Recognition of Arabic Phonetic Features Using Neural Networks and Knowledge-Based System: a Comparative Study

Sid-Ahmed SELOUANI
L.C.P Lab. Institute of electronics,
U.S.T.H.B, Algiers, Algeria
e-mail:parole@ist.cerist.dz

Jean Caelen
CLIPS/IMAG Lab.
U.J.F, Grenoble, France
e-mail:jean.caelen@imag.fr

Abstract

This paper deals with a new indicative features recognition system for Arabic which uses a set of a simplified version of sub-neural-networks (SNN). For the analysis of speech, the perceptual linear predictive (PLP) technique is used. The ability of the system, has been tested in experiments using stimuli uttered by 6 native Algerian speakers. The identification results have been confronted to those obtained by the SARPH knowledge-based system. Our interest goes to the particularities of Arabic such as geminate and emphatic consonants and the duration. The results show that SNN achieved well in pure identification while in the case of phonologic duration the knowledge-based system performs better.

1. Introduction

The essential characteristics of neural networks are generally their learning capability from examples, their adaptability, their strength to the spoilt or missing data and in speech recognition, their discriminating power in order to divide the acoustic parameter space in phonetic classes. Numerous implementations of these networks have been dealt within literature [7][8][10][23]. The most used structure is the multi-layer perceptron (MLP). This type of network is capable to learn and generalize on complex and non-linear relationships linking the acoustic vector space and the phonetic classes that we desire to recognize.

In this paper, we are concerned with the automatic recognition of phonetic macro-classes of the Arabic language by multi-layer sub-neural-networks (SNN). We will focus our experimentation on specific Arabic phonetic aspects. These phonetic aspects are: long and short vowels, the emphasis as well as the gemination. We are comparing the results obtained by this purely automatic approach to the one which uses the phonetic knowledge expressed by rules of the SARPH system (Arabic Recognition System using Phones).

2. The SARPH system overview

SARPH [19][20] is an analytical system of recognition (based on rules). It is organized around a segmentation module in homogenous phones and finite state networks (FSN) for the macro-classes phonetic identification of the Arabic language. There are five macro-classes: vowels (V), fricatives (S), plosives (Q), nasals (N) and liquids (L).

The SARPH structure is analog to the one adopted in the DIRA system for the French language [5]. The following sections are presenting briefly the different modules of SARPH.

2.1. Acoustic attributes and distinctive features

The energy as well as the zero crossing rate (ZCR) of the signal are calculated on 10 ms frames. The fundamental frequency is evaluated and corrected by the modified ambiguity technique [18].

The formants extraction is operated on the linear predictive coding (LPC) spectrum. A 24 channel spectrum (24 coupled filters) using Caelen ear model [4] is generated. From a particular linear combination of the outputs of the channels, 7 cues are derived: acute/grave, open/close, diffuse/compact, sharp/flat, mat/strident, continuous/discontinuous and tense/lax.

We proved in [2] that these static acoustic indicative features are very relevant to characterize the Arabic phonemes.

2.2. Segmentation and identification strategy

A delta coding of the acoustic indices is done in order to find out their variation. A function which is the sum of absolute outputs of delta coders is evaluated. It quantifies in such way the discontinuity between two successive frames. If this amount is over a threshold, which is variable in time, a mark is attached to the current
the frames. The frames between two successive marks make an homogenous phone. For each phone, the mean, the maximum and the minimum values of each parameter are calculated. The static and dynamic acoustic indices are coded (in a non-linear way) in 5 levels: --, -, 0, +, ++. An order relation exits between these 5 coding degrees. Thus, TL++ means very tense and TL-- very lax. Other acoustical parameters such as the friction cue and the vocalic indicative feature, which provide some robust information, are calculated. They constitute, with other parameters, the knowledge base of SARPH. To each macro-class is associated a phonetic network, representing the knowledge on the macro-class. This phonetic network is applied independently on the sequence of phones. A network consists of a set of states and of a set of transitions (cf. figure 1). The states represent all the possible realizations of the various acoustic phases of the phonetic macro-classes. To each transition (or arc) is associated a constraints list (rules) to check, an action list to operate in any case of success and at the end a score for each passage by the transition. A phone may be labeled by one or several networks as it can be rejected by all. The number of phones is not known first and the labeling of a phone of N rank is done only if the N-1 preceding phones have been labeled.

