# Classification of the Arabic Emphatic Consonants using Time Delay Neural Network

Kamel Ferrat
Ecole Nationale Polytechnique (ENP)
Algiers Algeria
CRSTDLA, University of Algiers2, Algeria

Mhania Guerti
Ecole Nationale Polytechnique (ENP)
Algiers Algeria

## ABSTRACT
This study concerns the use of Artificial Neural Networks (ANNs) in automatic classification of the emphatic consonants of the Standard Arabic Language (SAL). It reinforces the few works directed towards the speech recognition in Standard Arabic. We have applied the Time Delay Neural Network (TDNN) approach which permits a classification of the phonemes by taking into account the dynamic aspect of speech and consequently to overcome problems of coarticulation phenomenon. We have conducted a supervised training method based on Bayesian Regularization (BR) backpropagation coupled with the Levenberg-Marquardt (LM) optimization algorithm, to adjust the synaptic weights in order to minimize the error between the computed output and the desired output for all samples. Based on the results, the proposed Neural Network provides a higher percentage of recognition accuracy of the emphatic phonemes (92.25%). The choice of our study is quite important. Indeed, efficient phoneme classifiers lead to efficient word classifiers and the ability to recognize phonemes accurately provides the basis for an accurate recognition of words and continuous speech in the future.

## General Terms
Speech Processing, Neural Networks.

## Keywords
Arabic phonemes; emphatics; Speech Recognition; Neural Networks; TDNN.

## 1. INTRODUCTION
This paper deals with the recognition of the emphatic phonemes of the Standard Arabic Language (SAL) using Artificial Neural Networks (ANNs). The main idea of the ANNs is to simulate the organization of the human biological neurons and their interconnections to process the data [1,2]. The ANNs are a widely used classifier in various domains, in pattern classification [3,4]. Since the early eighties, the number of scientific papers reporting on artificial neural network (ANN) applications in speech recognition has been quickly increasing [5-7]. Because ANNs are efficient for pattern recognition, and because speech recognition is essentially a pattern recognition problem, many researchers have tried to apply neural networks for speech recognition.

For our study, we have applied the Time Delay Neural Network (TDNN), which takes into account the dynamic aspect of speech and therefore the process of coarticulation, in order to discriminate the emphatic phonemes of the SAL. By their architectures, TDNN networks provide a perfect adaptation of data presenting temporal characteristics, such as speech [8, 9]. During the training phase, we have used the Bayesian Regularization (BR) backpropagation technique, coupled with the Levenberg-Marquardt (LM) optimization algorithm. For the recognition tests, the input vector is classified with the number associated with the class that gives the minimum total distance.

In Arabic language, the automatic translation orthographic-phonetic is not systematic. A large number of phonetic and phonological rules characterize the language, and permit to replace, in the pronunciation, some consonants with other consonants, in order to avoid articulatory heaviness [10]. Those rules include essentially the sound:

-Assimilation: [man yaqūlu] (who says) is pronounced [mayyaqūlu];

-Substitution: [min bayni] (from between) is pronounced [mim bayni] ;

-Deletion: [razaqakum] (gave you) is pronounced [razaqkum];

- Gemination: [al šams] (the sun) is pronounced [aššams].

Furthermore, some consonants do not occur contiguously in the same word. Thus, the sibilant sounds [s, ṣ, š, z] do not combine each other and are never used contiguously within the same root. The same is true of the consonants (ğ/k, q/k, k/Ɛ, ħ/h, ħ/ɣ, s/ṣ, t/ṭ, d/ḍ). This shows the importance of processing in low-level (phoneme) to take account of all those rules of pronunciation. Therefore, phonemes recognition may be quite important. It will improve and perfect the recognition system and eliminate ambiguities that may arise when we automatically translate speech into written text.

## 2. EMPHASIS PROCESS IN STANDARD ARABIC LANGUAGE (SAL)
This phenomenon that presents only a simple means of expression in many languages, like French, is very pertinent in Arabic language. Indeed, the pronunciation of the word [tīn] (fig) is semantically different to the pronunciation of the word [ṭīn] (clay), by a simple emphasization of the first consonant. Likewise, the word [tāb] (he repent) is semantically different from the word [ṭāb] (it's cooked).

