Automatic Selection of Segmental Acoustic Parameters by means of Neural-Fuzzy Networks for Reordering the N-Best HMM Hypotheses.

T. Moudenc [1] *and G. Mercier*
France- Télécom - CNET/DIH/RCP 2 avenue P. Marzin, 22307 Lannion -France

ABSTRACT

We present a neural fuzzy network architecture devoted to the recognition of specific segmental phonetic features.. A neural fuzzy network allows us to select the best acoustic parameters associated with eachfeature and to compute an phonetic segmental plausibility score. Segments result from the alignements provided by an allophone based Markov model.

These segmental scores are then processed by a statistical post-processing system for reordering the N-best HMM hypotheses. This post-processing is based on the computation of segmental scores for each solution under the hypotheses of a correct solution and of an incorrect solution.

Moreover, we present comparison results between these neural fuzzy network architecture and a classical one, on 3 speaker-independent telephone databases.

1 INTRODUCTION

One way to improve an HMM-based automatic speech recognition system is to integrate segmental parameters in a post-processing stage. Different kinds of segmental parameters like duration measures or measures of stationarity [1] can be computed and integrated in such a post-processing stage [2]. In this paper we propose to use segmental parameters associated with phonetic features. It is assumed that each segment composing a word is characterized by a set of phonetic features like voiced, vocalic, nasal, lateral, and so on.

For each segment of an alignment associated to a word hypothesis, a specialized neural network, trained to recognize a specific phonetic feature, is able to compute an phoneticfeature segmental plausibility score. The aim of the post-processing module is to integrate these segmental evidences in its architecture to improve the word recognition rate. Previous work has already shown the feasibility of this approach [3].

Neural-fuzzy networks (N.F.N.) [4] are very useful tools to implement (or even to discover) acoustic-phonetic decoding rules and to adjust parameters like thresholds. The following sections describe how these networks can be built and used for computing phonetic segmental scores, on correct and incorrect segments provided by the HMM Viterbi alignment. Moreover, an efficient way to select a subset of robust acoustic segmental parameters, for identifying phonetic features, is described. Finally, recognition experiment results carried on 3 speaker-independent telephone data bases are presented.

2 NEURAL FUZZY NETWORKS ARCHITECTURES

Four main steps are necessary to build the final neural *fuzzy* network associated with each phonetic feature.

1. An initial neural network composed of four layers is defined, designed, trained and evaluated. Figure 1 represents such an initial network.

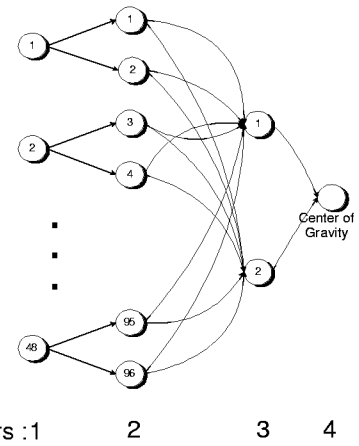

Fig. 1 : Architecture of the initial neural network

---

now at IRISA Lannion, rue E. Branly, 22300 Lannion-France, moudenc@iut-lannion.fr

The *first* layer is composed of cells representing the input parameters. The role of the *second* layer is to implement the membership values associated with each input ; this is done by means of sigmoidal functions. Two membership functions per input are usually sufficient. In the initial architecture, the *third* layer is composed of two cells only and each cell is connected to every cell of the preceding layer ; no a priori rule is implemented in this initial system. Each neuron of the third layer is also composed of a sigmoidal activation function. The *last* layer is composed of a unique cell. It is designed to compute the final output score by means of the center of gravity method. The parameters of membership functions and the weights between layers 2 and 3 and between layers 3 and 4 are adjusted with the classic back propagation algorithm.

2. The less useful connections are then pruned using the Karnin pruning procedure [5], leading to select the best subset of input parameters. Thus, a new network architecture is designed with less input parameters and with specialized cells and connections in the second and third layers. Different kind of architectures can be defined by keeping for instance two cells in the third layer and full connections between the second and the third layer and by reducing the number of input parameters and/or the number of membership functions on the basis of the pruning procedure. Figure 2 represents the selected neural fuzzy network architecture.

