Lleida and José B. Mariño.

E.T.S.E. Telecomunicació. U.P.C.
C/ Jordi Girona Salgado, s/n. 08034 Barcelona - Spain.

ABSTRACT. It is well know that the pattern matching approach to word-based recognition yields good results when the time alignment is carried out by means of dynamic programming. However, the high computational expenditure needed by it, specially for large vocabularies, has motivated a number of works proposing the use of techniques that reduce the pattern lenght.

In this paper, three different frame or spectrum reduction methods are investigated in a LPC isolated word recognition framework. The reduction performance with different degrees of nonlinearity is tested, and reasons showing its relationship with some kind of actual endpoint detection innacuracies are given.

1. INTRODUCTION

In the isolated or connected word recognition systems based on pattern matching [1] the time alignment procedure plays a central part as it compensates for differences in speaking rate between two utterances of the same word. In this task, the dynamic programming method has proven to be more efficient than a simple linear time alignment [2], due to the nonlinear nature of the speech rate variations in a word. However, the computational expenditure can be much higher because it grows proportionally to the square of the pattern lenght.

This fact has motivated various papers ([3-10] are among the more recent) that propose methods of carrying out a nonlinear time normalization before the dynamic programming is applied. All of them start from a sequence of spectra, each one corresponding to a frame, and then produce a redistribution of these spectra along the time axis. The aim is a reduction of their number in order to lessen the amount of storage required by the references and, as a consequence, the recognition speed.

All these data compression techniques have in common the fact of exploiting the similarity between spectra belonging to a stationary segment of speech signals, so that one or few spectra can be representative of the complete segment. Since speaking rate greatly affects these stationary portions, such techniques perform a sort of time normalization. Thus, the dynamic programming procedure can search the warping path within a more restricted area, so that an additional reduction in the computational time required by the alignment process is obtained.

Furthermore, some reported techniques not only discard frames or spectra but they also insert additional ones in the transitional portions, where consecutive spectral shapes are more separated. Both factors emphasize transitions in front of stationarities. However, is this always suitable?. Obviously, it must be useful, for instance, when differences between words lye in plosive consonants [9]. Anyway, the degree of this emphasys (nonlinearity) should be different for each vocabulary.

In this paper, three different techniques are described and tested within the framework of an isolated word recognition system, where spectra are estimated through the autocorrelation method of linear prediction coding [11]. By using several degrees of nonlinearity, it is shown that a linear or quasi-linear normalization to a fixed number of spectra per utterance may be the best choice depending on certain factors among which there are the kind of vocabulary, the speaker and the accuracy of the endpoint determination.

2. THE REDUCTION TECHNIQUES

Two of the methods are of fixed lenght class, i.e. they aim to transform the initial sequence, which has a variable lenght or number of spectra, into a new one of lenght N, being N the same for all utterances. The first and last spectra are equal for both sequences and the N-2 remaining spectra of the new sequence are found with different procedures.

The final objective of the reduction techniques is to obtain a more uniform distribution of distances between consecutive spectra of the sequence. However, the result depends on the way the distance D(n,m) between two non-consecutive spectra is defined. If d(i,j) is the Itakura's distance between spectra i and j, we allow for a cumulative distance (CD)measure

$$D(n,m) = \sum_{k=n}^{m-1} d(k,k+1)$$

and a real distance (RD) measure D(n,m)=d(n,m).

Then, the algorithm goes from the first to the last frame of the initial sequence acquiring a spectrum for the new sequence each time that D(n,m) exceeds a threshold. For the CD method, the threshold is set to D/(N-1), where D is the total cumulative distance of the initial sequence. The threshold for the RD method that obtains a lenght N is not previously known; an iterative procedure is used in this case.

We also take into account a variable lenght method that uses the same threshold for all utterances and the real distance. We will refer to it as the fixed threshold (FT) method. This technique was considered by some authors [4-8,10]; it gives good results in connected-word recognition [10,5], where the fixed lenght techniques have less meaning. Several versions of the CD method were already considered as well [3,6-10], obtaining the highest scores in isolated word recognition.

To be able to perform tests with different degrees of nonlinearity, we will change the distance d(i,j) by $d^r(i,j)$ so, when r<1, the transitions are less favoured with respect to the stationary parts than when r=1. We particularly mention that, in the CD method, r=0 means a linear lenght normalization.

Interpolation may be a suitable possibility to compute each spectrum of the reduced sequence from the spectra of two consecutive frames, specially in the transitions and when r is high. The use of linear interpolations of the autocorrelations proved to give good results for all the methods [12].

3.- EXPERIMENTAL RESULTS

Speech analysis and data base.