Sibawayh, one of the most prominent medieval Arab linguists, had given an exhaustive description of the phenomenon [itbāq] to indicate the curved form of the back of the tongue at time of pronunciation of some consonants. The movements that can describe the emphasis process on the articulatory plan are: [itbāq], the "fact of covering" and [istiɛlā?], the "fact of raising" [11]. Thus, consonants are covered "[mutbaq]" or discovered "[munfatiħ]". The mechanism of [mutbaq]

concerns the contact of the front part of the tongue and the hard palate while the back part of the tongue is raised towards the 'upper' part of the palate. For Cohen (1969), the emphasis concerns a movement of the vocal tract including a pharyngeal constriction by the projection toward the rear of the tongue root that increases the volume of the oral cavity [12]. For D.H. Obrecht (1968), the back of the tongue is raised toward the velum during the phoneme articulation [13]. So the emphasis consists in a velar constriction (velarization), added instead of the typical place of articulation of the phoneme. According to Al-Ani (1970), Arabic emphasis is synonymous with pharyngealization which is a secondary articulation of consonants by a constriction of the pharyngeal cavity [14]. Pharyngealization is considered a secondary articulation because it is an added constriction to a primary place of articulation in another location in the vocal tract.

Physiologically, the emphatic consonants are pronounced with retraction of the root of the tongue and raising of the back of the tongue towards the velum. The emphasis process concerns a pronunciation which involves a widening of the oral cavity and a constriction of the pharynx [15-16].

The emphatic consonants of the SAL are respectively, the voiced alveodental plosive [ḍ], the unvoiced dental plosive [ṭ], the voiced interdental fricative [ḏ] and the unvoiced alveolar fricative [ṣ]. So, the SAL presents four emphatic consonants: two plosives (one is voiced and the other unvoiced) and two fricatives (also the one is voiced and the other unvoiced), as shown in Table1.

**Table 1. Inventory of the emphatic consonants of SAL**

| Phoneme | Arabic Character | Place of articulation | Manner of Articulation | | |
|---|---|---|---|---|---|
| | | | voiced | plosive | fricative |
| [ ḍ] | ض | alveodental | + | + | - |
| [ṭ] | ط | dental | - | + | - |
| [ḏ] | ظ | interdental | + | - | + |
| [ṣ] | ص | alveolar | - | - | + |

Acoustic analysis shows a lowering of the $F_2$ acoustic formant due to the widening of the oral cavity and an increase of $F_1$ acoustic formant due to the constriction of the pharyngeal cavity. $F_1$ is close to $F_2$, as shown in Figures 1, 2 and 3.