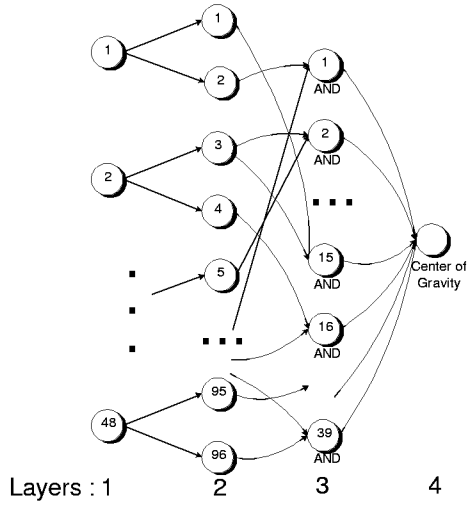

Fig. 2 : Architecture of the final neural network

This *final neural fuzzy network* is also composed of four layers but the hidden layers are completely modified. On the basis of the pruning procedure, the weights of the initial architecture which have no influence on the global error cost function are discarded. This sensitivity factor for each coefficient w_{ij} is a simple function of the partial derivative of this cost function (that is the components of the gradient). It is easily estimated during the training procedure. Deletion of connections between layers 2 and 3 is equivalent to deletion of cells (membership functions) in layer 2. The remaining useful cells of the second layer are now connected two by two to a new set of cells in the third layer, in order to implement the truth value α_k of rules defined by the following equation :

$$\alpha_k = and(\mu_{A_{ki}}(x_i) \cap \mu_{A_{kj}}(x_j))$$

where $\mu_{A_{ki}}(x_i)$ represents the membership value of x_i to the subset A_{ki}. Each truth value α_i is computed by means of the Lukaziewicz's max operator and approximated by a sigmoidal function.

3. This new neural network is then trained and evaluated. Weights between layers 2 and 3 remain constant during the training phase and other weights are modified by means of the back propagation procedure.

Results obtained with initial and final architectures are presented in section 4.

3 EXPERIMENTAL FRAMEWORK

3.1 Speech databases

3 speech databases were used, comprising 800 speakers, collected over the telephone network. The three databases are composed of the digit database (10 French digits), the Tregor database (36 French words) and the number database (from 00 to 99). Each database is split into two equal parts, a training set and a test set.

3.2 Post-Processing

As described in [2], four segmental post-processing scores are computed for each of the 5-best solutions provided by an allophone-based HMM model [6]. Each of these scores corresponds to one of the following phonetic features :

voiced/unvoiced, consonant/vowel, stop or fricative. These scores are linearly combined with the HMM score in order to provide the final score of each solution and to re-order the 5-best hypotheses. For each phonetic feature, the segmental post-processing scores are a function of the corresponding neural network outputs. The score of a solution X_i is computed as the likelihood ratio between the probability of the phonetic segmentation X_i to be correct over the probability of this segmentation to be incorrect. The probability of a solution X_i to be correct (resp. incorrect) is computed as the product of the probability of each segment *Si* to be correct (resp. incorrect). These segmental probabilities are obtained through a segmental modeling procedure of each NFN outputs associated with each phonetic feature, taking into account left and right phonetic contexts.

3.3 Neural Network (NN and NFN) architecture selection

Two types of NN architectures have been evaluated:

- the *initial architecture* has 48 input parameters (13 Mel filter bank energies, plus 26 first and second temporal derivatives of these energies ; the 9 last parameters represent 3 spectral centers of gravity followed by their first and second derivatives) ; the second layer, composed of 96 cells, and the third layer of 2 cells are fully connected (cf. Fig. 1).

- the *final architecture,* which is completely fuzzified, implements about thirty rules, with about 24 input parameters automatically extracted by the pruning algorithm. (cf. Fig. 2)

Four phonetic features have been selected : voiced/unvoiced, consonant/vowel, stop and fricative. The parameters of each specific phonetic NN are trained on each segment of the *correct* alignments of the training sets. The target is 1 when the phonetic feature is present and 0 if not. These trained networks are applied to each segmentt (correct or incorrect) composing a solution.

4 EXPERIMENTAL RESULTS

The modeling of the neural network outputs has been conducted for each contextual phonetic segment of each vocabulary. This modeling has been conducted for phonetic segments issued from *correct* HMM alignments and for phonetic segments issued from *incorrect* HMM alignments. Figures 3-a and 3-b represent an example of such models. Figure 3-a represents the model of a *correct* phonetic -vowel-segment "an" preceded by a "d" and followed by an ending silence. Figure 3-b represents the model of the same contextual phonetic segment but, issued from *incorrect* HMM alignments of the training set. The value ranges of the segmental output scores lying between 0 and 1 have been divide into 15 intervals from 1 to 15. These intervals are indicated on the abscissa (figures 3a and 3b).