The access is than operated sequentially and justifies the use of a chained list for the modeling of the phonetic network. Thus, each rank of the list represents a phone and contains the information relative to this one. The chained linear list is bi-directional in order to permit the backward action when the deep exploration of the network is done. Figure 1 gives an example of the progress in the SARPH vowel network. In this case, if for the current phone, an “Onset” is noticed, we will make an attempt to associate to the following phone the same phase or the “semi-standing state” or “standing state” phases. We do the same process for the following phones, taking into account the permitted passages in the network until we get out of it in any case of solution. Many ways of labeling are possible thanks to the reverse process. Any solution offering the highest score will be validated.

The labeling process of SARPH is built into two parts. The first one deals with the localization of the macro-classes with the help of the FSN. The context constraints are supposed realized. The second one is the accurate recognition of the emphasis and the gemination for consonants and of the temporal opposition for vowels. For all consonants networks, the average of flat/sharp indicative feature is computed over all the phones. One of 5-level non-linear code values is assigned to the feature average. In this way, if the code ‘++’ is assigned, it means that the consonant is very flat and hence it will be labeled as emphatic. The same mechanism is carried out for the vowels where, in addition to the macro-class detection, a complete classification of the six Arabic vowels is performed. The long/brief opposition is detected by means of the encoded tense/lax feature.

![Figure 1. Finite state network of vowels.](image)

3. Macro-class recognition by neural networks

The connectionist hierarchical structure that we propose (see figure 2) consists of a simplified neural network set to which a classification sub-tasks have been assigned in order to globally identify macros-classes and Arabic phonetic features. The training of each “specialized unit” is operated on all the learning corpus. The optimization of the parameters, such as the number of the network cells, the learning constant, the number and the quality of the inputs as well as the iteration number is done individually on each sub-network. It is by the means of a cross validation that the adjusting of all these parameters is realized. The principle is to notice their success rate for different values of the parameter to optimize. This task is very easy thanks to the simplicity of the sub-network to train.

3.1. Acoustic analysis

Different acoustic analyses have been tested. The purpose is to determine the one which will give the best
compromise between the learning speed and the generalization capability. For this purpose, a cross validation corpus has been established. It consists of 414 vowels, 246 fricatives, 214 plosives, 106 nasals and 101 liquids. The sub-network validation has been operated by using the linear predictive coding cepstral coefficients (LPCC), the perceptual linear predictive coefficients (PLP) [11], the energy (En) and the zero crossing rate (ZCR) as well as their first derivatives (of En and ZCR). The validation is operated on the vowel classification sub-network. The assigned task is the classification of the short vowels of the Arabic language. The PLP coefficients combined with the Energy, ZCR and their derivatives are those which give the best result as it is shown in table 1.

<table>
<thead>
<tr>
<th>Acoustical analysis</th>
<th>Number of inputs</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>36 LPCC</td>
<td>36</td>
<td>89 %</td>
</tr>
<tr>
<td>15 PLP</td>
<td>15</td>
<td>87 %</td>
</tr>
<tr>
<td>36 LPCC+3ZCR+3 En</td>
<td>42</td>
<td>90 %</td>
</tr>
<tr>
<td>15 PLP+3ZCR+3En</td>
<td>21</td>
<td>91 %</td>
</tr>
<tr>
<td>36 LPCC+3ZCR+3En+3dEn+3dZCR</td>
<td>48</td>
<td>91 %</td>
</tr>
<tr>
<td>15 PLP+3ZCR+3En+3dEn+3dZCR</td>
<td>27</td>
<td>92 %</td>
</tr>
</tbody>
</table>

Table1. Performance of the vowel sub-network with different kinds of acoustical analysis.