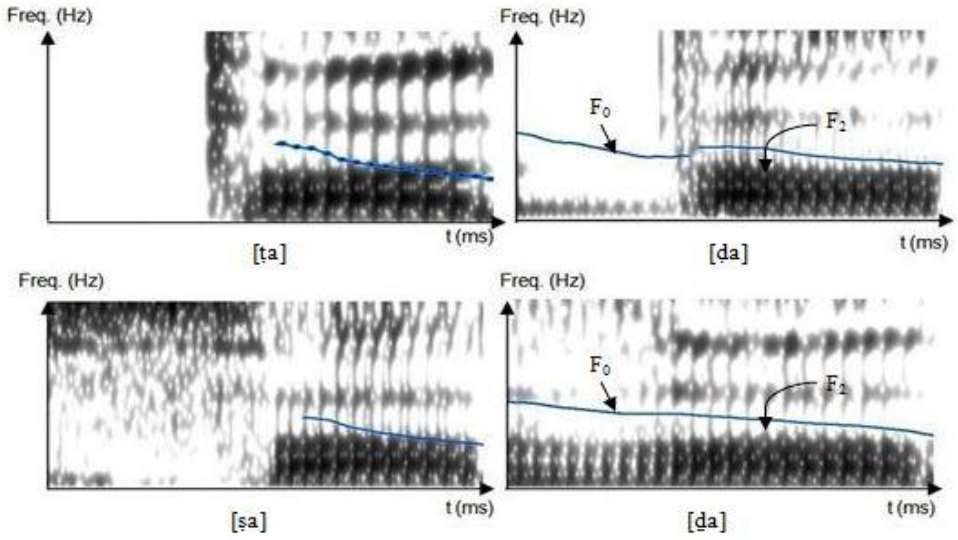


**Fig 1: Lowering of $F_2$ in adjacent vowels in [$C_e$a] context**

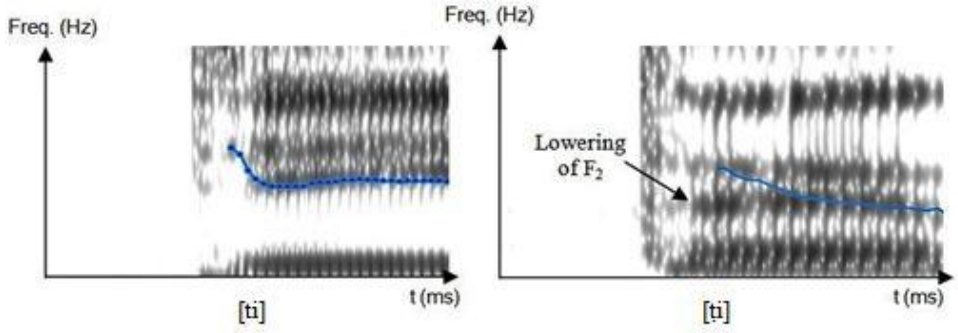


**Fig 2: Spectrogram of the emphatic plosive [ṭ] compared to its non emphatic counterpart [t], in [$C_e$i] context**

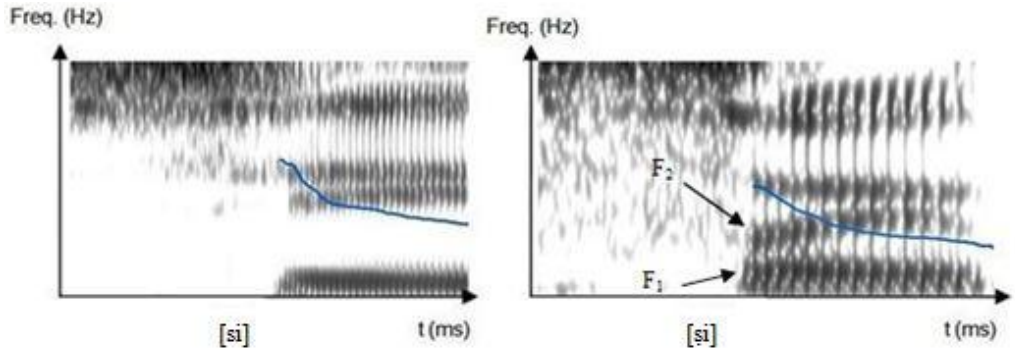


**Fig 3: Spectrogram of the emphatic fricative [ṣ] compared to its non emphatic counterpart [s], in [$C_e$i] context.**

## 3. TIME DELAY NEURAL NETWORK (TDNN)

The TDNN is a shared weights neural network which was first introduced for speech recognition by A. Waibel et al. [17]. They assume that for modeling dynamic signals such as speech, it's necessary to introduce memory in the network. The TDNN is distinguished from classical neural networks, such as MultiLayer Perceptron (MLP), by the fact that it takes into account the notion of time. Therefore, it takes into account the dynamic aspect of speech and consequently the phenomena of coarticulation. The difference between MLP & TDNN is also in the organization of the inter-layers liaisons: The MLP takes into account all the neurons of the input layer simultaneously (global view) while the TDNN only takes a window of the spectrum (local view) and then performs delayed buffers which discretely temporal shift and accumulate the input data (Figure 4). One or more layers consist of units. Each unit is connected to a unique local window in the previous layer (local information processing). All units are connected to their windows using the same set of weights (shared weights). Furthermore, the weight sharing constraint reduces the number of parameters in the system, facilitating thus the generalization process. One of the advantages of a TDNN is its capacity to use small training sets. Tebelskis quotes the findings of several papers which indicate that the TDNN, when exposed to time-shifted inputs with constraint weights, can learn and generalize well even with limited amounts of training data. This kind of network has been used for more advanced areas, such as mobile telephony, semiconductor engineering, handwriting recognition, motion recognition, control processing, forecasting of rainfall, and video quality assessment [18-23].

