MODULAR CONNECTIONIST SYSTEMS FOR IDENTIFYING COMPLEX ARABIC PHONETIC FEATURES

Sid-Ahmed Selouani[1] *Jean Caelen*[2]

[1]Houari Boumedienne University of Science and Technology
Speech Laboratory, Institute of Electronics, BP 32 El Alia-Algiers
[2]Informatique et Mathématiques Appliquées de Grenoble
CLIPS, BP 53, 38041 cedex 9 Grenoble
{sid-ahmed.selouani@imag.fr, jean.caelen @imag.fr}

ABSTRACT

This paper presents an approach using a mixture of connectionist experts for the identification of complex Arabic phonetic features such as the emphasis, the gemination and the relevant duration of vowels. These experts are typically time delay neural networks using a version of autoregressive backpropagation algorithm (AR-TDNN). A serial and parallel architectures of AR-TDNN have been implemented and confronted to a monolithic system. The parallel configuration achieved much fewer error rate (13% vs. 16% and 28%) than other architectures. This leads us to develop a hybrid system based on hidden Markov models (HMM) and a parallel configuration of AR-TDNN. Binary discrimination sub-tasks are assigned to these neural experts in order to enhance HMM classification capabilities of complex phonemes. Results show that 10% reduction of error rate is obtained by the hybrid system in comparison with a baseline system.

1. INTRODUCTION

Arabic phonetics is mainly funded on the relevance of lengthening in the vocalic system and on the presence of emphatic and geminated consonants. Designers of automatic speech recognition systems (ASR) dedicated to Arabic language have unanimously observed that emphasis, gemination and vowel's lengthening constitute the main root of failure [3][11].

In the case of an emphatic vs. non-emphatic opposition, we require the ASR system to distinguish, for example, between the two words: /nasaba/ (he imputed) and /nasaba/ (he erected), where an emphasis is observed over /s/. The present ASR systems cannot easily raise this ambiguousness. In the example of the words: /θabata/ (it was proved) and /θab:ata/ (he stabilized), the opposition resides in the gemination of /b/ plosive. Through this example, we measure the importance and the difficulty to automatically perform this feature detection. A similar problem is encountered in the vocalic system. For example, the two words: /jama:l/ (beauty) and /jamal/ (camel) differ only by the lengthening of the second vowel. We require the recognition system to detect this vowel without altering its temporal property. The temporal normalization performed by classical systems will penalize this detection.

In the particular case of Arabic, the system must be capable to distinguish between a time lengthening due to a variation of speech rate (utterance speed) and the one due to the presence of long vowels or geminate consonants. Another problem encountered by present ASR systems is the identification of a feature as subtle as emphasis.

The approach presented here is based on time delay neural networks (TDNN) structure using an autoregressive (AR) version of backpropagation algorithm. The aim is to identify complex Arabic phonetic features in order to improve ASR performances. This structure consists of serial or parallel disposition of connectionist experts. Binary classification sub-tasks are individually assigned to this ensemble of sub-neural-networks (SNNs). These experts can be introduced in a hybrid system in order to "boost" the capability of an HMM-based system to deal with complex phonemes.

Serial and parallel architectures of SNNs are presented in section 2. In section 3 we describe the integration of SNNs in a hybrid structure using HMM as main system. Results of the implemented systems are presented and commented in section 4.

2. MODULAR CONNECTIONIST ARCHITECTURES

A hierarchical structure of experts is introduced by Jakobs and Jordan [5] in order to solve problems of non-linear regression. They were inspired on principle of divide and conquer where a hard problem is broken up into a set of easier problems to solve. In the case of spontaneous telephonic speech, this structure was revealed more efficient than traditional monolithic networks [2]. In the case of complex Arabic phonetic features detection and identification, we propose a similar approach where binary sub-tasks have been assigned to a set of time delayed SNNs using an autoregressive version of backpropagation algorithm [9].