For testing, a speaker dependent LPC isolated word recognition system like the classical one from Itakura [11] was used. The speech was recorded in a quite room, sampled at 8KHz., pre-emphasized and analyzed every 15 ms with a 30 ms Hamming window, LPC spectra were estimated through the autocorrelation method and order 8.

1	/u/	6	/sis/
2	/dos/	7	/sɛt/
3	/trɛs/	8	/vuit/, /buit/
4	/kuatrə/	9	/nou/
5	/siŋk/	0	/zɛru/, /sɛru/

Table 1. Pronunciation of the catalan digits.

The speech material consisted of ten repetitions of the catalan digits (table 1) uttered by six male and three female speakers (900 words). The beginning and end of every utterance were automatically detected by means of an algorithm based on the signal energy. There were a number of innacuracies in this endpoint detection, mainly due to the fact that only a small part of final plosives were considered within the word by the algorithm.

In the recognition tests, each repetition of the ten word vocabulary was alternatively taken as reference and the other nine as test, so obtaining 900 tests for each speaker.

Discussion of results.

Table 2 shows noticeable differences in the recognition performance of the nine speakers. After careful observation, we could impute to the innacurate endpoint detection a great number of recognition errors. For instance, speaker 4 has the worst recognition rate because approximately half of the final plosives /t/ were not included within the word interval.

Speaker	JL	MM	AB	PF	JG	JM	EP	NF	CB
% error	0	0.55	0	1.88	0	0	1.11	0.66	0.55

Table 2. Percentages of recognition errors without reduction. The six first speakers are male and the other three female.

Results for all speakers (8100 tests) when frame reduction is used are plotted in Fig. 1. As can be observed in Fig. 1a, the FT method gives worse results than the other two methods, a fact that agrees with past experiments [5,6]. Furthermore, when N is lower than 11 the difference increases mainly due to the digit 1 whose templates are reduced to very few (2 or 3) spectra. Concerning the CD and RD methods, they show an opposite behaviour above and below N=11 [13].

Furthermore, Fig. 1b shows that the percentage of recognition errors with the CD method decreases when the exponent r takes values lower than 1, i.e. when the time normalization is closer to the linear case, and this percentage approaches 0.5 (the no reduction score) at roughly the half of 29, the average number of frames. Global results when dynamic programming is not used are plotted in Fig. 1c, where we observe that the decreasing with r is even more pronounced.

Surprisingly, the linear normalization case r=0 gives the best results. Furthermore, the use of reduction techniques does not obtain for any N a performance improvement upon the no reduction case (the same occurs in [6] within a similar context). We maintain that there is a relation between these facts and actual innacuracies created by the automatic endpoint detection. With the purpose of showing this relation, we will closely observe the particular results of two speakers with opposite scores.

Fig. 2 shows how utterances corresponding to speaker JL allow an easy recognition since a simple linear reduction (case r=0) to 5 spectra gives an excellent score (0,33%); even without dynamic programming the recognition error is as small as 0,44%. Conversely, speaker PF exhibits the worst score when reduction is not used. However, as can be observed in Fig. 3, all of the reduction techniques achieve lower error

for some N. Furthermore, the best results are obtained with the CD method and r=0.5 instead of r=0.

The latter speaker clearly shows that reduction techniques diminish the difficulties caused by the abovementioned suppression of final consonants /t/ because they perform a heavy compression of the plosive silences, specially when r is high, i.e. close to 1.

A particular experiment was carried out so as to gain more evidence about this assertion. Because speaker JL has not serious errors in its endpoint detection, we added 120ms. of signal (it usually will be silence) at both ends of half of its utterances and the 900 recognition tests were carried out again, either with and without frame reduction. If we compare the results shown in table 3 with Fig. 2, we observe that reduction really improves scores and again is not 0 but 0.5 the best value for r.

	Without reduction	CD, N=10 r=1	r=0.5	r=0
120 ms	1,88	1,55	1,11	1,55
50 ms	0,11	0,66	0,11	0,11

Table 3

A similar experiment is reported in table 3 as well; in this case, only 50 ms. of signal were added. It seems clear that the emphasys given to transitions between speech and silence when r=1 has a negative effect on the recognition rate. Thus, if the endpoint detection algorithm only includes initial or final speech-silence transitions in a part of the utterances, a performance degradation is produced so the higher is the nonlinearity degree of the time normalization the higher is the degradation. Therefore, the innacurate endpoint determination involved in the recognition tests showed in Fig. 1 is a factor that helps the linear lenght normalization method to achieve the better global results.

4. CONCLUSIONS

Three different frame reduction techniques have been evaluated in a LPC-based isolated word recognition framework. The fixed lenght methods have shown to yield the best results, being the CD method preferable to the RD method on account of its greater simplicity. This technique has been tested with several degrees of nonlinearity, observing that there is a close relation between this degree and the kind of innacuracies produced by the algorithm that isolates words from silence. In our particular instance, the simple linear lenght normalization case (r=0) gives the best results both with and without dynamic programming.