In Figure 4, the section "classification" acts as an MLP where each neuron of the layer is connected to all the neurons of the next layer. The first layer of the section "classification" corresponds to the last layer of the section "feature extraction"

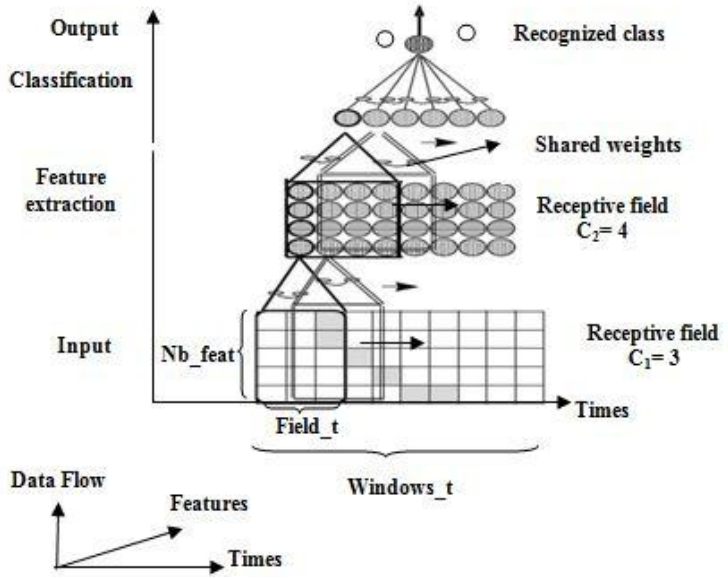


**Fig 4: Architecture of our TDNN**

## 4. EXPERIMENTAL METHOD

We have applied the TDNN network to classify the emphatic Arabic phonemes. To do this, we have followed several steps: signal preprocessing (preemphasis, endpoint detection and windowing), extraction of acoustic parameters, training, and finally automatic discrimination of the Arabic phonemes. The steps of signal preprocessing, extraction of acoustic vectors, training and automatic classification are implemented in Matlab. For the network parameters, we have used the following characteristics:

- Each layer has a characteristic direction and a temporal direction. The input layer consists of (14*18) neurons: 14 neurons as size of the observation window (window_t) and 18 neurons as size of the features (Nb_feat) ;
- two hidden layers of 12 and 8 frames respectively ;
- an observation window (window_t) corresponding to a duration of 160 ms, considered sufficient to take into account the size of the longest phoneme, so 14 frames with the frame length and the frame step size fixed to 30 and 10 ms respectively;
- a receptive field (field_t) of the input layer corresponding to $C_1$= 3 (three frames are connected to one frame in the first hidden layer) ;
- a receptive field (field_t) of the hidden layer corresponding to $C_2$ = 5 (every 5 frames are connected to one frame in the second layer) ;
- a time delay between two successive windows, Delay = 1.

A function under Matlab allows us to simulate this training. A hyperbolic tangent sigmoid function is used in each node of the hidden layers and a pure linear function in the output layer as transfer functions.

## 4.1 Speech Database

We have exploited a corpus of speech sounds extracted from the KAPD sound database, conceived at the Phonetics Laboratory of the Sciences and Technologies university of King Abdul Aziz (Saudi Arabia). KAPD contains more than 46000 files of Arabic sounds in various contexts, uttered by eight native Arabic speakers from Saudi Arabia [24]. In order to evaluate the performances of our system, we have exploited a set of 960 speech sound samples, uttered by five Algerian Arabic speakers. These recordings were achieved in the laboratory, with natural environment containing surrounding noise. We have used the *Kay Elemetrics CSL* (Computer Speech Lab) *4300B* as recording tool.

## 4.2 Pre-processing of the speech signal

Before extraction of the acoustic parameters, some processes on the speech signal are necessary, such as:

- Preemphasis by filtering with a first-order filter FIR (Finite Impulse Response) to spectrally flatten the speech signal. This type of filter boosts the magnitude of the high frequency components, leaving relatively untouched the lower ones. For that, we use one of the most widely used preemphasis filter given by:

$$F(z) = 1 - 0.97\, z^{-1} \qquad (1)$$

- Windowing with Hamming window of size *N*, expressed as:

$$h(n) = 0.54 + 0.46\cos(2\pi n/N) \qquad (2)$$

The preemphasised speech signal is blocked into frames of 30 ms. Each individual frame is windowed to minimize the signal discontinuities at borders of each frame.

- Endpoint detection of the speech signal in the environment with background noise. In other words, we must remove all frames which are not speech and delimit the useful speech signal from beginning to end, as shown in figure 5. For this task, the length of sound is divided into frames of 30ms of length, and then energy is calculated in each frame. When the energy exceeds the minimum threshold in frame, we consider that useful speech starts from this frame onward. All the other preceding frames are removed. The same procedure is applied at the end of speech sound. A Matlab function is used to do this procedure. This function uses a minimal threshold of average energy which we have calculated on the basis of recordings the ambient noise.

