Asian Journal of Information Technology 5 (1) : 89-95, 2006

Recognition of Arabic Pronunciation of Numbers Using Neural Networks

[1]N. Boukezzoula, [1]F. Djahli and [2]Y. Bouterfa
[1]Neural Networks laboratory, Electronic Department,
Engineer's Sciences Faculty, Setif University, Algeria
[2]Electronic Department, Louvain la neuve University, Belgium

Abstract: In this study we present an extraction method of some parameters contained in the Arabic pronunciation of numbers from 0 to 9 (sifr to tessea), which are then used as an input vector of the proposed neural networks.Globally, the methodology is concentrated about three important steps. The first step is the extraction of characteristics as the formants and their respective bandwidths of some interesting particular woofs. A learning step using two cycles for adjusting, computing and saving the weights of the neural networks interconnections. Finally a recognition step or test verifying the credibility of the system.

Key words: Formant extraction , learning phase, recognition phase, back-propagation, vocal number

INTRODUCTION

Speech is one of the most used ways of communication. Since many years, researchers attempt to conceive devices allowing vocal man-machine dialogue. Their hope is to control many electronics systems by speech, or to control a robot, or to ask a database using the phone.

The computer developement has allowed a direct communication man-machine and many great project of speech recognition have been developed in the last years. Today, the speech recognition is in full rapid expansion and we observe the multiplication of the applications domain (robot control, security systems, rolling and moving seats control for handicapped motor bodies, vocal phone number, etc.).

Using the speech recognition device, we can do many vocal control operations[1-3].

We know that the speech recognition is treated by many methods as the statistics methods (the Hidden Markov Model, the hybrid Model, etc.)[5,6] .

This communication mode facilitates the task to the users and particularly to the handicapped bodies, excepted the deaf dumb bodies.

The objective of this study was to develop a device allowing the handicapped bodies to use the phone, only by pronouncing numbers to obtain the correspondent.

The developed system is a recognition device of isolated words. The applied strategy is the back-propagation algorithm[7-12].

Our work concerns the Arabic pronunciation numbers recognition using the back-propagation method seeing which results can we obtain and if possible making a comparison with other methods (the HMM, etc.)[5,6].

The isolated words pronounced are: sifr, wahed, ithnani, thalatha, arbaa, khamsa, setta, sabea, thamania, tessea.

The corpus is constituted of ten words(numbers), 20 speakers pronounce each word in Arabic language where 50% of the corpus is used for learning and 25% for tests.

The corpus is recorded with a sampling frequency of 16 kHz in a size of 16 bits.

The objective is to reduce the redundancy of the speech signal and preserving from the available data, only a cluster of pertinent parameters reducing the calculation time and the storage time at the learning and recognition treatment.

The realized system is constituted of two different and independent programs.

The first program concerns the speaker signal analysis from which we extract the parameters characterizing the signals of the Arabic numbers pronunciation from 0 (sifr) to 9 (tissaa).

The second program is consisting of two phases, the learning phase and the recognition phase of a finite number of speakers.

We also present the developed system, the used characteristics signal and the learning and recognition phases.

Corresponding Author: N. Boukezzoula, 14, rue Dehbi Mohamed cité Tlidjene, 1900, Setif (Algeria)

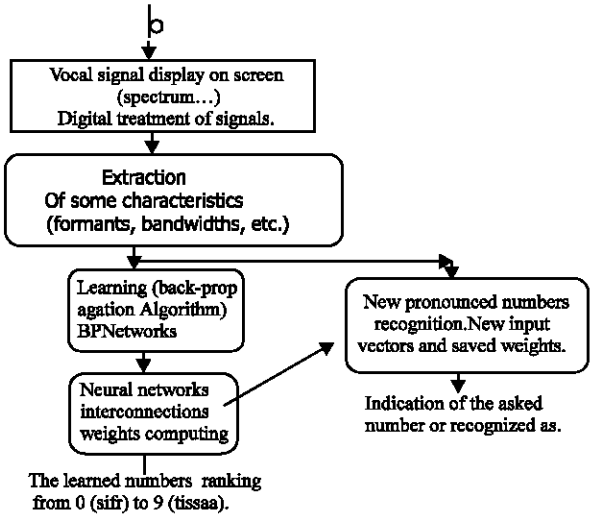

Fig. 1: Recognition system organization

GENERAL DESCRIPTION OF THE DEVELOPED SYSTEM

Before detailing the different parts of the system, wepropose a brief description of its general organization. Thedeveloped device is composed of many distinct phases as shown in the algorithm of (Fig. 1).

In a first time, we carry out a digital treatment of the recorded signals, then we make an analysis which allows the computing of some interesting characteristics as the formants, the bandwidths, etc.

In a second time, each Arabic pronounced number undergoes a learning operation by neural networks. The neural networks weights interconnections are calculated for the pronounced numbers cluster, which are exploited in the recognition phase.