2.1. Auto-Regressive Time Delay Neural Networks (AR-TDNN)

The tone variations characterizing emphatic and geminated consonants and long vowels free us of the explicit computation of duration parameter. As it is shown in different studies [1][3][10][11], these variations influence the phonetic context of the phoneme to recognize. Thus, it reinforced us in the idea that this tone 'perception' must be previously 'learned' by a system which simultaneously 'memorizes' the phonetic contexts of the sequence to identify. Russel [9] showed that the use of an

autoregressive version of backpropagation algorithm (AR-back propagation) gives the neural network this memorization ability. In the version we propose, in the network input layer, a delay component similar to the one used by Waibel TDNN [12] is integrated in addition of AR component. In the detection of relevant phoneme duration, we expect that this combination increases the ability of the system to discern the phonological length even in a strong coarticulation context. The output (y_i) of the neuron we use is given by the following expression:

$$y_i(t) = f\left(bias_i + \sum_{j=1}^{P} \sum_{m=0}^{L} w_{i,j,m}\ x_j(t-m) \right) + \sum_{n=1}^{M} a_{i,n} y_i(t-n)$$

f(x) being sigmoid function. *P* is the number of input units. *L* is the delay order at the input. *M* is the order of autoregressive prediction. Weights $w_{i,j,m}$, biases and $a_{i,n}$ coefficients are parameters to optimize.

The AR component of the network gives it the ability to recognize series of sequences in a certain phonetic context which is implicitly learned. The discrimination of emphatic, geminated consonants and the long vs. short vowels is performed using the previous values stored in the delays as well as in the feedback.

The learning phase is performed such as if a phone of the target-phoneme appears in the speech continuum, the network activation arises gradually in one of its two outputs. In the case of geminated consonants detection/classification example, as it is illustrated in figure 1, the task is to learn to recognize this sequence: LCG-GEM-RCG: LCG is the left phonetic context of the geminated consonant (noted GEM) and RCG is its right phonetic context.

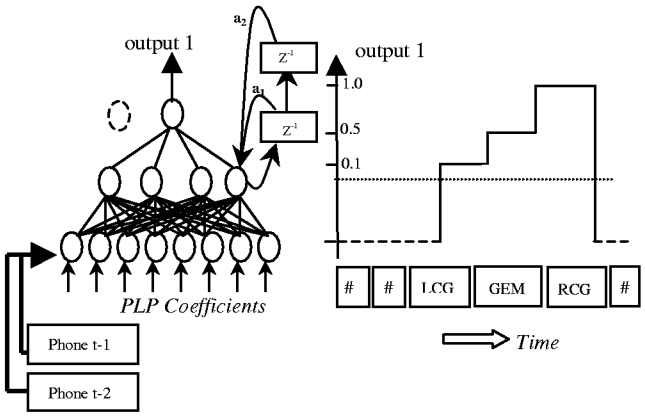

Figure 1: AR-TDNN phone-based identification process

GEMI_NET (gemination expert network) receives three input token at a time t and it must detect a geminated sequence from any other sequence combination The learning consists in setting at the high level (+1) the first output when the end of the LCG-GEM-RCG sequence is attained. Low level (-1) is set otherwise. The second output is set at the high level if a scrolling (stream) of non-geminated phone sequences is observed. An autoregressive order of 2 is chosen and a delay of 2 phones is also fixed. These lower values of delay and order are justified by the fact that phones are used instead frames. Consequently the stability of AR nodes is insured.

2.2. Serial Disposition of Experts

A serial hierarchical disposition of AR-TDNN is firstly tested. The progress in the structure (cf. figure 2.a) is conditioned by the activation of the two outputs of a given sub-network. After a macro-classes (vowels, fricatives, plosives, nasals and liquids) identification, a finer classification is performed. A discrimination between long and brief vowels and detection of gemination and emphasis are achieved. The process starts the identification of the emphatic feature or the geminate feature if one of the consonant macro-classes is detected. Otherwise an activation of the contiguous network is operated. A failure of the overall system is accounted if the last network is attained without any discrimination. Success rate during the cross-validation step [6] determines the rank of a given expert. Thus, superior levels are occupied by the most accurate systems.

The hierarchical structure of this system seems inadequate since the networks located deeply are penalized. An architecture putting networks in the same level of competence seems less constraining.