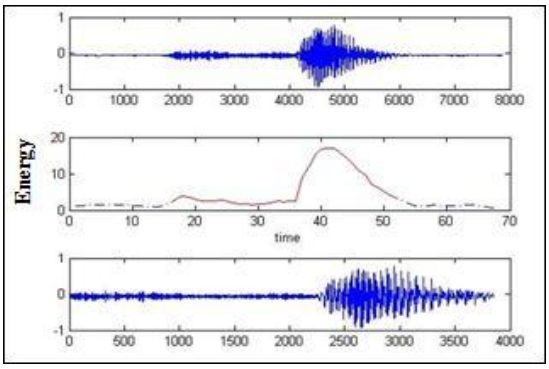


**Figure 5. Endpoint detection of the sound [kataba]**

## 4.3 Feature extraction

The feature extraction is crucial to prepare data for the classification process. For that, we search to obtain the most representative signal form in order to minimize the recognition error, so we use MFCCs (Mel Frequency Cepstral Coefficients) parameters which allow a modelling of the speech signal by filters corresponding to human hearing system. We note that MFCCs are commonly used as features in speech and speaker recognition systems [25-26]. We complete those MFCCs coefficients by the temporal derivatives, first ΔMFCCs and second ΔΔMFCCs, which permit to take into account the temporal variability of speech, and therefore its dynamic aspect. We use 13 MFCCs coefficients with the 1st coefficient corresponding to energy, and their derivates, so a total of 39 coefficients per frame, considered as a compact representation of the speech signal.

## 4.4 Training step

The central problem in neural networks is the development of training and recognition algorithms, which can perform the desired interaction with a changing or unknown and fuzzy environment such as speech processing and pattern recognition. To do this, we have exploited the Bayesian Regularization (BR) backpropagation coupled with the Levenberg-Marquardt (LM) optimization algorithm, to adjust the synaptic weights in order to minimize the error between the computed output and the desired output for all samples [27-29]. The BR permits to overcome the problems of under and over training of the system. The LM algorithm gives best training performances with a small number of iterations compared to other backpropagation algorithms [30]. For the network training, we iteratively adjust all the weights to minimize the Mean Squared Error (MSE) which represents the average squared difference between desired outputs and real obtained outputs:

$$MSE = \frac{1}{2}\sum_i^n (d_i - y_i)^2 \qquad (3)$$

With $d_i$ : desired output for neuron *i* ;

$y_i$ : obtained real output.

After random initialization of the connection weights, we compute the output of the network according to the input (computing of the potential) by the following equation:

$$y_i = f(p_i) \qquad (4)$$

$$p_i = \sum_{i=1}^{N}\sum_d^D W_{ij}^D * e_i\,(t-d) \qquad (5)$$

With $P_j$ : Potential of neuron j ;

$W_{ij}^D$ : Weight of the connection from the i-th neuron of the lower layer and the j-th neuron of the upper layer ;

$\sum_{i=1}^{N}$ : Sum of input neurons of neuron i;

$\sum_{d}^{D}$ : Sum of delays (receptive field);

$e_i$ : state of the neuron i ;

$t$ : the present instant ;

$d$ : delay.

Input features are rescaled to the interval [-1, 1]. These rescaled values are near transition regions of sigmoid function, which allows a faster training of the system. In addition, inputs of the neural net are often of different types and different scales, so it's necessary to normalize the data by centering and reducing the variables, in order to have the same impact on the model [29, 31]. For this, we have transposed the original variables to new centered and reduced variables:

$$\overline{x_i} = \frac{1}{N}\sum_{n=1}^{N} x_i \quad (6)$$

$$\sigma_i^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(x_i - \overline{x_i}\right)^2 \quad (7)$$

$$x_i = \frac{x_i - \overline{x_i}}{\sigma_i} \quad (8)$$

With $N$ number of samples, and $\sigma_i$ variance.