REFERENCES

|1| L.R.Rabiner, S.E. Levinson, "Isolated and connected word recognition. Theory and selected applications", IEEE Trans. on Comm. Vol. 29, pp. 621-659, May 1981.

|2| G.M. White, R.B. Neely, "Speech recognition experiments with linear prediction, bandpass filtering and dynamic programming", IEEE Trans. on ASSP, vol. 24, pp. 183-188, 1976.

|3| M.H. Kuhn, H.H. Tomaschewski, "Improvements in isolated word recognition", IEEE Trans. on ASSP, Vol. 31, pp. 157-167, Feb. 1983.

|4| F. Soong, "A non-uniform sampling approach to isolated word recognition", J. Acoust. Soc. Am. Suppl. 1, Vol. 72, Fall 1982.

|5| J.S. Bridle, N.D. Brown, "A data-adaptive frame rate technique and its use in automatic speech recognition", Inst. of Acoust. Autumn Conf., Nov. 1982.

|6| C.K. Chuang, S.W. Chan, "Speech recognition using variable frame rate coding", Proc. ICASSP-83, pp. 1033-1036, April 1983.

|7| J.L. Gauvain, J. Mariani, J.S. Lienard, "On the use of time compression for word-based recognition", Proc. ICASSP-83, pp. 1029-1032, April 1983.

|8| R. Pieraccini, R. Billi, "Experimental comparison among data compression techniques in isolated word recognition", ICASSP-83, pp. 1025-8, 1983.

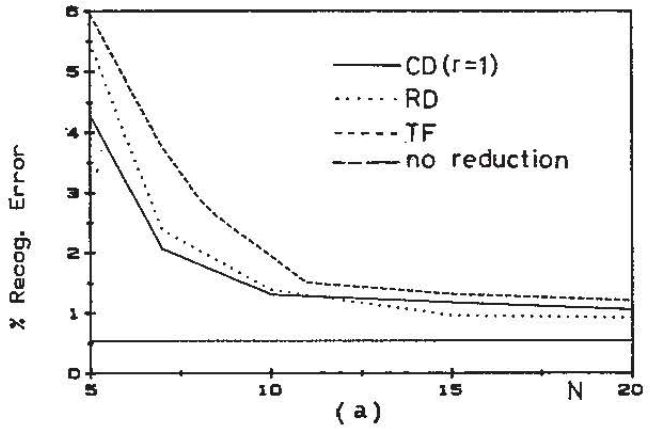
|9| J.S. Lienard, F.K. Soong, "On the use of transient information in speech recognition", Proc. ICASSP-84, pp. 17.3.1.-17.3.4, April 1984.

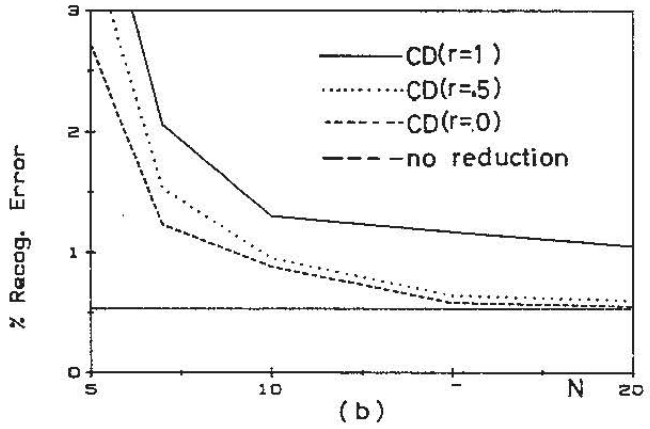
|10| J.L. Gauvain, J. Mariani, "Evaluation of time compression for connected word recognition", Proc. ICASSP-84, pp.35.10.1-4, April 1984.

|11| F. Itakura, "Minimum prediction residual principle applied to speech recognition", IEEE Trans. on ASSP, Vol. 23, pp. 67-72, Feb. 1975.

|12| E. Lleida, "Compresión de información en un sistema de reconocimiento de palabras aisladas por ajuste de plantillas", Graduation Project Report, E.T.S.E. Telecomunicació, Universitat Politècnica de Catalunya, 1985.

|13| C. Nadeu, E. Lleida, M.E. Santamaría, "Trace segmentation in a LPC-based isolated word recognition system", Proc. MELECON-85, Vol. II, pp. 111-114, Oct. 1985.