2.3 Parallel Disposition of Experts

This configuration is established without any fixed condition concerning the individual performances of each expert. Experts have the same activation potential when a sequence to identify is presented at the input (cf. figure 2.b). This property allows a flexible use and avoids the failure case. A pruning module based on *a priori* rules and grammar constraints performs the final identification. This module manages cases where several experts are simultaneously activated. The pruning procedure is summarized as follows:

Let S_{j1} be the output 1 of the expert *j*. Ideally, it takes state +1 if the feature (assigned to that expert) is detected. In that case, we say that the expert is activated.

Let S_{j2} be the output 2 of the expert *j*. Ideally, it takes state +1 if the aimed phonetic feature is absent. A distance between the distributions of the two-classifier outputs is estimated as follows:

$$Dist_j(S_{j,1},S_{j,2}) = (\mu_{j,1} - \mu_{j,2}) - (\sigma_{j,1} + \sigma_{j,2})$$

Where $\mu_{j,1}$, $\sigma_{j,1}$ are respectively the mean and the standard deviation of observed values at output 1 for the j^{th} binary classifier. If the distance of a classifier is negative, this classifier is not retained for the final decision. In the case of positive distance, the corresponding classifier is retained to make the decision about recognized pattern. The identified class is the one having its corresponding network verifying: $Argmax_i(Dist_i)$. The *i* index covers only retained classifiers. Afterwards, the long/brief expert is activated if (and only if) the vowel output is at high level. Otherwise, in the consonant case, whatever the activated expert, the "gemination expert" is solicited. The emphatic expert is activated only if inferior levels have detected a plosive or a fricative. This is carried out according to Arabic grammar properties.

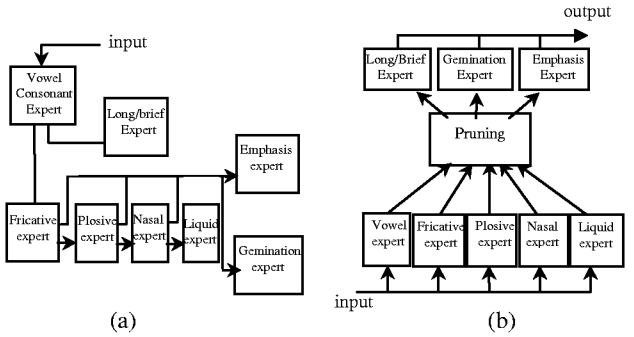

Figure 2: Serial (a) and parallel (b) connectionist structures for the identification of Arabic phonetic macro-classes and features.

3. HYBRID HMM-SNN SYSTEM

The SNNs are used in a hybrid structure to enhance the accuracy of a HMM-based system. The overall structure of the proposed system is shown in Figure 3.

It was showed in [8][11] that if auditory models are used for acoustical analysis, it is not necessary to integrate an explicit time parameter and consequently ASR systems acquire more robustness towards time variation. Two auditory models are used in the hybrid system we propose: Caelen ear model [10][11] for homogenous phone segmentation and PLP (perceptual linear predictive) [4] technique as acoustical analyzer for SNNs. The Caelen ear model consists in the determination of a 24 channel spectrum (24 coupled filters) by modeling the basilar membrane. From a particular linear combination of the channels outputs, 7 cues are derived: acute/grave, open/close, diffuse/compact, sharp/flat, mat/strident, continuous/discontinuous and tense/lax. A delta coding of these acoustic indicative features is done in order to find out their variation and to perform homogeneous phone segmentation. For each phone, the log-energy of the 24 channels outputs are used as parameters by HMMs. A mean vector of PLP coefficients over frames constituting a phone is fed into SNNs. This is justified by the fact that this acoustical analysis gives the best cross-validation results and conducts to a useful (fewer inputs) structure of the network.

Training the HMM/SNN on an utterance proceeds in two steps. The first step performs optimal alignment between the acoustic models of phones and the speech signal. In the second step the SNNs act as post-processor to HMMs and refine their recognition results. The global task is then divided between the main system constituted by HMM and the "booster" system composed of SNNs. We require HMMs to achieve phone identification without discriminating between long and short vowels and between emphatic and non emphatic consonants. The gemination detection is also not required. The hand-labeled data set presented to HMMs presents a single label for phonemes belonging to these macro-classes. For example, in the case of /a/ short vowel and /a:/ long vowel, a unique /A/ label is given. The /A/ sequence of phones is presented to the AR-TDNN mixture which confirms that the sequence is vowel and makes a final and finer decision by the network specialized in the long/short vowel discrimination. The same process is conduct for the other complex phonemes.