EXTRACTION OF THE CHARACTERISTICS

Analysis with formants: The vocal pipe presents some obvious numbers of natural pulsations for a voiced sound.

The natural frequencies are noticed F_k (k=1, 2, …, 5) and the amortizations are defined by the relative bandwidths at 3 dB noticed B_k. These natural frequencies constitute the formant of vocal pipe.

The principal objective of the analysis with formants is the extraction of the parameters representing the important signal characteristics, which are the formants. A formant is generally defined as a sinusoidal component of the vocal pipe response to an acoustic pulse. In the classical speech model, a formant can be also defined as

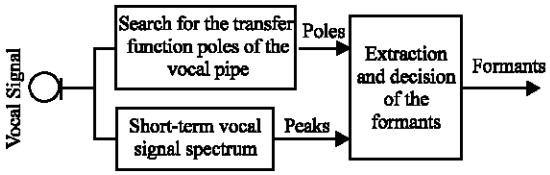

Fig. 2: Analysis with formants

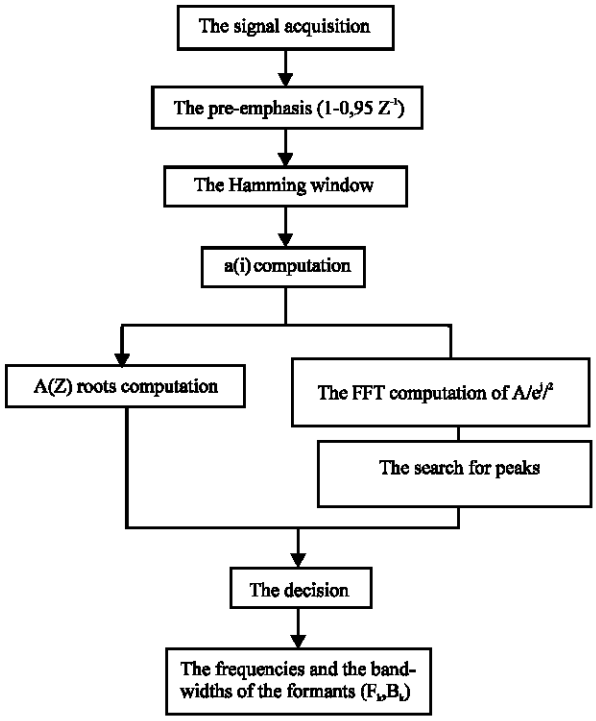

Fig. 3: The formants trajectory valuation

a pair of complex poles of the vocal pipe transfer function [13].

The analysis with formants (Fig. 2) consists to extract from the vocal signal, the formants frequencies, their (Fig. 3). amplitudes and their bandwidths[14].

We commonly use the LPC analysis[13,14]. The schematic organization chart of the linear prediction method LPC is represented in (Fig. 3)

DESCRIPTION OF THE DIFFERENT STEPS OF THE ANALYSIS OPERATION

Notions of the used steps in analysis: The Hamming window is the commonly used. It is defined by[14]

$$\text{Ham}_N(k) = 0,5\ 1+\cos\ \frac{2\pi k}{N} \quad \text{for } |k| \le n/2 \qquad (1)$$

$$\text{Ham}_N(k) = 0 \quad \text{otherwise}$$

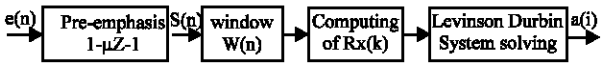

Fig. 4: The organization chart of the coefficients a(i) calculation

The window duration N (samples number) depends directly on the sampling frequency F_e, the better choice is given by[14] :

$$N=F_e S \quad (2)$$

Where S varies between 15 and 20 ms and F_e in Khz The used method for the coefficients a(i) valuation is the self-correlation method shown in (Fig. 4)

The good vocal pipe modification for the voiced sound consists to take the prediction order m equal to[14]:

$$m = F_e \, 2\frac{L}{C} + D \quad (3)$$

where

F_e: is the sampling frequency (KHz)
L: is the vocal pipe length, generally taken equal to 17 cm
C: is the sound speed C=34 cm/ms
D: takes the value 4 or 5

The model order used is at least equal to 16. After the coefficients a(i) calculation, we use the Fast Fourier Transform to compute the spectrum model.

The formants extraction: After the pre-emphasis, the windowing, the coefficients a(i) calculation and the obtaining of the signal spectrum, the frequencies F_i and eventually the bandwidth B_i estimation can be done by searching the spectrum model maximums. Many methods allow this calculation; among these methods, we notice the factorization method where the spectrum model maximums can be determined by the polynomial A(z) factorization. For computing the formants, we only resolve the following equation:

$$1+\sum_{k=1}^{p} a_k Z^{-k} = 0 \quad (4)$$

This method guarantees that all the formants and the possible amortization will be extracted.