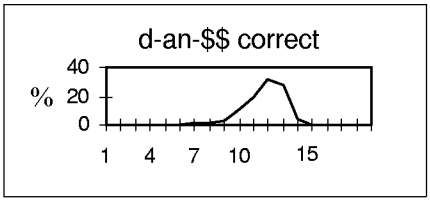

Figure 3-a : Model of the *correct* contextual phonetic segment "d-an-$$"

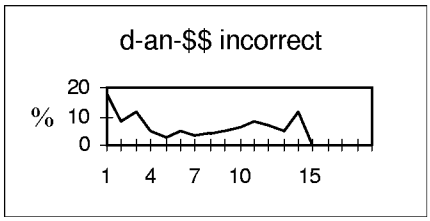

Figure 3-b : Model of the *incorrect* contextual phonetic segment "d-an-$$"

This example shows how two models of the same phonetic segment, one for the *correct* utterances and one for the *incorrect* utterances, may be discriminative in the post-processing task.

Table 1 presents the results obtained on the test sets with the initial NN architecture.

Table 1 : Recognition error rates before and after a post-processing combining HMM scores and segmental phonetic scores provided by standard NN.

	HMM alone	HMM + 4 NN Post-Processings	
	Error rate	Error rate	Error rate reduction
Digit	1.13 %	0.91 %	19 %
Number	3.58 %	3.45 %	3 %
Tregor	0.83 %	0.71 %	14 %

Results obtained with the final neural fuzzy network are presented in Table 2.

Table 2 : Recognition error rates before and after a post-processing combining HMM scores and segmental phonetic scores provided by NFN.

	HMM alone	HMM + 4 NFN Post-Processings	
	Error rate	Error rate	Error rate reduction
Digit	1.13 %	0.94 %	17 %
Number	3.58 %	3.43 %	4 %
Tregor	0.83 %	0.65 %	22 %

These results show that it is possible to reduce markovian recognition error rates by using phonetic features and fuzzy neural nets. Extracting rules resulting from the pruning procedure enables to reduce the number of input parameters by a factor of 2 without any loss of performances on digits and numbers and better performances on the Trégor database (improvement of 8 %). By looking at the parameters selected by the pruning procedure, we have also observed that they differ from one phonetic feature to another. For instance it could be observed that first and second derivatives and spectral centers of gravity are very important parameters for distinguishing consonants from vowels. It can also be noticed that better post-processing performances (table 3) are obtained with the consonant/vowel and stop features, than with fricative (badly identified in the telephone frequency range) or voiced features (the voiced feature is very context-sensitive and even speaker-dependent).

Table 3 : Recognition error rates reduction measured on each database for each phonetic feature

	Voiced Unvoiced	Fricative	Vowel Consonne	Stop
Digit	0 %	7 %	15 %	17 %
Number	4 %	1 %	3 %	6 %
Trégor	8%	0 %	15 %	5 %

5 CONCLUSION AND FUTURE WORK

In this paper, we have described an automatic selection method of relevant segmental parameters for each phonetic feature by means of neural fuzzy networks.

Applying this hybrid architecture (neural fuzzy networks + statistical post-processing of HMM N-best solutions) for rejecting extraneous speech or for rescoring word hypotheses in a word lattice are natural extensions of this work. Working on new phonetic features, adding new segmental input parameters are other possible directions of research.

References

[1] André-Obrecht R., "A New Statistical Approach for the Automatic Segmentation of Continuous Speech signal", *IEEE Trans. on ASSP*, Vol 36, N° 1, pp. 29-40, 1988.

[2] Moudenc T., "On using an a priori segmentation of the speech signal in an N-best solutions post-processing", *ICASSP 95*, Detroit, USA, pp. 580-583, 1995.

[3] Moudenc T., Sokol R., Mercier G., "Segmental Phonetic Features Recognition by means of Neural-Fuzzy networks and integration in an N-Best Solutions Post-Processing", *ICSLP 96*, Philadelphia, USA, Vol 1, pp. 338-341.

[4] Glorennec P.Y., "A general class of fuzzy inference systems : application to identification and control", *Proc. ESSC 93*, Prague, Czech, 1993.

[5] Karnin, E.D., "A simple procedure for pruning back-propagation trained neural networks", in *IEEE trans. Neural Networks*, vol 1, n° 2, 1990.

[6] Jouvet D., Bartkova K. & Monné J., "On the Modelization of the Allophones in an HMM based Speech Recognition System", *Proc. EuroSpeech* 91, pp. 923-926, Genova, Italy.