For each phone of input speech at time *t*, we feed phone(t-1) and phone(t-2) into all the SNNs. Each sub-network is trained such it achieves this finer discrimination as it is described in section 2.1.

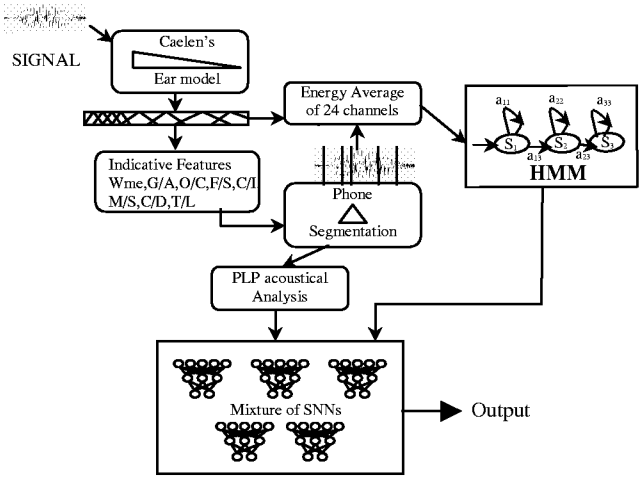

Figure 3: HMM-SNN hybrid structure for Arabic speech recognition.

4. EXPERIMENTAL RESULTS

The test database consists of 60 VCV utterances and 50 phrases. The corpus has been pronounced by six Algerian native speakers (3 men and 3 women). As a whole, the test concerns 3724 vowels (1348 long), 1197 fricatives (182 geminated, 193 emphatics), 1089 plosives (215 geminated, 273 emphatics), 573 nasals and 413 liquids. The semi-vowels are assimilated to their corresponding vowels.

We firstly compare performances of serial and parallel configuration of SNNs to a monolithic (simple) connectionist structure in order to find out the most efficient architecture. The monolithic architecture performs recognition of all macro classes and features by a simple neural network using the standard backpropagation learning procedure. We must recall that each AR-TDNN component is trained independently using Nguyen initialization and conditions [7].

The results given in Table 1 show that either serial and parallel architectures with respectively 15% and 12% of mean error rate, are more efficient than the standard backpropagation-based system with 30% error rate. In the serial architecture, nasals and liquids are detected with an error rate of 26% and 28% respectively. We have remarked that these cases of bad detection are generally due to the failure of previous levels. In the case of fricatives parallel and serial structures have the same accuracy with an error rate of 10% while the simple back propagation-based system performs 12% less than first systems. Among all the errors made for the fricatives, approximately 50% were due to rear fricatives. For plosives the shortness of /q/ and /?/ velar and glottal sounds involve errors due to deletion or

insertion. The monolithic system achieved the identification of geminated consonants with a relatively low success rate of 61%. At the opposite, serial and parallel neural systems with a correct rate of approximately 88% increase dramatically the recognition rate of these complex phonemes. This appears clearly in the case of long/short discrimination of vowels where an improvement of 30% in the recognition rate is observed.

Class / System	Long/brief Vowels	Plos.	Fri.	Nas.	Liq.	Emp.	Gem.
Single NN	38.7	24.3	22.8	26.6	28.2	30.0	39.1
Serial NN	8.1	16.8	10.9	16.4	20.9	15.8	11.2
Parallel NN	8.9	13.3	10.2	12.3	15.5	14.7	11.8

Table 1: Error rate (%) of serial and parallel neural network structures and single neural network. (Vow: Vowel, Plos: Plosives, Fri: Fricatives, Nas: Nasals, Liq: Liquids, Emp: Emphatic, Gem: Geminate).

