Segmental Phonetic Features Recognition by means of Neural-Fuzzy Networks and Integration in an N-Best Solutions Post-Processing

T. Moudenc, R. Sokol, G. Mercier

France-Télécom - CNET-LAA/TSS/RCP
2,avenue P. Marzin, 22307 Lannion Cedex - France
e-mail : moudenc@lannion.cnet.fr

ABSTRACT

In this paper, we present investigations on using segmental phonetic features in an N-best solutions post processing of an HMM based ASR system. These phonetic features are extracted by means of neural-fuzzy networks.

Specialized neural-fuzzy networks are defined to recognize specific phonetic features (consonant/vowel, voiced/un-voiced, ...). Each of these neural networks furnishes a segmental coefficient (resulting from the output layers) which enables the computation of a segmental post-processing score for the N-best solutions of an HMM based ASR system. This post-processing is based on the computation of segmental score for each solution respectively under the hypotheses of a correct solution and an incorrect solution.

Preliminary experiments were conducted on 3 speaker-independent telephone databases. An error rate reduction up to 20 % was achieved on the Digit corpus.

1. INTRODUCTION

Integrating segmental parameters in a post-processing stage may be of interest to recover the remaining errors of an HMM-based automatic speech recognition system. Previous studies have already shown the interest of such an approach. For example, Gish [1] introduces a segmental discontinuity score as a function of time based on a similarity measure between adjacent frames; Schwartz [2] uses segmental neural networks trained to distinguish between correct and incorrect phonetic segments issued from the HMM. He also uses different knowledge sources such as stochastic segment models [3], high-order language models, ...

In the KEAL system [4], acoustic-phonetic knowledge is represented by a set of rules. This knowledge may also be modeled by a set of neural-fuzzy networks built to detect segmental phonetic features [5][6]. Each neural-fuzzy networks is then able to furnish a parameter which characterizes a particular phonetic feature, like *consonant* or *vowel* phonetic segments.

The particularity of the CNET segmental post-processing approach resides in the formalism used, which considers a rescoring of each solution under the hypotheses of a correct solution and of an incorrect solution respectively [7]. This post-processing formalism has been applied in [8] with a modeling of the abrupt changes in the speech signal and in [9] with a segmental duration modeling of the phonetic segments of each of the N-best solutions.

In this paper, we propose to apply a segmental post-processing based on a modeling of the neural network outputs. To integrate these parameters in the post-processing task, it is necessary to estimate two discrete probability distributions. One of these distributions represents the probability distribution of the output scores on the correct segments (resulting from the correct solutions) while the other represents the distribution on the incorrect segments (from the incorrect solutions).

The next section describes the neural-fuzzy network architecture, together with some of the features we take into account. Section 3 sums up the post-processing rescoring and the estimation of the discrete probability distributions associated to each considered phonetic feature. Section 4 details the experimental results carried on 3 speaker-independent telephone databases. Conclusion and future work are addressed in the last section.

2. DESCRIPTION OF THE NEURAL-FUZZY SYSTEM

Multiple acoustic cues and several rules are often necessary to identify a phonetic event in the acoustic representation of speech. For instance, in the KEAL knowledge-based approach [4], more than thirty rules are used to separate *consonantal* segments from *vocalic* ones and about thirty acoustic parameters are treated in this set of rules to sum up the various acoustic images related to these segments. These parameters are filter bank energies, cepstral coefficients, formants, centers of gravity, linear and non-linear combinations of these basic parameters but also local temporal variations like spectral derivatives or more global temporal variations like linear (or non-linear) regression coefficients. These global temporal parameters allow to represent the temporal directions of formants or the shape of valleys or hills visible on any energy curve between two consecutive vocalic nuclei.

In a neural-fuzzy system, each rules

if (cond_1 ... *and* cond_n) *then* decision_n)

can be formalized by the following form :

$$\textit{if}\ (\mu_{A_1}(x_1)\ \ldots\ \textit{and}\ \ldots\ \mu_{A_n}(x_n))\ \textit{then}\ \mu_B(y_s)$$

where x_i are input variables (parameters), y_s the output variable, and $\mu_A(x)$ a membership value representing the membership of x in the subset A. This membership function takes its values between 0 and 1 and can be implemented by gaussian or sigmoïdal functions. The structure of our neural-fuzzy network is based on preliminary studies described in [5][6][10]. This system is in fact a four layer network where each layer plays a specific role (figure 1).

