Arabic Phonetic Features Recognition using Modular Connectionist Architectures

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

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

ABSTRACT

This paper proposes an approach for reliably identifying complex Arabic phonemes in continuous speech. This is proposed to be done by a mixture of artificial neural experts. These experts are typically time delay neural networks using an original version of the autoregressive backpropagation algorithm (AR-TDNN).
A module using specific cues generated by an ear model operates the speech phone segmentation. Perceptual linear predictive (PLP) coefficients, energy, zero crossing rate and their derivatives are used as input parameters. Serial and parallel architectures of AR-TDNN have been implemented and confronted to a monolithic system using simple backpropagation algorithm.

I. INTRODUCTION

The recognition of Arabic is confronted with difficulties inherent to its linguistic particularities. Bearing in mind the advances and the predominance of the automatic methods, it can be inappropriate to hypothesize the correlation of a system performance with linguistic particularities. The reliability reached by the present automatic speech recognition (ASR) systems (HMM, neural networks and hybrids) permits to assume that performances are conditioned by the only availability of conveniently segmented and labeled corpus [3][9][24]. In the Arabic language particular case, this statement is from far not tested. Indeed, the available products in the commerce have turned up with inconsistencies as far as their capability to overcome the problems due to the strong inflection of the language. Designers of the systems dedicated to the Arabic language are unanimously observing that emphasis, gemination and vowel's lengthening[1] constitute the main root of failure [6][7].
The presented approach 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 those experts. We are inspired in this method on principle of divide and conquer where a hard problem is broken up into a set of easier problems to solve [23][26]. Consequently, learning phase does not need a large number of cycles and each sub-network is independently trained to recognize a specific feature.

II. PROBLEMATIC OF ARABIC PHONETIC FEATURES DETECTION

The originality of the Arabic phonetics is mainly based on the relevance of lengthening in the vocalic system and on the presence of emphatic and geminated consonants. These particular features play a fundamental role in the nominal and verbal morphological development.

A. Emphasis

The emphasis is in the case of the Semitic languages a phonetic feature characterizing consonants. On the

[1] In Arabic, the vowel duration is semantically relevant.

articulatory plan, these consonants are achieved in the rear part of the oral cavity. During their realization the tongue root is carried against pharynx. There are four emphatic consonants in the Arabic language: 2 plosives: /t̲/, /d̲/ and 2 fricatives: /δ̲/, /s̲/. In the example of the two words /nasaba/ (imputed) and /nas̲aba/ (erected), an emphatic vs. non-emphatic opposition is observed on /s/. The ambiguousness is far than can be raised by the present systems even those providing linguistic processes.

B. Gemination

This particular feature compensates the poverty of the Arabic vocalic system. The geminated consonant arises by sustaining the plosive closure. In the example of the words /faʃala/ (he failed) and /faʃ:ala/ (he thwarts), the opposition resides in the gemination of /ʃ/ fricative. Through this example, we measure the importance and the difficulty to perform this feature detection by automatic systems. In the classical approach, the gemination is simply considered as the doubling of consonant duration. We proposed in [2][19] to compute the tense/lax feature in order to detect the gemination. So, it is not necessary to integrate an explicit time index in the acoustical analysis. Consequently ASR systems acquire more robustness in the case of speech rate changes.

III. PRE-PROCESSING BY AUDITORY MODELS

Two Auditory models are used: Caelen ear model [4] for homogenous phone segmentation and PLP (perceptual linear predictive) [10] technique as acoustical analyzer. This type of pre-processing is privileged regarding to their capabilities to improve recognition performances of ASR systems. 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. We have shown in [20] that these static acoustic indicative features are very relevant to characterize the Arabic phonemes. A delta coding of these acoustic indicative features is done in order to find out their variation and to perform phone segmentation. A function computes the sum of absolute outputs of delta coders. In such way, it quantifies the discontinuity between two successive frames. If this amount is over a time variable threshold, a mark is attached to current frames. The frames between two successive marks are considered as a homogenous phone. Over each phone, an average of PLP coefficients combined with energy (En), zero-crossing rate (ZCR) and their derivatives are calculated. This multi-component vector is used as input in the networks. This type of acoustical parameters have been retained because it gives the best cross validation results as it is shown in [21].
Automatic labeling of these phones is performed by connectionist architectures described in the following sections.

