An Hybrid Parallel Associative Memory / DTW Based System for Speech Recognition

F. CURATELLI, O.MAYORA-IBARRA
Dipartimento di Ingegneria Biofisica ed Elettronica
Universitá di Genova
Via Opera Pia 11A, 16145, Genova
ITALIA

Abstract:- This paper presents an hybrid, Parallel Associative Memory (PAM) - Dynamic Time Warping (DTW) method, for speech recognition. The present method is the core of the SPEAR speech analysis and recognition tool developed at the University of Genova, Italy. The proposed method works in two different stages for phone and word recognition respectively. Both recognition levels are detailed in a separate way using constrained tests in noisy conditions for speech dependent and independent modes. Speech recognition tests are presented using a subset of the IRST Italian language database for recognizing the ten digits. The obtained results show that the PAM - DTW method performs in a competitive way with respect to other traditional speech recognizers (MLP - HMM) for speaker dependent and independent tests under semiclean environments (S/N $\geq$ 17dB of white noise).

Key-Words:- Speech Recognition, Associative Memories, Dynamic Time Warping. Proc.pp..1751-1755

1 Introduction

Speech perception is one of the most difficult tasks performed by the human brain. The mechanism of understanding speech has been widely study in the last years [1] [2]. In all these studies, the perception and understanding of speech is explained in terms of the association capabilities of the human brain, where audio signals are codified into meaningful information. This knowledge motivates the application of computational models such as artificial associative memories to simulate the human speech perception and recognition process.

Associative memories, work as stimulus-response entities for target matching [3]. Their main characteristics are the capability of identifying pattern similarities across a set of pre-stored vectors. In this way, the speech signals have to be codified into a pattern representation for being trained to the associative memory. For this purpose, segmented speech representations have demonstrated to be the best way of modeling such signals [4].

In general, the recognition process takes place in two different levels. The first one is related to a segmented pattern recognition stage that outputs a set of frame by frame phone associations, and the second one to a concatenated recognition of such phone streams into complete words. For doing this, there have been proposed several methods and their combinations. The most used ones are:

- Dynamic Time Warping (DTW) [5],
- Hidden Markov Models (HMM) [6],
- Artificial Neural Networks (ANN) [7]
- Hybrid models [8]

In this work, a Parallel Associative Memory (PAM)

is proposed for phone recognition and the DTW algorithm for complete word matching. This kind of approach takes advantage especially in the training stage where the learning time strongly decreases in comparison to other NN-MLP based methods. This is due to the associative memory nature of the method that implies a single pass training process with a consequent reduction of time consumption.

This paper is organized as follows: Section II presents the PAMs-DTW technique in a detailed way. Section III, explains the design of speech recognition tests under clean and noisy environments. Section IV summarizes the obtained results in an separate way for phone and word recognition under both testing environments. Finally Section V outlines some concluding remarks and future perspectives.

2 PAM-DTW Method

The PAM-DTW method is the core of the SPEAR speech recognition tool developed at DIBE of the University of Genova, that combines a parallel associative memory approach with the dynamic time warping technique for speech recognition. First, the speech signal is processed with a feature extraction method, in this case, the Perceptual Linear Prediction (PLP) analysis [9] was used for extracting relevant parameters for modeling speech frames (6 parameters per frame). Every segmented speech vector was associated to its correspondent phone for constructing pattern association pairs.

Every different phone ph was considered as a possible target for the pattern matching classifier. In this way, a set of N PAMs were constructed (with *N = number of different phones in the dictionary*) for storing the information of each class of pattern associations (See figure 1).

The testing procedure is performed by presenting a new input pattern P:

$$P = [plp_1, plp_2, ..., plp_6] \tag{1}$$

where plp_i corresponds to each PLP coefficient of the test frame. The pattern matching takes place by computing distance scores between the presented vector and the set of N PAMs (See figure 2).