The first layer corresponds to the input variables : spectral parameters, acoustic and temporal cues. The weights between the input layer and the second layer implement the evaluation of the membership values. The truth value α of each rule, that is the measure that the observation belongs to some specific phonetic feature is given by the value of the following expression :

$$and\left(\mu_{A_1}(x_1) \cap \mu_{A_2}(x_2) \cap \cdots \cap \mu_{A_n}(x_n)\right)$$

and is evaluated in the third layer by means of the Lukasiewicz's conjunction operator given by :

$$and(A,B) = \max\left(0, \mu_A(x) + \mu_B(y) - 1\right)$$

In this case it is possible to approximate the *max* function by a sigmoïdal function but every rule which includes more than two conditions has to be decomposed in a set of rules including two conditions only. The weights between layer 2 and 3 remain constant during the training phase. The weights between layer 3 and 4 give the evaluation of each individual rule and the output value is computed by the center of gravity method :

$$y_S = \sum \alpha_i W_i \,/\, \sum \alpha_i$$

In this case, the output neuron is not standard and the back propagation algorithm has to be modified in order to manage this special output function [6].

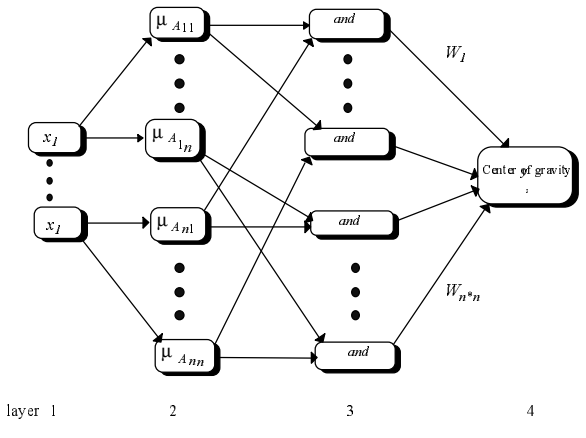

Figure 1 : Neural fuzzy network architecture

This is a flexible architecture giving us many possibilities :

1. The possibility to introduce specific parameters (extracted from any knowledge-based rules), in the first input layer.

2. The possibility to implement a set of appropriate rules representing the a priori knowledge, for each phonetic feature. In this case, the conjunctions and the w_i parameters are determined by the rules and remain constant. The membership values are the only parameters modified by the learning algorithm.

3. The possibility to use this system as a classical neural network with full interconnection between layers 2 and 3. The segmental post-processing results presented in section 4.3 are obtained with such a network.

4. The possibility to discover the most important combinations of the membership values by pruning irrelevant connections between layers 2 and 3 and then to retrain the new network composed of a reduced set of rules, which corresponds to the most important connections.

3. SEGMENTAL POST-PROCESSING

3.1. Segmental rescoring

The segmental post-processing score of each of the N-best solutions is computed as follows [8] :

$$Sc_{PP}(X) = -\text{Log}\left(\frac{\Pr(X \text{ is correct})}{\Pr(X \text{ is incorrect})}\right)$$

where X represents the phonetic segmentation for the current solution. For a given alignment X_i composed of S_i segments, the probability to have a correct alignment is then computed as the product of the probability of each segment $V_{i,j}$ composing the alignment using the correct models.

$$\Pr\left(X_i \,/\, M_{i,correct}\right) = \prod_{j=1}^{S_i} \Pr\left(V_{i,j} \,/\, M_{V_{i,j},correct}\right)$$

$$X_i = \left(V_{i,1}, \cdots, V_{i,S_i}\right)$$

As such, segments are presumed to be independent. The probability for the alignment to be incorrect is computed in a similar way using the "incorrect" models that are associated to incorrect segments.

3.2. Segmental modeling

The probability of each segment is obtained using discrete probability distributions. The observation is the output of the neural-fuzzy network. Two discrete probability distributions are estimated from histograms distributions computed separately on the set of the correct segments and on the set of the incorrect ones.