IV. MODULAR CONNECTIONIST ARCHITECTURES

Jakobs and Jordan [11][12] introduced a hierarchical structure of experts in order to solve problems of non-linear regression. In the case of spontaneous telephonic speech, this structure was revealed more efficient than traditional monolithic networks [5]. 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 sub-neural networks

A. Auto-regressive time delay neural networks (AR-TDNN)

The temporal component of the speech signal is difficult to capture by a connectionist system [27]. Besides this, in the particular case of the Arabic, a temporal alignment can be prejudicial. 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.
The tone variations characterizing emphatic and geminated consonants and long vowels free us of the explicit calculus of duration parameter. As it is shown in different studies [1][6][13][17] these variations influence the phonetic context of the phoneme to recognize. 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 [18] showed that the use of an autoregressive version of backpropagation algorithm (AR-backpropagation) gives the neural network a memorization capacity in the case of temporally unstable event identification. In the version we propose, in the network input layer, a delay component similar to the one used by Waibel TDNN (Time Delay Neural Networks) [25] is integrated.
In the detection of relevant phoneme duration, this combination increases the capacity of the system to discern the phonological length even in a strong coarticulation context.
The AR version of backpropagation extends the classical learning algorithm to discrete time varying systems by including feedback weights.

In the neuron we use, a delay component is added at the input layer in order to boost the capacity of the classifier to deal with temporally distorted phonemes. Figure 1 shows this new type of neuron.

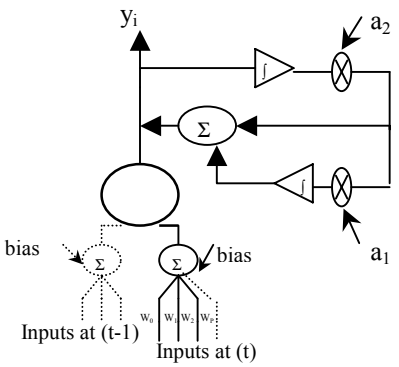

Figure 1- A neuron of the AR-TDNN

y_i 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 coefficients $a_{i,n}$ are parameters to optimize. Initial conditions have been chosen as it is proposed by Nguyen in [16].

The identification phase is performed over homogeneous phones. If a phone of the target-phoneme appears in the speech continuum, the network activation arises gradually in one of its two outputs. Besides the detection/classification task, the configuration of the AR-TDNN permits the learning of the phoneme context. This property is more suitable in the case of the temporally unstable phonemes where the duration plays a fundamental role. The AR component of the network gives it the ability to recognize series of sequences in a certain context.

The discrimination of emphatic, geminated consonants and the long short vowels is performed using the previous values stored in the delays as well as in the feedback.

In the case of emphasis detection/classification example (EMPHA_NET network), as it is illustrated in the figure 2, the task is to learn to recognize this sequence: LCE-EMP-RCE: LCE is the left phonetic context of the emphatic (noted EMP) and RCE is its right phonetic context of the emphatic. EMPHA_NET receives three input token at a time t and it must detect an emphatic sequence from any other sequence combination.

The learning consists in setting at the high level (+1) the first output when the end of the LCE-EMP-RCE sequence is attained. Low level (-1) is set otherwise. The second output is set at the high level if a scrolling (stream) of non-emphatic phone sequences is observed. An autoregressive order of 2 is chosen and a delay of 2 frames is also fixed. These lower values of delay and order are justified by the fact that phones are used instead frames. An important advantage of this approach resides in the stability of AR nodes.