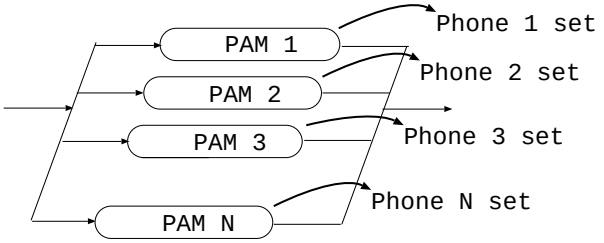

Figure 1: Parallel associative memories (PAMs) set.

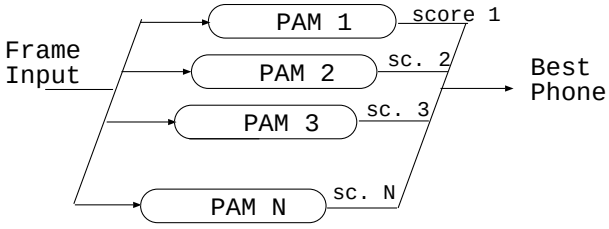

Figure 2: PAMs phone recognition.

The scores S are calculated obtaining the Euclidean nearest neighbour distance between the P pattern and every vector inside each PAM_n. In this way, a total of N scores will be obtained for the test pattern (one for each possible phone) and the best one will be the first estimated associated phone response R.

$$S_n = min[PAM_n(l) - P] \tag{2}$$

where $n = 1, 2, ..., N$ and $l = 1, 2, ..., len[PAM_n]$

$$R = ph[min(S_n)] \tag{3}$$

This process is repeated for all the test frames P_j of a word (with $j = 1, 2, ...$,*Num. of frames per word*) in order to obtain the correspondent R_j phone associations as shown in figure 3.

This last figure shows a window of the SPEAR interface, where the upper part shows the frame by frame target value in the first column and in the next ones, the best sequence of scored matches from left to right, so the second column will indicate the closest response association for the test input pattern, the third column

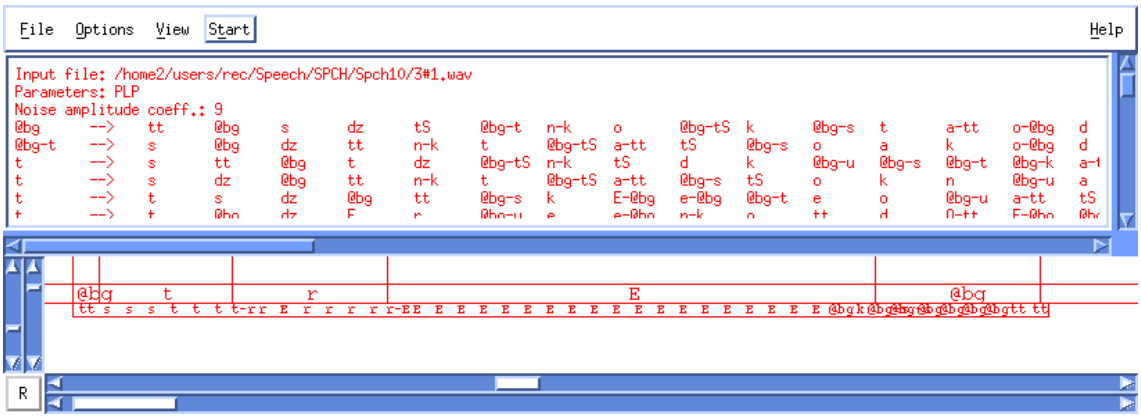

Figure 3: Frame by frame matching of phones.

will be the second best score and so on. The lower part of the window shows a two-row script. In the first row appears the correct target phone for a speech interval and in the second, the resulting chain of best frame by frame matches (so the second column's association).

This result, while sufficient in some cases (speaker dependent recognition without noise), is not frequently applicable. For this reason, a more robust recognition must include an overall estimation applied to a sequence of phones (e.g. words). This is done by suitable using the whole information provided by the score vectors.

The whole chain of score vectors is input to the DTW algorithm for selecting the word with the lowest distance when compared with the allowed words of the data dictionary (see figure 4).