## 4.5 Recognition test step

For the recognition test of the output vector, the rule is to compute the minimal distance corresponding to target vector of reference matrix which is close the test vector. The input vector is, therefore, classified with the number associated with class that gives minimum total distance. Also, it's necessary to note that for the coding of target vectors (desired outputs), we have chosen the classic method which consists in assigning "1" for a single element of vector and "0" for all the others, so that all the vectors possess a single activation in "1"and all the other activations in "0". We introduce the selected acoustic parameters at the input of the system, taking into account the weights obtained from the training phase [32]. A simulation was then performed with a set of tests, in order to evaluate the performance when unknown samples are presented. The classifier behavior is assessed in terms of the percentage of correct classification in the test set. The method to calculate the Global Recognition Accuracy (GRA) is given by:

$$RA\ (\%) = (Correct\ cases\ /Total) \times 100 \quad (9)$$

$$GRA\ (\%) = (\Sigma\ RA)/M \quad (10)$$

*where M is the number of emphatic phonemes*

Overall, 92.25 % of the emphatic consonants in the database test samples were located and correctly classified. The table 2 shows the confusion matrix and results of recognition tests of all the emphatic consonants.

**Table 2. Confusion matrix, Recognition Accuracy (RA) and Global Recognition Accuracy (GRA) for the emphatic consonants**

| **Confusion (%)** | [ṭ] | [ṣ] | [ḏ] | [ḍ] | RA (%) |
|---|---|---|---|---|---|
| [ṭ] | **98.00** | 01.03 | 00.24 | 00.73 | 98.00 |
| [ṣ] | 00.12 | **99.60** | 00.08 | 00.20 | 99.60 |
| [ḏ] | 03.24 | 00.36 | **76.38** | 20.02 | 76.38 |
| [ḍ] | 00.50 | 00.00 | 04.50 | **95.00** | 95.00 |
| | | | | **GRA** | **92.25** |

# 5. RESULTS AND DISCUSSION

From the obtained results shown in Table 2, we note that:

- The emphatic phonemes [ṭ] and [ṣ] have a highly RA (≥98%). The other emphatic phonemes are recognized with proportions of 76.38% and 95%;
- Confusion has been observed between the phonemes [ḍ] and [ḏ]. That is probably caused by the confusion in pronunciation of these two phonemes in the Maghreb countries, contrary to the Middle East. We notice that we have used for the training phase, a set of samples pronounced by Saudi Arabian speakers and for the testing phase, we have used a set of samples pronounced by Algerian speakers. This also shows that the training step has been well done by the system.
- Nevertheless, our system permits to have an appreciable GRA (92.25%) of the emphatic phonemes.

# 6. CONCLUSION

In this study, we give a general description on main features of the eight emphatic phonemes of Standard Arabic language. After that, we show the contribution of the main method of artificial neural networks ANNs for the automatic recognition of these emphatic phonemes. To achieve this, we apply a Time Delay Neural Network TDNN. This experience shows us that the contribution of Neural Network methods for the automatic recognition of the emphatic phonemes of Arabic Language is very interesting. This method permits us to have appreciable recognition accuracy of the emphatic consonants, with notably a recognition accuracy (≥98%) of the consonants [ṣ] and [ṭ]. Some confusions persist between the consonants [ḍ] and [ḏ] which presents many similarities in pronunciation in the Maghreb Countries

# 7. ACKNOWLEDGMENTS

The authors would like to thank the Computer and Electronic Research Institute King Abdulaziz City for Science and Technology (Saudi Arabia), particularly Mr. Mansour AlGhamdi, for providing the database Kacst Arabic Phonetic Database (KAPD) of the recorded Arabic sounds.

# 8. REFERENCES

[1] Kremer, S. 2001. Spatiotemporal connectionist models: A taxonomy and review. Neural Comput. 13, 249-306.

[2] Dreyfus, G. 2004. Réseaux de neurones- Méthodologie et Application. Ed. Eyrolles, France.

[3] Duda, R.O., Hart, P.E., and Stork, D.G. 2001. Pattern Classification", John Wiley and sons, second edition.

[4] Bishop, C.M. 1995. Neural Networks for Pattern Recognition. Oxford University Press.

[5] Tebelskis, J. 1995. Speech Recognition Using Neural Networks. PhD. Dissertation, School Of Computer Science, Carnegie Mellon University.

[6] Ahad, A., Fayyaz, A., and Mehmood, T. 2002. Speech recognition using multilayer perceptron. Proc. of the IEEE Conference ISCON'02, 1 , 103-109.

[7] Dede G., and Sazlı M.H. 2010. Speech recognition with artificial neural networks. Digit. Signal Process. 3(20), 763-768.

[8] Gatt, E., Micallef, J., Micsllef, P., and E. Chilton. 2001. Phoneme Classification in Hardware Implemented Neural Networks. Proceedings of the 8th IEEE International Conference on Electronics, Circuits and Systems, Malta.

