SPEECH RECOGNITION WITH A *NEURAL NETWORK TRACE-SEGMENTATION*

Euvaldo F. Cabral Jr.
São Paulo University, Polytechnic School, Department of Electronic Engineering
São Paulo - SP - Brazil
Tel: +55 11 818-5267; Fax: +55 11 818-5718
email: euvaldo@lcs.poli.usp.br

ABSTRACT

Trace-segmentation (TS) is a method for non-linear time-normalization of a sequence of speech representation frames prior to recognition of the sequence. It has been shown in a recent work [1] that an *Individual Trace-Segmentation* (ITS), i.e. a separate segmentation of the trajectory described by each individual coefficient in the speech frame leads to a much improved recognition which exceeds the performance provided by DTW recognition on the same database.

This paper describes a follow on work on the ITS technique where a Multi-layer Perceptron has been used to perform an internal mapping in the original ITS input space in order to provide a tighter set of clusters of the speech sequences. This novel technique is called *Neural Network Trace-Segmentation* (NNTS) and has produced a significant improvement on the ITS original performance.

1. INTRODUCTION

Exploring temporal variability in representations of speech is one of the outstanding problems in speech recognition. *Individual Trace-Segmentation* (ITS), i.e. a separate segmentation of the trajectory described by each individual coefficient in the speech frame has been proposed [1] to deal with temporal variability in representations of speech.

This paper describes a follow on work on the ITS technique where a Multi-layer Perceptron has been applied in a special way over ITS trajectories to perform a suitable mapping onto the input space containing the speech clusters. The objective of this mapping is to transform the speech templates to produce a tighter set of clusters of the speech sequences in order to reduce the inaccuracy of the Euclidean distance metric when used to compare two sets of ITS vectors. This novel technique is called *Neural Network Trace-Segmentation* (NNTS) and has produced a significant improvement over the ITS original performance.

2. THE STRATEGY BEHIND THE NNTS TECHNIQUE

The strategy behind the technique is the mapping of the ITS trajectories to a higher dimensional space in a way that a subsequent application of the ITS produces a better selection of the vectors along the new trace, i.e. the MLP - through the dimensional expansion carried out - is able to help the ITS finding the 'hidden' distance metric that it needs to properly select the vectors which, once back to the lower dimensional space, are more separable by class with a conventional classifier.

3 THE SPEECH CORPUS

All the analysis and experimental work described in this paper was based on a subset of the Connex Alphabet Database (binary version 0.1) from British Telecom Labs. This speech consists of three examples of each letter of the alphabet uttered by a total of 104 speakers. The speech was recorded in a silence cabinet through a high quality handset, digitally sampled at 20 KHz using a 16 bits A/D. The subset chosen for the evaluation of the *NNTS* was the same used in the work ITS [1] aiming a consistent comparison of results. It is composed by 6 letters, respectively B, P, M, R, S and T. Fifty two speakers have been designed

training talkers and other fifty two selected for the test phase.

The time domain samples were firstly converted to eight dimensional Mel Frequency Cepstral Coefficients (MFCCs) at a one millisecond rate. This unusually high frame was originally chosen [1] to ensure that none of the coefficients in the frame were undersampled for one of the objectives in [1] was to find the appropriate Nyquist rate for each frame aiming an improvement in the performance of the ITS method. Secondly, sequences of MFCCs with a variable number of vectors (different trace length were converted to a sequence of 20 ITS vectors as described in [1]. At this stage, a couple of NNTS experiments were carried out as commented below.

4. THE TWO NNTS EXPERIMENTS

4.1 General Description

Two different experiments were performed. In the first experiment, normal ITS trajectories are produced as described in [1] and used as inputs to the MLP; the outputs of the MLP have the same dimension of its inputs. In the second experiment, the outputs of the MLP are made to have a greater dimension (which outputs are considered 'traces' here) and a second process of ITS is carried out to produce new trajectories presenting the original small dimension. Figure 1 illustrates the experiments.