The segmental modeling takes into account the phonetic context of each segment. A reliability threshold R is used for estimating these context-dependent probability distributions. If the number of context-dependent examples is lower than this threshold R, a smoothing of the context-dependent model is carried out using the context-independent model. R specifies how much data is presumed to be necessary in order to obtain a reliable estimate for the discrete probability models.

3.3 Combination with the HMM Score

Once the post-processing score is computed for a solution, it is linearly combined with the Markovian score, thus providing the final score for the solution :

$$Sc\left(X_i\right) = \alpha\, Sc_{HMM}\left(X_i\right) + \left(1-\alpha\right) Sc_{PP}\left(X_i\right)$$

The solution which delivers the lowest combined score is regarded as the answer.

If multiple post-processing scores are computed for each segment, then the final score is given by the following equation :

$$Sc\left(X_i\right) = \alpha_0 Sc_{HMM}\left(X_i\right) + \sum_{k=1}^{NbPP} \alpha_k Sc_{PP_k}\left(X_i\right)$$

$$\text{with : } \alpha_0 + \sum_{k=1}^{NbPP} \alpha_k = 1$$

The optimal set of coefficients α_p is estimated over the training set by means of the Powell algorithm [11][12]. This algorithm finds the parameters that minimize for instance the recognition error rate computed over the training set.

4. TESTS AND RESULTS

In the following experiments, the N-best solutions are determined using a context-dependent phoneme-based HMM. The acoustical input vectors are formed by 8 cepstral parameters plus the energy, and their first and second order temporal derivatives, computed on a 5-frame window.

4.1 Databases

In the experiments, 3 speech databases were used, comprising 800 speakers, collected over the telephone network, covering different regional accents. Each database was split into 2 parts : the training set for training model parameters, and the test set for evaluating recognition performance.

The 3 databases are composed of the "Digit" database (10 digits), the "Trégor" database (36 French words) and the "Number" database (numbers between 00 and 99).

4.2 Phonetic Feature Recognition

The first set of experiments concerns the evaluation of the neural fuzzy networks and their ability to provide reliable segmental parameters. The phonetic feature recognition performances of the neural-fuzzy system were evaluated on the "Trégor" database.

Phonetic feature	Feature recognition error rates	
	basic neural network	neural-fuzzy network
Voiced/unvoiced	13.0 % (±0.5)	10.3
Vowel/consonant	14.3 % (±0.3)	12.8
Fricative	12.5 % (±0.4)	12.0
Stop	14.8 % (±0.4)	12.8

Table 1 : Recognition error rates on phonetic features with neural-fuzzy networks.

In this table, the second column represents the phonetic feature recognition error rate obtained on the test set with the basic neural networks. Each percentage is given with its half confidence interval measure (95%). Each neural network is provided by the same input parameters. This normalized input vector is composed of 42 basic components including the 14-channel filter bank energies and the temporal first and second derivatives of each channel energy. These parameters are computed at the frame level and the output of the neural network is averaged over the frames belonging to the current segment. The weights of each neural net are estimated by the modified back-propagation algorithm and for this purpose, the training set is itself divided into two parts; half of the data being kept for cross-validation experiments. The last column reports the results obtained by using the neural-fuzzy networks built as follows: a priori knowledge (specific rules and temporal parameters) are introduced in the basic network and the pruning technique is applied to reduce the number of connections.

4.3 Segmental Post-Processing Test Results

The last set of experiments concerns the evaluation of the neural-fuzzy network output modeling for the segmental post-processing task.

The next table reports the results obtained on the test sets of each used database. For each of the N-best solutions, four segmental post-processing scores are computed. Each of these scores corresponds to one of the following phonetic features : voiced/unvoiced, vowel/consonant, stop and fricative. These 4 scores and the HMM score are linearly combined to provide the final score of each solution. The coefficients of this linear combination are themselves computed on each training set.

	Digit (0 to 9)	Trégor (36 words)	Number (00 to 99)
Error-rate HMM alone	1.13 %	0.83 %	3.58 %
Error-rate HMM+Post-Processing	0.91 %	0.71 %	3.45 %
Error-rate reduction	**19 %**	**14 %**	**3 %**

Table 2 : Results obtained on the test sets.

The first row in Table 2 gives the error rates obtained with the HMM alone, the second row reports the error rates obtained after the post-processing task, and the last row indicates the reduction in the error rates.