[9] Hou, J., Rabiner, L., and Dusan, S. 2008. Parallel and Hierarchical Speech Feature classification using frame and segment-based methods. Interspeech2008, Brisban, Australia.

[10] Alfozan, A.I. 1989. Assimilation in Classical Arabic A phonological study. Thesis Doctorat of Philosophy, Faculty of Arts of the University of Glasgow.

[11] Roman, A. 1983. Etude de la phonologie et de la morphologie de la koiné arabe. Tome I, Université d'Aix-En-Provence, France.

[12] Cohen, D. 1969. Statut phonologique de l'emphase en arabe. Word 25, 59-69.

[13] Obrecht, D. 1968. Effects of the second formant on the perception of velarization consonants in the Lebanese Arabic. Ed. Mouton, The Hague.

[14] Al-Ani, S.H. 1970. Arabic Phonology. An Acoustical and Physiological Investigation. Ed. Mouton, The Hague.

[15] Al-Tamimi, F., Alzoubi, F., and Tarawnah, R. 2009. A Videofluoroscopic Study of the Emphatic Consonants in Jordanian Arabic. Folia Phoniatr Logop. 61, 247-253.

[16] Ferrat, K. 2005. Acoustical study of the Tachdid and the Idgham in Standard Arabic- Application for speech synthesis. Int. Conf. Sci. of Electronic, Technologies of Information and Telecommunication, SETIT2005, Susa, (Tunisia), IEEE France.

[17] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. 1989. Phoneme recognition using time-delay networks. IEEE Trans. Acoustics, Speech and Signal Process. 37(3), 328-339.

[18] Lin, C.T., Nein, H., Lin, W.C. 1999. A space-time delay neural network for motion recognition and its application to lipreading. Int. J. Neural Syst. 9 (4), 311-334.

[19] Stokes, D., and May, G.S. 2000. Real-time control of reactive ion etching using neural networks. IEEE Trans. Semicond. Manuf. 13(4), 469-480.

[20] Pfister, M., Behnke, S., Rojas, R. 2000. Recognition of Handwritten ZIP Codes in a Real-World Non-Standard-Letter Sorting System. Appl. Intell. 12(1-2), 95-115**.**

[21] Le Callet, P., Viard-Gaudin, C., Barba, D. 2006. A convolutional neural network approach for video quality assessment. IEEE Trans. Neural Netw. 17(5), 1316-1327.

[22] Stegmayer, G., Chiotti, O. 2006. Neural networks applied to wireless communications. Artificial Intelligence in Theory and Practice. IFIP 19th World Computer Congress, Santiago, Chile.

[23] Htike, K., Khalifa, O. 2010. Rainfall Forecasting Models Using Focused Time-Delay Neural Networks. Int. Conf. Computer Communication Engineering (ICCCE 2010), Kuala Lumpur, Malaysia.

[24] Alghamdi, M. 2003. KACST Arabic Phonetic Database. The Fifteenth International Congress of Phonetics Science, Barcelona.

[25] Chia Ai, O., Hariharan, M., Yaacob, S., and Sin Chee, L. 2012. Classification of speech dysfluencies with MFCC and LPCC features. Expert Syst. 39(2), 2157-2165.

[26] Sahidullah, M.D., and Saha, G. 2012. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Commun. 54(4), 543-565.

[27] Marquardt, D. 1963. An Algorithm for Least-Squares Estimation of Nonlinear parameters. SIAM J. Appl. Math. 11, 431-441.

[28] Fun, M., and Hagan, M. 1996. Levenberg-Marquardt Training for Modular Networks. International Conference on Neural Networks, Washington, USA.

[29] Foresee, F.D., and Hagan, M.T. 1997. Gauss-Newton Approximation to Bayesian Regularization. International Joint Conference on Neural Networks, Houston, USA.

[30] Dhar, V.K., Tickoo, A.K., Koul R., and Dubey, B.P. 2010. Comparative performance of some popular artificial neural network algorithms on benchmark and function approximation problems. Pramana-J. Phys. 74(2), 307-324.

[31] Demuth, H., and Beale, M. 2000. Neural network toolbox. User's Guide.from. http://www.mathworks.com/access/helpdesk/help/pdf_doc/nnet/nnet.pdf.

[32] Ferrat, K., Guerti, M. 2011. Apprentissage et Reconnaissance Automatique de la Parole par Réseaux de Neurones Artificiels. Rev. Sciences de l'homme. 4, 57-71.