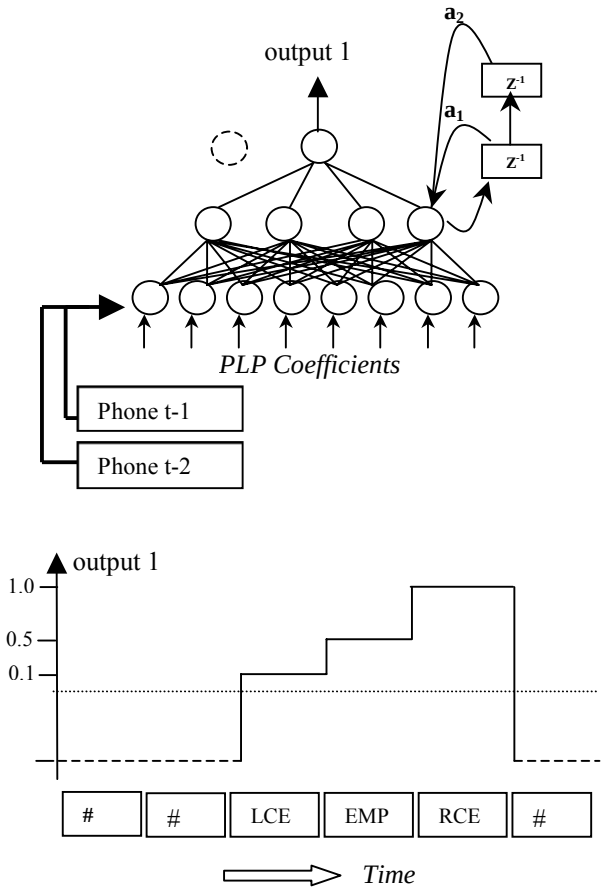

Figure 2- AR-TDNN phone-based identification process

B. Serial structure

This structure is made up of serial disposition of AR-TDNN experts. Two types of unknown sequence classification are accomplished. The first is a rough classification performing macro-classes detection such as vowels, fricatives, plosives, nasals and liquids. The second classification, finer, discriminates between long and brief vowels and detects the gemination and the emphatic feature. The progress in the structure (cf. Figure 3) is conditioned by the

opening of the logical gates (and/or). According to the activation of the two outputs for a given expert network, the process starts the identification of the emphatic feature or the geminate feature if the macro-class is detected. Otherwise an activation of the contiguous network is operated. A failure of the overall system is accounted if the last network is reached without any discrimination.

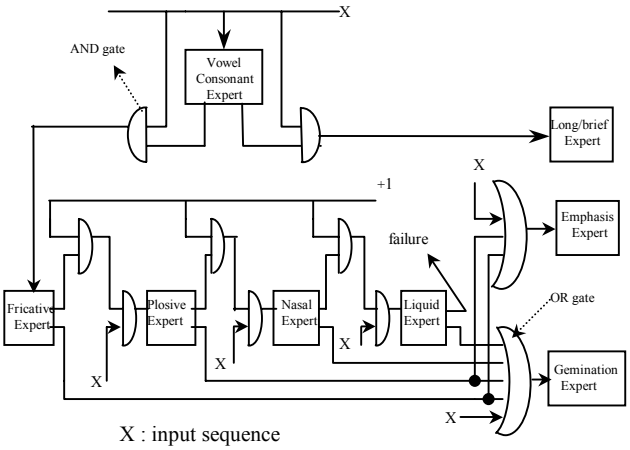

Figure 3- Serial connectionist structure for the identification of Arabic phonetic macro-classes and features.

The hierarchical structure (pipeline) of this system seems inadequate since the networks located deeply (the far right in the Figure 3) are penalized. An architecture putting networks in the same level of competence seems less constraining. However, we can note that the experts of the highest levels (the far left in Figure 3) justify their position by their high success rate during the cross validation [14][22]. So, the penalization of the following levels is minimized.

C. Parallel structure

This configuration is established without any fixed condition concerning the relative individual performances of each expert. Experts have the same activation potential when a sequence to identify is presented at the input (cf. Figure 4). This property allows a flexible use and permits to avoid the failure case. During the recognition phase it is required from each expert only a specialization in the scanning of one phonetic class. A logic module based on pruning and on *a priori* rules performs a final identification (decision).

D. Pruning procedure

In the parallel configuration, a pruning module manages cases where several experts are simultaneously activated. The pruning procedure is summarized as follows:
Considering S_{j1} and S_{j2}.
S_{j1} being 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.
S_{j2} being 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 eliminated from the competition. In the case of positive distances, the corresponding classifiers are retained to make the decision about recognized pattern. The identified class is the one having its corresponding network (winner) 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.