4.2 The Selection of the Targets

The targets are desired to be the best possible representatives of their respective classes. An obvious choice would be the centroids of the set of traces for each class produced by the ITS algorithm. Those could be determined, for instance, by the K-means clustering algorithm running on the training samples. This would produce six targets (corresponding to the six letters of the database). Unfortunately, it was found that these representatives cannot be guaranteed to be sufficiently far from each other to avoid ambiguity in the MLP training, i. e. the clusters are too much overlapped.

A more suitable approach is as follows: the K-means algorithm is applied to the ensemble of trajectories of all classes to generate 6 new clusters. The membership of each cluster is then examined and the class most frequently present in that cluster is used to label the cluster. The set of labels are examined and if there is repetition in term of the class, the label (correspondent to the repeated class) with the smallest membership assigned to it is discarded. The process is repeated until six labels are obtained which are taken as the representatives of each one of the six classes, i. e. of each letter belonging to the database used; the trajectories associated with the selected labels are now used as targets in the process of training the MLP.

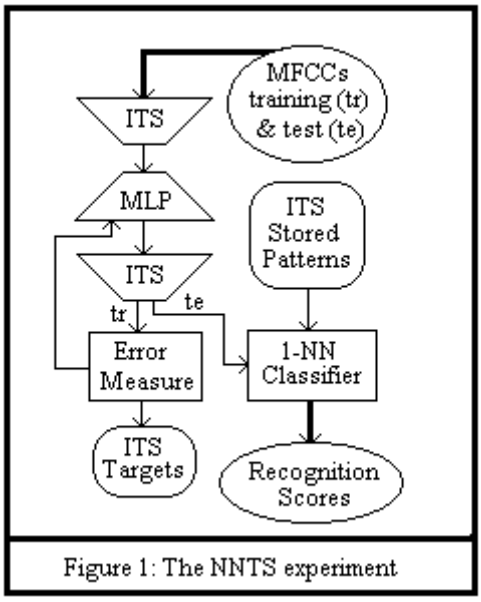

Figure 1: The NNTS experiment

4.3 Evaluation of the Derivatives for Backpropagation

It is necessary to perform backpropagation of the error in order to train the MLP in both experiments. The input vector to the MLP represents a complete trajectory by concatenating the M p-dimensional vectors along the trajectory. Thus the dimension of the input vector to the MLP is M.p.

In the first NNTS experiment, the output vector from the MLP is also of dimension M.p and can be visualized as a concatenation of M, p-dimensional vectors. The same is valid for the second NNTS experiment where the vector is of dimension N.p (N>M).

Again in the first experiment, the dimension of the input and output vectors to the MLP are the same and target vectors of dimension M.p are also available from the K-means clustering process described in the previous section. Backpropagation can therefore be easily accomplished in this system using the conventional approach to the evaluation of derivatives.

A problem with this approach may be the incapacity of the MLP of mapping the input vectors to the targets performing only shifts and rotations within the M-dimensional input space. Therefore, the second experiment is proposed in which the dimension of the output pattern from the MLP is deliberately made greater (N.p) than the input dimension; this correspond to the second NNTS experiment. It is opportune to comment that the value of N is determined experimentally.

The output of the MLP in the second experiment can be viewed as an 'oversampled' trajectory. However, the only targets available have dimension M.p and so the N.p dimensional MLP outputs must be mapped to a M.p dimensional vector by a further stage of Individual Trace-Segmentation. Since we shall eventually need to backpropagate an error through this stage of ITS in order to adjust the MLP in the second experiment, it is necessary to derive a differentiable expression for the ITS mapping; this calculation includes not only the ITS mathematical expression but also the final derivatives for backpropagation. The expressions obtained are too long to be reproduced here and will be given somewhere else [6].