3 Speech Recognition Tests

The speech recognition tests were carried out using a specific IRST database [10], containing speech registrations of the ten Italian digits. A training memory file was created using 5 female speakers each one registering 4 sessions of the 10 digits. Every session consisted in registering 4 repetitions of each digit, so a total of 800 different registrations were used for creating the training file.

This training set was processed with the SPEAR feature extraction module that constructed a 6 dimensional pattern for every 10 ms. frame. Every pattern was associated to one of 58 possible targets. The 58 classes were constructed by correlating every speech frame to an associated phone. For example, for the Italian word *zero* (pronounced *dz-E-r-o*), the system decomposes it in several frames associating a class number for each different frame. The first and last utterances of the word, were the background noise *bg* that exist in the unvoiced fragments before and after the speech start. The second utterance was the transition between the silence part and the *dz* sound and so on, constituting each sub-sound and/or transition a different class. This process is illustrated for the word *uno* (one) in figure 5.

If this process is repeated for the 10 digits from 0 to 9, the number of possible associated classes will be in total 58 as mentioned before. This process is illustrated in table 1. Every one of this 58 classes will have a different number of members depending on the repetitions of each speech sub-phone among all the words.

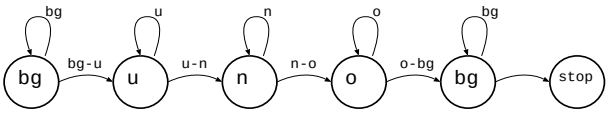

Figure 5: Sub-phone chain for word "uno"

The final training set, included phones corresponding to: A) Vowels and consonants needed in the pronunciation of Italian digits (18 classes), B) background

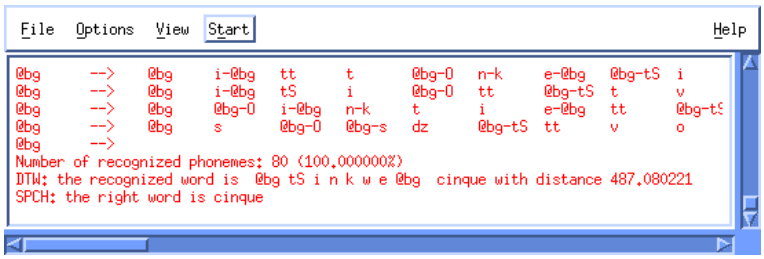

Figure 4: Recognized word after DTW analysis.

Word: zero sound: dzero		Word: uno sound: uno		... ...	Word: nove sound: nove	
class	phone	class	phone	...	class	phone
1	bg	1	bg		1	bg
2	bg-dz	11	bg-u		55	bg-n
3	dz	12	u		14	n
4	dz-e	13	u-n		15	n-o
5	e	14	n		9	o
6	e-r	15	n-o		56	o-v
7	r	9	o		57	v
8	r-o	10	o-bg		58	v-e
9	o	1	bg		5	e
10	o-bg				20	e-bg
1	bg				1	bg

Table 1: Class construction for the 10 digits.

(1 class) and C) any possible transition between the previous phonetic entities (39 classes).

The speech recognition tests were carried out for speaker dependent and independent modes. In both cases, different levels of white noise were added to verify the robustness of the PAM-DTW recognizer.

4 Recognition Results

The PAM-DTW technique was tested separately in two stages for phone and word recognition respectively. The experiments were designed considering trained and untrained patterns as data input for the system. The obtained results are listed below:

4.1 Phone Recognition

The first part of speech recognition is the correct estimation of speech basic units. Accurate phone estimation, permits to set the basis for more sophisticated recognition stages. The phone recognition obtained with the PAM method, was comparable to other NN approaches such as [11] when testing untrained data, and was superior in all the cases when testing with trained patterns. This later result is explained in terms of the intrinsic property of the associative memory paradigm that allows 100% recall when tested with trained patterns. The previous is an encouraging result for going ahead in word recognition. The detailed phone recognition results for clean and noisy experiments are shown in table 2 for either speaker dependent and independent tests:

Test data	S/N	Rec. Rate user depen.	Rec. Rate user indep.
Trained data	∞	100%	$---$
untrained data	∞	81.2%	60.1%
white noise	20 dB	79.7%	60.0%
white noise	15 dB	73.7%	57.0%
white noise	10 dB	43.7%	27.0%

Table 2: Phone recognition with PAMs under a white noise environment.