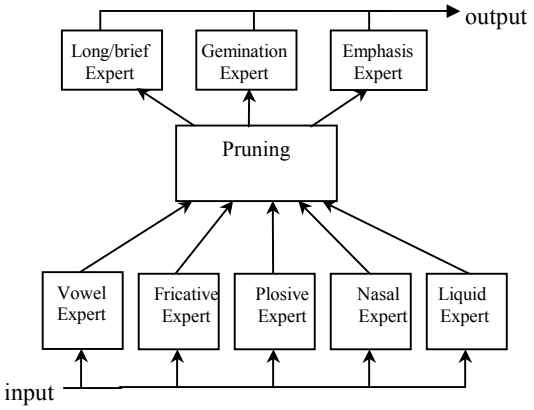

Figure 4- Parallel connectionist structure for the identification of Arabic phonetic macro-classes and features.

V. RESULTS AND COMMENTS

The test corpus has been pronounced by six Algerian native speakers (3 men and 3 women). These speakers have participated to the learning and the cross validation. The stimuli are composed of 60 VCV utterances and 50 phrases.

The test concerns:

- fricatives: /f/, /s/, /s̲/, /z/, /h/, /ħ/, /ʃ/, /θ/, /χ/, /ð/, /ð̲/, /γ/, /ε/, /ʒ/;
- plosives: /t/, /t̲/, /k/, /b/, /d/, /d̲/, /q/, /ʔ/;
- liquids: /l/, /r/;
- nasals: /m/, /n/;
- short vowels: /a/, /u/, /i/;
- long vowels: /aa/, /uu/, /ii/.

As a whole, the test has concerned 3724 vowels 1197 fricatives, 1089 plosives, 573 nasals and 413 liquids. The semi-vowels are assimilated to their corresponding vowels. Additional 397 geminated consonants have been tested [15].

Either serial or parallel architectures realize mediocre scores in the particularly case of /ʔ/ and /q/ glottal and velar plosives. The rear fricatives (/h/, /ħ/, /γ/, /ε/) also cause problems. Their shortness and their sensibility to the utterance speed (co-articulation effects) make them merged into the vocalic context. We can conclude that VCV (Vowel-Consonant-Vowel) utterances are unfavorable material for the learning of this type of sounds (the omission percentage is very high).

We have noticed the total failure of all systems in the identification of emphatic feature for /d̲/ consonant. The explanation does not reside in the difficulty inherent to the consonant's acoustical proprieties, but rather in the capability of the speaker to pronounce it correctly. In fact, in a VCV context, it is very difficult to keep the emphatic character of /d̲/ and more often, it is its opposite by this feature (/d/) which is achieved[2].

Considering the obtained results (cf. Table 1), it seems clear that the serial and parallel configuration with respectively 15 % and 13 % of mean error rate, are more efficient than the classical simple backpropagation system with 30 % error rate.

In the serial architecture, nasals and liquids are respectively correctly detected in 84 % and 79 % of cases. We have remarked that these cases of bad detection are generally due to the failure of previous levels.

For all macro-classes and features, the difference between scores is always in favor of the mixture of experts. Parallel connectionist structure remains more reliable than serial structure with a positive difference of 5%, 4%, 3% and 1% respectively for liquids, nasals, plosives and emphatic consonants.

In the case of fricatives parallel and serial structures have the same accuracy with a rate of 90% of correct identification while the simple backpropagation based system performs 12% less than first systems.

The monolithic system achieved the identification of geminated consonants with a relatively low rate of 61%. At the opposite, serial and parallel neural systems with a same correct rate of approximately 88% increase dramatically the recognition rate of these complex phonemes.

These results confirm that the integration of delays and prediction feedback in the used networks give them the ability to capture unstable and relevant temporal component of speech. 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 Vow.**	**Plos**	**Fri.**	**Nas.**	**Liq.**	**Emp.**	**Gem.**
Simple 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 simple backpropagation system. (Vow:Vowel, Plos:Plosives, Fri:Fricatives, Nas:Nasals, Liq:Liquids, Emp:Emphatic, Gem: Geminate).

VI. CONCLUSION

We have presented the identification results of Arabic macro-classes by two systems based on a mixture of neural experts. 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.

Our objective is to test the ability of autoregressive time delay neural networks (AR-TDNN) to detect Arabic complex phonemes. In regard of obtained results, we can conclude that parallel and serial structures of AR-TDNN overpass monolithic configuration. The parallel disposition constitutes the most reliable system. The proposed mixture of neural experts approach is also advantageous by the fact that it eases the learning because the binary discrimination does not need a large number of cycles.