The same experimentation conditions (iteration number, learning constants, ...) have been used during the various experiments. The validation results show a distinct superiority of the auditory modeling opposed to the classical modeling. A gain in the learning time (reduction of the number of input units) is also noticed. Since the size of the network has been also diminished, the number of weights and biases to store will decrease strongly.

3.2. Input normalization

The dynamic temporal management by the MLP-type networks remain their main weak points [23]. The networks are not capable to manage the temporal distortions which have not been learned. In the speech case, each segment consists of a variable number of frames. In the case of a static classification where the network architecture is stiffened, this difficulty must be discarded [25]. The system proposed here avoids this difficulty: it divides each segment into three intervals (onset-stabilization-end) on which we calculate the mean of the acoustic vectors. The first and the last interval include the contextual information (right and left). In the case in which the division result is not an integer, the middle interval is extended by the same number of remaining frames (the stable phases are favored). Consequently, the number of parameters presented at the input is always fixed whatever the length of the segment is. It will be always equal to three times the size of acoustic vector of the frame.

If m is the number of frames by a segment and p the acoustic vector size, then:

\[ n_1 = n_3 = m/3 \]

and

\[ n_2 = m/3 + (m \mod 3) \]

\( n_1, n_2, n_3 \), being respectively the number of frames on the first, the second and the third interval on which is operated the average of the parameters vectors. The inputs \( E_j \) presented to the sub-network are given by the following expression:

\[
E_j = \sum_{k=1}^{m} C_{ij} \quad \text{for} \quad j = \{0, \ldots, P - 1\}
\]

\[
E_j = \sum_{k=1}^{2} C_{ijk} \quad \text{for} \quad j = \{0, \ldots, P - 1\}
\]

\[
E_j = \sum_{k=1}^{3} C_{ijk} \quad \text{for} \quad j = \{0, \ldots, P - 1\}
\]

\( C_{ikj} \): Kuple component of the vector of the frame i, participating to the input j calculation. The number of input units will be 3p.

3.3. The identification task

During the learning, a flow of data segmented in macro-classes is presented at the network input. Since the learning is supervised, the data base is also labeled in macro-classes (V, S, Q, N, L). During the identification phase, we will ensure that the segments to classify are the same as those presented to SARPH.

Two types of classifications of unknown sequences are done. The first one is a rough classification whose the objective is to detect the macro-classes (V, S, Q, N, L). The second classification is more refined and it tries to detected the emphasis and the gemination on the consonants. A refining of the vowel detection is also operated by a specialized network. After having been submitted to a temporal normalization the acoustic vectors are first injected in the V network (noted 1 in figure 2) in charge of the rough discrimination of vowel/consonant. Then, they progress in the successive networks S, Q, N and L (from left to right and noted respectively 2, 3, 4, 5 in figure 2). According to the activation of two outputs of a given network, the process...
stops if the macro-class is detected. Otherwise an activation of the adjacent network is operated. A failure of the global system is counted if the last network is reached without any discrimination. When the Q and S macro-class is detected, the emphasis network (Noted 2’) is activated. The gemination network (noted 3’) is activated when a consonant is detected (Q, S, L, N). The disposal of the sub-networks in the global system is discussed in section 3.5.

3.4. Architecture and learning constant

The same material (learning and validation corpus) is used to determine the optimum number of the hidden units. We notice that the performances are decreasing from a certain threshold linked to the number of hidden units. For example, this value is approximately of 38 for the fricative networks. We used here at the input both LPCC coefficients, ZCR, energy and their derivatives. Table 2 gives the success rate in validation for each sub-network with a proper optimum architecture. The inputs being the PLP coefficients, the ZCR, the energy and their first derivatives. These results permit a posteriori to organize into hierarchy the sub-networks according to their performances.