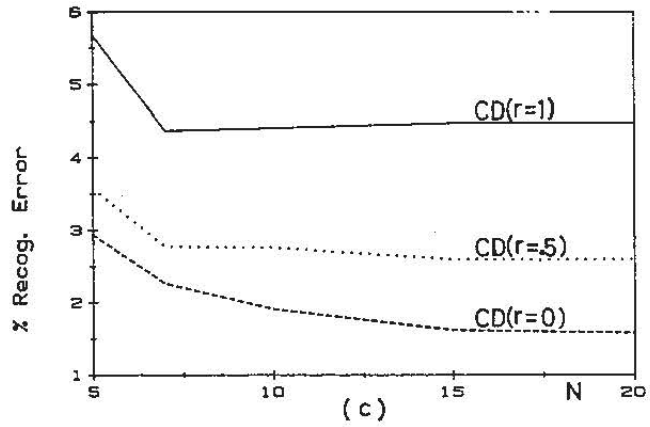
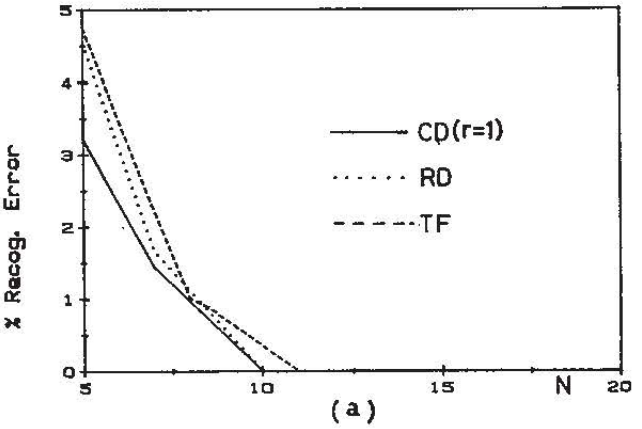


Fig. 1. Global results. (a)-(b) With dynamic programming. (c) Without dynamic programming.



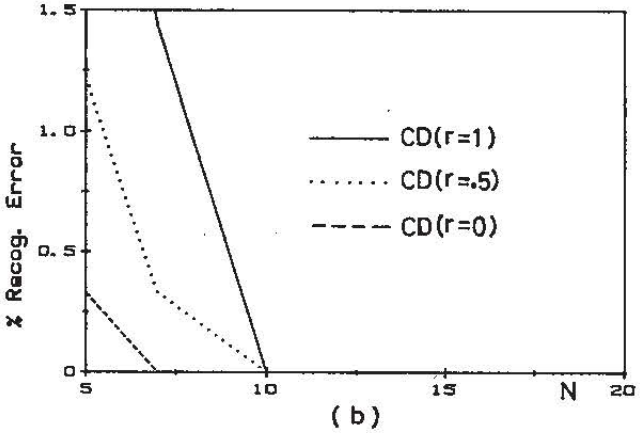
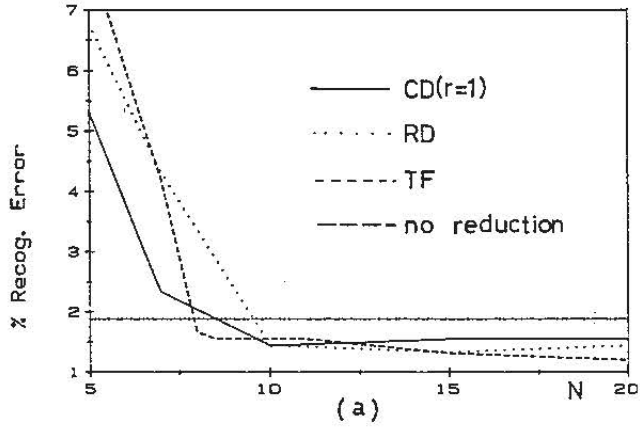


Fig. 2. Results corresponding to speaker JL.



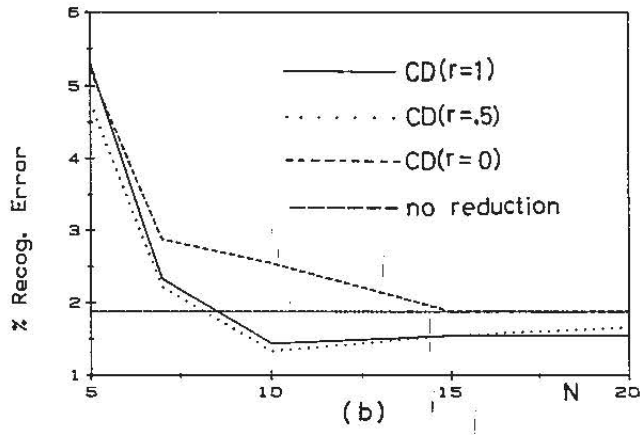


Fig. 3. Results corresponding to speaker PF.