4.4 Summary of the NNTS Experiments

4.4.1 NNTS first experiment

A - Training Phase:
Repeat for all training samples:

- obtain the ITS trajectory;
- use the trajectory obtained as input to the MLP;
- calculate the MLP output;
- obtain a measure of the error;
- based on the error measure obtained, correct the MLP's weights.
-

B - Testing Phase
Repeat for all test samples:

- obtain the ITS testing trajectory;
- use the trajectory obtained as input to the MLP;
- use the modified template obtained as input to the classifier.
-

4.4.2 NNTS second experiment

A - Training Phase
Repeat for all training samples: (keep all the MLP learning parameters from the NNTS first experiment)

- obtain the ITS trajectory;
- use the trajectory obtained as input to the MLP;
- calculate the MLP output;
- calculate the modified ITS trajectory;
- obtain a measure of the error;
- based on the error measure obtained, correct the MLP's weights using the appropriated equations.
-

B - Testing Phase
Repeat for all test samples:

- obtain the ITS testing trajectory;
- use the trajectory obtained as input to the MLP;
- obtain the ITS modified trajectory;
- use the modified template obtained as input to the classifier.

5. ASSESSMENT OF THE NNTS TECHNIQUE

The usefulness of the novel NNTS algorithm was accessed by conducting recognition experiments and comparing the results obtained with the ones for ITS [1] and *Dynamic Time Warping* (DTW) working on the same database. For both ITS and NNTS experiments, ten templates per class were generated using the Modified k-means Clustering Algorithm [4] and classification was then attempted using the Bell version of the k-nearest neighbor [5].

Table 1 shows the recognition scores for ITS and NNTS (benchmark scores for an interframe rate of 1 ms and 1-NN) Comparing the scores it can be noted that a improvement of 12.5% was obtained over classical TS. It is worth mentioning that the training set performance reached 100 % for both NNTS experiments. Table

2 presents the recognition scores per letter. Note that NNTS is better than DTW for all letters, although has performed worse that ITS for letter p. Informal tests have indicated that this difference can probably be corrected by fine tuning the training parameters of the MLP.

6. CONCLUSION

A combination of an artificial neural network technique with a modified version of the conventional trace-segmentation algorithm (the ITS), has produced a novel approach called *here Neural Network Trace-Segmentation (*NNTS). An evaluation of the NNTS on a subset of the English alphabet has shown that the technique performs much better than classical DTW on a speech recognition experiment and improves the performance of the ITS. The improvement obtained over the previous ITS algorithm open the possibility of using NNTS in a number of practical systems for isolated word recognition.

7. ACKNOWLEDGMENTS

The author is indebted to British Telecom PLC for access to the Conex Database and to Mr. Graham D. Tattersall of the University of East Anglia (UEA-UK) for his invaluable comments and suggestions.

8. REFERENCES

[1] Cabral Jr., E. F. and Tattersall, G. D., *Trace-Segmentation of Isolated Utterances for Speech Recognition*, Proceedings of the ICASSP, Michigan, Detroit, pp. 365-368, May 1995.

[2] Kuhn, M. H., Tomaschewski, H. and Ney H., *Fast Non-linear Time Alignment for Isolated Word Recognition*, Proceedings of the ICASSP, Atlanta, GA, pp. 736-740, March 1991.

[3] Linford, P. W. and Tattersall, G. D., *Non-linear Time Normalization of Utterances for Speech Recognition using MLP's*, Proc. of the Inst. of Acoust. Vol. 12, Part 10, 1990.

[4] Wilpon, J. G. and Rabiner, L. R., A *Modified K-means Clustering Algorithm for Use in Isolated Word Recognition*, IEEE Trans. on ASSP, Vol. ASSP-33, no. 3, June 1985.

[5] Cox, S. J., *Clustering Technique for Speaker-Independent Recognition*, BT memo G172, May 1987**.**

[6] Cabral Jr., E. F., *Individual Trace-Segmentation and Neural Network Trace Segmentation: A Formal Development*, to be submitted.

Method	Classical TS	Benchmark	ITS	NNTS (exp 1)	NNTS (exp. 2)
Ave. score (%)	71.1	73.2	76.7	78.5	83.6

Table 1: Comparative average recognition scores

letter	B	M	P	R	S	T
TS	46.0	90.4	44.2	98.1	96.1	51.9
Benchmark	66.0	88.5	46.2	98.1	98.0	42.2
ITS	62.0	92.3	59.6	98.1	96.1	51.9
NNTS	71.2	92.0	53.8	98.1	96.1	90.4

Table 2: Comparative scores per letter