Another problem resides in the choice of the learning constant of the gradient (generally noted $\eta$). A sub-network validation has been done for various values of $\eta$ with the same number of iterations. The learning time depends strongly on the choice of this constant. The results given in figure 3 show that a value of 0.4 for $\eta$ is a good compromise between the learning time and the generalization ability.

![Hierarchical connectionist structure for the Arabic macro-classes detection.](image)

**Figure 2. Hierarchical connectionist structure for the Arabic macro-classes detection.**

<table>
<thead>
<tr>
<th>sub-network</th>
<th>Architecture</th>
<th>failure</th>
<th>Success</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel/consonant</td>
<td>27-15-2</td>
<td>12</td>
<td>634</td>
<td>98%</td>
</tr>
<tr>
<td>Vowel /a/,/u/,/i/</td>
<td>27-25-3</td>
<td>17</td>
<td>397</td>
<td>96%</td>
</tr>
<tr>
<td>Fricatives</td>
<td>27-18-2</td>
<td>20</td>
<td>226</td>
<td>92%</td>
</tr>
<tr>
<td>Plosives</td>
<td>27-20-2</td>
<td>39</td>
<td>175</td>
<td>82%</td>
</tr>
<tr>
<td>Nasals</td>
<td>27-20-2</td>
<td>22</td>
<td>84</td>
<td>79%</td>
</tr>
<tr>
<td>Liquids</td>
<td>21-15-2</td>
<td>23</td>
<td>78</td>
<td>77%</td>
</tr>
</tbody>
</table>

**Table 2. Macro-classes identification rate obtained in the cross validation.**
The various forms to learn are presented in an alternate way. The approach advantage is that it allows an easy learning because the binary discrimination task does not need a large number of cycles (they are 400). Each sub-network is “initiate” independently of the others. During the recognition phase, a specialization in the scanning of one (and only one) phonetic class among the others is required for the sub-network.

![Figure 3. Effect of learning constant on the error.](image)

3.5. Classification by the global system

The cross validation method, useful for the refining of the network architecture and determining the experiments optimum conditions, permits us also to organize into hierarchy the sub-networks. This is carried out according to their competence to detect the considered macro-class. Indeed, regarding the results of the validation obtained by each sub-network (see table 2), it appears that some tasks are more easily done than others. In the test system, the different sub-networks are solicited in a competence decreasing order.

In the hierarchical structure that we advocate (see figure 2), a simple detection of the vowels is done first. This classification is straightforward thanks to the fact that the Arabic language is essentially consonantal and contains a very few vowels: 3 short vowels (a, u, i) and their long corresponding vowels (aa, uu, ii). A second argument pleading for this discrimination is the fact that the Arabic language does not contain any nasal vowel. Hence, the vowel/consonant confusion due to the nasality does not exist. The architecture and the macro-class disposal in the global system depend on the cross validation results obtained by each sub-network. The structure (organized into hierarchy) of this system may appear less constrained [6][12][21]. However, in the particular case of the Arabic language some arguments plead oppositely for the structure that we have adopted. These arguments are the following:

- fricatives (they are 14) represent by themselves 50% of the Arabic consonantal system;
- plosives (they are 8) are realized in scattered articulation spots (uvular, glottal, velar alveolar and bilabial) [1]. Note that there are no /p/ nor /g/ in the Arabic language;
- liquids and nasal classes contain few elements (2 liquids, 2 nasals).

Thus, the tasks devoted to the superior levels (the left in figure 2) are characterized by three important aspects. These latter justify their position (of superior levels) in the test system, and reduce the penalization of the following levels:

- the appearance frequencies of the macro-classes when speaking (both the 2 first levels: vowels and fricatives deal with approximately 90% of the speech cases) [16];
- the simplicity of their discrimination task for the cases dealt with. These latter are scattered in the articulation spot (for fricatives and plosives) [1];
- their success rate when we operate the cross validation.

When an identification arrives to the last network (the one of the liquids) without a possibility of classification, an ambiguity is announced. In our experiments, this last case is considered as a failure.