To compare with previous experiments, Table 3 reports the results that were obtained using a segmental post-processing of the N-best solutions based on the abrupt changes in the speech signal [8].

Error-rate reductions HMM+Post-Processing	Digit (0 to 9)	Trégor (36 words)	Number (00 to 99)
abrupt changes	**15 %**	**5 %**	**9 %**
neural-fuzzy networks	**19 %**	**14 %**	**3 %**

Table 3 : Comparison with segmental post-processing results previously reported in [8].

These results show that the use of phonetic features - extracted by neural networks - improves the recognition performances up to 20 %. Moreover, on the different corpus, this improvement does not behave as the one obtained with the abrupt changes post-processing : worse results on the Number database but better results on the Digit and Trégor databases. Thus, a combination of these 2 sets of parameters may be useful.

5. CONCLUSIONS AND FUTURE WORK

This study focuses on two points : the investigation of segmental parameters based on phonetic features and their implementation in a post-processing speech recognition system. For this purpose, a neural-fuzzy network architecture has been built which enables us to include specific segmental parameters, to evaluate them and to extract a set of reduced rules. The output of the network is then incorporated in the post-processing system. Four segmental phonetic features only were used and preliminary results presented in table 2 are encouraging.
Investigations will be conducted in several directions : the phonetic feature set and the corresponding fuzzy-neural networks have to be optimized, and then combined with other segmental knowledge sources.

ACKNOWLEDGMENTS

This study was partially subsidized by the Conseil Régional de Bretagne. The authors wish to extend their appreciation to Katarina Bartkova, Denis Jouvet, and Jean Monné for their helpful comments about this work.

REFERENCES

1. Gish H., Kenney Ng, Rohlicek J. R., "Secondary processing using speech segments for an HMM word spotting system", *Proc ICSLP 92* , pp. 17-20, Banff, Canada, 1992.

2. Schwartz R., Austin S., Kubala F., Makhoul J., Nguyen L., Placeway P., Zavaglios G., "New Uses for the N-best Sentence Hypotheses within the Byblos Speech Recognition System", *Proc. ICASSP 92*, pp. 1-4, San Francisco, USA, 1992.

3. Ostendorf M., Roucos S. "A stochastic segment model for phoneme-based continuous speech recognition", *IEEE Trans. on ASSP*, vol 37, pp.1857-1869, 1989.

4. Mercier G., Bigorgne D., Miclet L., Le Guennec L., Querré M., "Recognition of Speaker-Dependent Continuous Speech with KEAL", *Readings in Speech Recognition*, pp. 225-234, Waibel and Lee editors, 1990.

5. Sokol R., Mercier G., "Neural-Fuzzy Network for Phonetic Features Recognition", *Proc. EuroSpeech* 95, pp. 1579-1582, Madrid, Spain, 1995.

6. Sokol R., "Réseaux neuro-flous et reconnaissance des traits phonétiques pour l'aide à la lecture labiale", *thèse de doctorat* de l'université de Rennes 1, 1996.

7. Lokbani M. N., Jouvet D., Monné J., "Segmental Post-Processing of the N Best Solutions in a Speech Recognition System", *Proc. EuroSpeech*, Berlin, Germany, pp. 811-814, 1993.

8. Moudenc T., Jouvet D., Monne J., "On Using A-Priori Segmentation of the Speech Signal in an N-Best Solutions Post-Processing", *Proc. ICASSP 95*, pp. 580-583, Detroit, USA, 1995.

9. Bartkova K., Jouvet D., Moudenc T., "Using segmental duration prediction for rescoring the N-best solution in speech recognition", *Proc. ICPhS 95*, Stockholm, Sweden, 1995.

10. Glorennec P. Y., "A general class of Fuzzy Inference Systems : Application to identification and control", *2th European System Science Congress*, Prague, Czechoslovakia, 1993.

11. Ostendorf M., Kannan A., Austin S., Kimball O., Schartz R., Rohlicek J. R., "Integration of Diverse Recognition Methodologies through Reevaluation of N-Best Sentence Hypotheses", in *DARPA Speech Natural Language Workshop*, Monterey, USA, 1991.

12. Press W. P., Flannery S.A., Teukolsky S.A., Vetterling W.T., "*Numerical recipes in C*", Cambridge University Press, 1990.