Parallel connectionist structure remains the most reliable system. Thus, in the hybrid structure presented in section 2.1, we incorporate a parallel architecture of SNNs. We compare hybrid system performances to a baseline HMM-based system. This latter is a left-to-right model with three states and has a single Gaussian density per state. It is parameterized with 24 coefficients derived from Caelen ear model. The results obtained by the two systems are presented in figure 4. These results concern the six vowels, the geminated and emphatic consonants. The analysis of the results revealed that hybrid configuration is more accurate in all cases of complex phonemes. We found that this system achieved 90% accuracy, which represents 10% fewer errors than the baseline system. Concerning the standard HMM, we noticed that even if it was relatively effective to detect long vowels, it failed dramatically in the detection of short vowels. An unbalance of performances which can reach 15% is observed. The same phenomenon is observed in the gemination case. The redundancy of this trend leads us to conclude, as it was expected, that standard HMM is not capable to perceive relevant phoneme duration changes. On the contrary, the hybrid system achieved successfully this task with the satisfactory rate of 90% for all vowels.

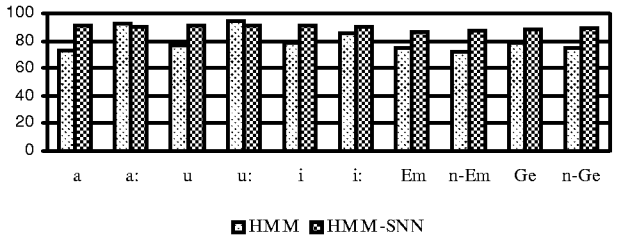

Figure 4: Results of HMM-based baseline system and HMM-SNN hybrid system for long and short vowels, emphatic and geminated consonants.

5. SUMMARY

We have presented the identification results of Arabic macro-classes and features by systems using a hierarchy of neural networks. These systems are composed of sub-neural-networks carrying out binary discrimination sub-tasks. Two types of architecture have been presented: serial structure of experts and parallel disposition of them. The obtained results confirm that the integration of delays and prediction feedback in the used networks (AR-TDNN) give them the ability to capture unstable and relevant temporal component of speech and emphasis. It appears that the parallel disposition of sub-neural-networks constitutes the most reliable system. The parallel mixture of experts was incorporated in an hybrid structure, in the order to enhance HMM accuracy in the case of complex Arabic phonemes identification. Significant improvements were obtained. This suggests that the generalization of the approach by devoting other sub-tasks to adequate experts provides an effective way to the robustness of present ASRs.

6. REFERENCES

1. El-Ani S.H, *Arabic phonology: an acoustical and physiological investigation*, Mouton ed., the Hague, 1970.

2. Cook G.D., Waterhouse S.R., and Robinson A.J, "Ensemble methods for connectionist acoustic modeling", ESCA, Eurospeech'97, Rhodes, Greece: 1959-1962, 1997.

3. Djoudi M., Fohr D., Haton J.P., "Phonetic study for automatic recognition of Arabic", European Conference on speech and technology: 268-271, 1989.

4. Hermansky H., "Perceptual linear predictive (PLP) analysis of speech", Jour. Ac. Soc. Amer. 87 (4): 1738-1752, 1990.

5. Jacobs R.A., Jordan M.I., Nowlan S.J. and Hinton G.E., "Adaptative mixtures of local experts", Neural computation 3(1): 79-87, 1991.

6. Krogh A. and Vadelsby J., "Neural networks ensembles, cross validation, and active learning", Advances in Neural information processing Systems, Vol. 7, MIT press, 1995.

7. Nguyen D., Widrow B., "Improving the learning speed of two-layer neural networks by choosing initial values of the adaptative weights", IJCNN, Vol. III: 21-26, San Diego CA., 1990.

8. Potapova R. K, "The auditory identification of long and short vowels in the Germanic languages", XVth Congress of linguistics: 166-169, Canada, 1992.

9. Russel R.L. and Bartley C., "The autoregressive backpropagation algorithm", IJCNN, Vol II:369-377, 1991.

10. Selouani S.A. and Caelen J., "Recognition of phonetic features using neural networks and knowledge-based system" 3rd IEEE Symposium on Intelligence and systems: 404-411, Washington D.C., 1998.

11. Selouani S.A. and Caelen J., "Experiment in automatic speech recognition of standard Arabic", Proceedings of KFUPM workshop on information and computer science, Dhahran Saudi Arabia: 161-171, 1996.

12. Waibel A., Hanazawa T., Hinton G. and Shikano K., "Phoneme recognition using time-delay neural networks", IEEE trans. ASSP 37: 328-339, 1989.