Networks specialized in the detection of emphasis and of gemination are added so as to measure the ability of the neural networks to detect this type of phonetic features (very refined) which are specific to the Arabic language. The architecture of these two sub-networks is in the plosive one (27-20-2). Their performances will be discussed in the following section.

4. Results and comments

The originality of the Arabic phonetics is founded for a large part on the relevance of the duration in the vocalic system and on the presence of emphatic consonants. Another determining characteristic is the geminate consonants. This latter has a fundamental part in the nominal and verbal morphological development. These particularities will focus our interest in the comparison of the two macro-class identification systems.

The test corpus has been pronounced by six native Algerian speakers (3 men and 3 women). These speakers have participated to the learning and the cross-validation. The stimuli are composed of 40 VCV utterances and 20 phrases (in standard Arabic), where the phoneme appearance frequencies are respected [16].

The test concerns:
- the 14 fricatives: /ʃ/, /ʂ/, /ʒ/, /z/, /θ/, /θ/, /ʃ/, /ʃ/, /ʃ/, /θ/, /θ/, /θ/;
- the 8 plosives: /t′/, /t′/, /t′/, /t′/, /θ′/, /θ′/;
- the 2 liquids: /l′/, /l′/;
- the 2 nasals: /n/, /n/;

The various forms to learn are presented in an alternate way. The approach advantage is that it allows an easy learning because the binary discrimination task does not need a large number of cycles (they are 400). Each sub-network is “initiate” independently of the others. During the recognition phase, a specialization in the scanning of one (and only one) phonetic class among the others is required for the sub-network.
As a whole, the test has concerned 852 vowels, 384 fricatives, 248 plosives, 164 nasals and 168 liquids. The semi-vowels are assimilated to their corresponding vowels. An additional sequence of 108 VCV utterances whose consonant is a geminate fricative has been tested. The choice to attempt the detection of the geminate consonant independently of the detection of the other consonants is deliberate, since in the Arabic consonantic system there is no geminate consonants (to use them in the general corpus will unbalance phonetically this one). The number of emphatic consonant (fricatives and plosives) tested is 104.

4.1. Evaluation of SARPH

We have shown in [18][19][20] that SARPH permits to confirm by experimentation three facts concerning the vocalic and consonantal system of the Arabic language, as theoretically already announced by Jakobson [14], El Ghazeli [9], Bonnot [3] and Boudraa & Selouani [2]:
- the acoustic indicative feature tense/lax, calculated and then coded by the system permits to make a distinction short vowels/long vowels;
- the emphasis aspect is detected thanks to a rule managing the flat/sharp cue;
- a distinction between a geminate consonant and its simple homologous is possible thanks to the tense/lax cue.

We must recall here that SARPH has permitted to validate experimentally (by the identification) some important phonetic aspects of the Arabic language. However, we have to note that the basis of its knowledge requires very often empirical thresholds which depend on the experimentation conditions (microphone, signal on noise ratio, utterance speed, etc.). This requires a rigorous management of an important number of parameters. The results obtained by SARPH (histograms per speaker) are shown in figure 4.

4.2. Evaluation of the sub-networks

For the plosives, particularly /?/ and /q/, mediocre scores have been realized. The rear fricatives (/h/, /?, /γ/, /ε/) cause problems. Their shortness and their sensibility, to the utterance speed (coarticulation effects) make them merged into the vocalic context. The VCV utterances are unfavorable material for the learning of this type of fricatives because the omission percentage is very high. As it is the case for some plosives (glottal and velar), we must attach importance to the segmentation in the learning and test phases. The nasality is detected in 85% of cases. The cases of bad detection are generally due to the previous levels. The average scores obtained individually on each of the six speakers for the different macro-classes are given in figure 5.