4.2 Word Recognition

The phone recognition stage, as mentioned, is only the first step towards speech recognition. A good phone estimation doesn't means that the word recognition stage will be done with total success. For doing this, it is important to consider an efficient technique for mapping phones into words.

A widely experimented technique for doing the last, is the DTW paradigm. For using it, a limited target

dictionary was created using the 10 Italian digits for testing the system. The obtained word matching results are shown in table 3

	Clean	30 dB	20 dB	17 dB	15 dB	10 dB
Sp. Dep.	100%	100%	91%	83%	30%	12%
Sp. Indep.	100%	91%	69%	50%	17%	8%

Table 3: Word recognition level under different white noise expositions.

In the last 2 tables, the noise threshold level for good recognition was founded to be of 25 dB for speaker independent and of 17 dB for speaker dependent tests.

5 Concluding Remarks

In this paper, the PAM-DTW technique was presented as a robust method for speech recognition. Experimental results with the IRST specific Italian language database, show a good performance of the system for speaker dependent and independent modes.

Some advantages of the PAM method in contrast with MLP-based systems, are 1) the faster training process (due to single pass training of associative memories) and 2) the better control and external supervision of the recognition process in the search space built during the training phase. This last characteristic is especially useful for optimizing the recognition stage.

The use of associative memories implies the possibility of achieving 100% recall values when testing with trained data. Of course, it can be argued that an MLP gains generalization at the expense of memory associations, but this increase of memory associations can be reduced with vector quantization and feature optimization algorithms. The PAM system was also tested for robustness under noisy environments. Results show accurate recognition over 15dB S/N rates of white noise.

References

[1] Junqua J.C. & Haton J.P. ”Robustness in Automatic Speech Recognition, Fundamentals and Applications”, Kluwer Academic Publishers, 1996.

[2] Picone W.J. ”Signal Modeling Techniques in Speech Recognition”, *Proceedings of the IEEE*, Vol. 81 N. 9, Sep. 1993, pp. 1215-1247.

[3] Hassoun M.H.: ”Dynamic heteroassociative neural memories”, Neural Networks, vol. 2, 1989, pp. 275-287.

[4] Rabiner L.R. & Schafer R.W. ”Digital Processing of Speech Signals”, Prentice Hall, Englewood Cliffs, N.J. 1978.

[5] Bottou L. & Fogelman F. ”Speaker Independent Isolated Digit Recognition: MLP vs DTW”, *Neural Networks*, Vol. 3, 1990, pp. 453 - 465.

[6] Rabiner L. & Juang B. ”An introduction to hidden Markov models”, *IEEE Trans. ASSP*, Vol. 3 (10), 1986, pp. 4-16.

[7] Lippmann R. ”Review of Neural Networks for Speech Recognition”, *Neural Computation* 1,1-38 MIT, 1989.

[8] Zavaliagkos G., Zhao Y., Schwartz R. & Makhoul J. ”A Hybrid Segmental NN/HMM System for Continuous Speech Recognition”, *IEEE Trans. On Speech and Audio Processing*, Vol. 1, part II, Jan. 1994.

[9] Hermansky H. ”Perceptual Linear Predictive (PLP) Analysis of Speech”, *Journal of Acoust. Soc. Am.*, April 1990, pp. 1738-1752.

[10] IRST Italian language digit database, *Istituto Ricerca Scientifica e Tecnologica*, Trento, Italia.

[11] Robinson A. ”An Application of Recurrent Nets to Phone Probability Estimation”, *IEEE Transactions on Neural networks*, Vol. 5, No. 2, Mar. 1994, pp. 298-305.