[2] this defect is mainly due to the characteristic of the Algiers regional accent.

The generalization to the identification of other features such as speaker gender and prosodic features may constitute a simple and powerful way to improve ASR systems performances.

REFERENCES

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

[2] B. Boudraa, and S.A. Selouani**,** "Matrices phonétiques et matrices phonologiques arabes", proceedings XXèmes JEP, Tregastel, France, 1994, pp. 345-350.

[3] H. Bourlard, N. Morgan, *Connectionist speech recognition: A hybrid approach*, Kluwer Publisher, 1994.

[4] J. Caelen, and H. Tattegrain "Le décodeur acoustico-phonétique dans le projet DIRA", proceedings XIIèmes JEP, Nancy,1988, pp 115-121.

[5] G.D. Cook, S.R. Waterhouse, and A.J Robinson, "Ensemble methods for connectionist acoustic modeling", ESCA, Eurospeech97, Rhodes, Greece, 1997, pp. 1959-1962.

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

[7] O. Emam, "Speech recognition of Arabic" Technical notice on www-page of IBM Cairo scientific center, 1997.

[8] P. Gallinari., S. Thiria, and F. Badran, "On the relations between discriminant analysis and Multi-Layer Perceptrons". neural networks Vol 4, 1991, pp. 349-360.

[9] J.P. Haton, "Modèles neuronaux et hybrides en reconnaissance de la parole: état des recherches", fondements et perspectives en traitement automatique de la parole, éditions H. Méloni, 1995, pp. 139-154.

[10] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", Journal Acouc. Soc. Am. N° 87 (4), 1990, pp.1738-1752.

[11] R.A. Jacobs "Methods for combining experts probability assessments", Neural computation, Volume 7(5), 1995, pp. 867-888.

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

[13] R. Jakobson, G.M. Fant, and M. Halle, *Preliminaries to speech analysis: The distinctive features and their correlates*, MIT press, Cambridge, 1963.

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

[15] M. Mrayati., *Statistical studies of Arabic roots*, Applied Arabic linguistics and signal and information processing, Hamshire publishing, 1987.

[16] D. Nguyen, B. Widrow, "Improving the learning speed of two-layer neural networks by choosing initial values of the adaptative weights", International Joint Conference on Neural Networks, San Diego, CA, Vol. III, 1990, pp.21-26.

[17] R.K. Potapova, "The auditory identification of long and short vowels in the Germanic languages", XVth International Congress of linguistics, Canada 1992, pp. 166-169.

[18] R.L. Russel, C. Bartley, "The autoregressive backpropagation algorithm", International Joint Conference on neural networks Vol II, 1991, pp. 369-377.

[19] S.A. Selouani, and J. Caelen, "Experiments on Arabic phone recognition using automatically derived indicative features", IVth International Symposium on Signal Processing and its Applications, Gold coast, Australia, 1996.

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

[21] S.A. Selouani and J. Caelen, "Recognition of phonetic features using neural networks and knowledge-based system: a comparative study," 3rd IEEE Symposium on Image, Speech, Natural Language Systems, Washington D.C., 1998, pp.404-411.

[22] M. Stone, "Cross validatory choice and assessment of statistical predictions", Journal of the royal statistical society series B, Volume 36, 1974, pp. 111-147.

[23] K. Takuya., and T. Shuji, "Simplified sub-neural-networks for accurate phoneme recognition', proceedings ICSLP, Yokohama Japan, 1994, pp. 1571-1574.

[24] J. Tebelskis, *Speech recognition using neural Networks*, PHD thesis, CMU, Pittsburgh, Pennsylvania, 1995.

[25] A.Waibel, T. Hanazawa, G. Hinton, K. Shikano, "Phoneme recognition using time-delay neural networks", IEEE trans. on ASSP, Vol. 37(3), 1989, pp.328-339.

[26] S.R. Waterhouse and G.D. Cook, "Ensembles for phoneme classification", Advances in Neural information processing Systems, Volume 9, MIT press, 1996.

[27] R.L. Watrous, and L. Shastri, "Learning phonetic features using connexionist networks: an experiment in speech recognition", proceedings ICASSP Vol 4, San Diego California, 1987, pp. 381-388.