4.3. Detection of emphasis

The emphasis is a phonetic feature characterizing 4 consonants, 2 plosives : /t/,/d/ and 2 fricatives : /??/, /s/). These consonants are articulated in the rear part of the mouth cavity, the root of the tongue is set back against the rear pharyngeal wall and a curve of the tongue is noticed. Acoustically, they are characterized by the elevation of the first formant transition and the fall of the second formant transition of the following and previous vowel.
As it is shown in figure 6, the correct detection rate is 81% for SARPH and 86% for the SNNs. We remark the total failure of the two systems in the identification of this feature for the /d/ consonant. The explanation does not involve an inherent difficulty of the acoustic properties of the consonant, but rather it is due to the capability of the speaker to pronounce correctly. In VCV context, it is very difficult to keep the emphatic character of /d/. Very often, it is its opposite by this feature (/d/) which is realized. This alteration is a typical characteristic of the Algiers regional accent.

4.4. Detection of gemination

The traditionalist school of the Arab grammarians [17] consider that the gemination is a doubling of the consonant (when we pronounce a consonant in a strained way, this involves this feature on it). We have shown in [2], confirming the Bonnot theses [3] that the tense/lax cue can detect the gemination. This has permitted to SARPH to detected it with a successful rate of 77%. At the opposite, the neural system has turned out to be a little failing with a rate of 68% (cf. figure 6). We think that it is due to the fact that the duration parameter which characterizes this feature is not integrated by this system type.

4.5. The long/short vowel distinction

In the SARPH system, the long/short vowel distinction is realized with 78% success cases. A complete vowel detection is performed by using an algorithm of formants tracking (on the LPC spectrum). For SNNs, we have tried to do this discrimination by adding to the system of figure 2, at the input of the vowel sub-network, a specialized network in that classification. Less than 68% of the success rate has been reached (cf. figure 6).

The duration parameter is very important in the Arabic language, it characterizes not only the vowels but also the geminate consonants. This characteristics counterbalances the poverty of the Arabic vocalic system. At the grammatical level as well as the semantic one, this parameter is fundamental. As far as this aspect is concerned, a double problem is to solve in the automatic recognition of the Arabic speech : we must detect the extended phonemes (long) while assuring that this extension is relevant i.e. by distinguishing the extension which are due to the speaker utterance, to a particular accent of the speaker, etc.

For instance, the two words : /jamal/ (camel) and /jamaal/ (beauty) are different only by the extension of the last vowel. We require from the recognition system to detect the two vowels without altering the temporal property. The finite state networks used by SARPH (the stagnancy duration can be known) managed by phonetic rules are much more adapted to this task than the neural networks. Even if the neural networks integrate the temporal component (time delay neural networks, recurrent,...), they perform a temporal alignement, and therefore penalize the detection of duration relevance.

Figure 6. Identification rate of macro-classes by SARPH and the SNNs.

5. Conclusions and perspectives

We have presented the identification results of Arabic macro-classes by two systems having completely different strategies. The first one is based on phonetic rules. The second is organized on a structure of sub-neural-networks to which some binary discrimination tasks (∉ the macro-class or ∉ to the macro-class) have been assigned. Our major interest in the second system is the simplicity, the ease of its learning and the versatility of use.

Our experiments have concerned the Arabic language, especially the ability of the automatic systems to operate a “blind” classification detecting the aspects as subtle as the gemination, the emphasis and the relevant extension of vowels. These sytems have been confronted to those which operate an “intelligent classification” monitored by a human expert (SARPH knowledge). Of course, the database is reduced but it is sufficient for the purpose of this article. According to the obtained results, we can conclude that in the detection of complex phonetic features such as the phonological duration (long vowels and the gemination), the expert systems remain more performing. At the opposite, when a rough discrimination is solicited (macro-class discrimination), neural networks are more adapted. A compromise exists perhaps, by
injecting at the neural networks inputs, not raw data, but data including some knowledge about the forms to classify. Our SNNs are ready to use, thanks to the very versatile structure that permits to add different nature information at the input. The proposed approach, if it is refined (by including networks by kinds of speakers, a voicing network, a parallel architecture, etc.), may serve as an additional system to others complementary techniques such as the Hidden Markov Models.

References