If $Z = R_e(Z)+j\, I_m(Z)$ is a root of the previous equation, the associated formant frequencies and their corresponding bandwidths are given by[14,15]:

$$F_i = \frac{F_e}{2\pi} \operatorname{arctg} Z_i = \frac{F_e}{2\pi} \operatorname{arctg} \frac{I_m(Z_i)}{R_e(Z_i)} \quad (5)$$

And

$$B_i = -\frac{F_e}{\pi} \ln|Z_i| \quad (6)$$

Where

F_e: is the sampling frequency.
Ln: is the neperian logarithm.
I_m (Z): is the imaginary part of the root.
R_e (Z): is the real part of the root.

Many algorithms have been developed to determine the roots of the polynomials of the different degrees of the polynomials for factorization; the used algorithm is from Bairstow[14]

The choice of the neural networks input vectors: We propose the Arabic pronunciation of zero (0). We show (Fig. 5) the signal time representation of zero (sifr), a weft formant values table and below, the FFT corresponding spectrum.

We noticed that in the given example, the spectrum is obtained for the second weft, which is between 256 and 512 samplers.

We present in (Table 1) the formants values of the sifr (0), khamsa (5) and tessea (9) numbers.

Table 1: The Formant frequencies of three numbers

The number of the formant	The format frequencies [Hz] for sifr (0)	The formant frequencies [Hz] for khamsa (5)	The formant frequencies Hz] for tissaa (9)
1	947.830	682.320	1126.327
2	1276.605	622.323	1173.619
3	2298.776	1299.853	1977.839
4	2774.288	1696.536	1954.316
5	5000.000	2079.361	2840.409
6	2298.776	2226.299	5000.000
7		2345.671	2840.409
8		5000.000	
9		5000.000	
10		2346.671	

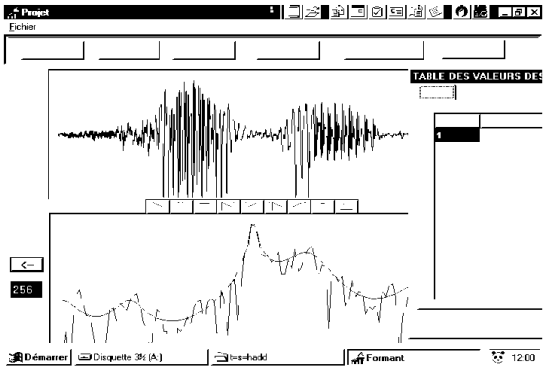

Fig. 5: Time and spectral representation of the Arabic zero (sifr) pronunciation

With the all registered formants of many wefts, we have made an empirical choice to select the components for the proposed neural networks input vectors.

THE NEURAL NETWORKS ARCHITECTURE

Learning and recognition: Using the neural approach as recognition system requires:

- Inputs (parameters extracted from the analysis, which constitutes the different wefts of a sampled and digitized vocal signal.
- One or many hidden layers.
- Outputs (ten relative numbers outputs) from 0 (sifr) to 9 (tesseaa) pronounced in Arabic language.

As for any recognition system, there is two distinct phases, the learning and recognition phases.

Globally, we adjust the weights w_{ij} for a fixed minimal error. The role of the learning is to adjust the interconnection weights between nodes of the different layers of the networks. This phase is applied on the cluster of the used phone numbers from 0 (sifr) to 9 (tessea), during five epochs.

We noticed that the weights initialization was delicate at the beginning of learning.

After making the learning on each number, we determine in the recognition phase, the maximal output corresponding to each number.

The outputs are numbered from 1 to 10 and correspond respectively to the number 0 (sifr) to 9 (tissaa) as in the registration used in the learning phase.

The input vectors have been chosen empirically. We also have arbitrary chosen, at the beginning of learning, the values of the learning rate, the momentum, the activation function slope (sigmoid or hyperbolic tangent) and the nodes number of the hidden layers. After many tests the values have been improved experimentally, to obtain the optimum results, of course with respecting an optimal convergence time.

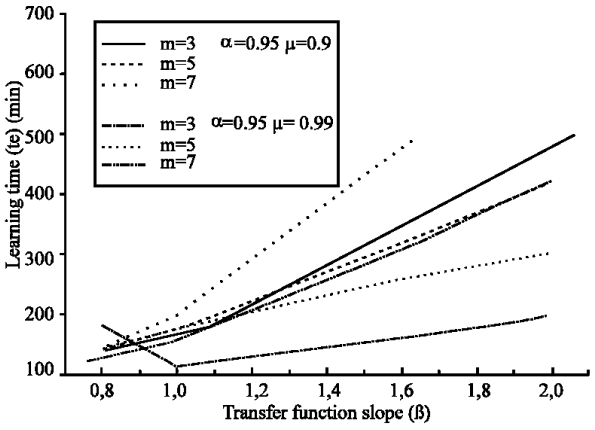

Fig. 6: The convergence duration $t_e = f(\beta)$ in the one hidden layer networks

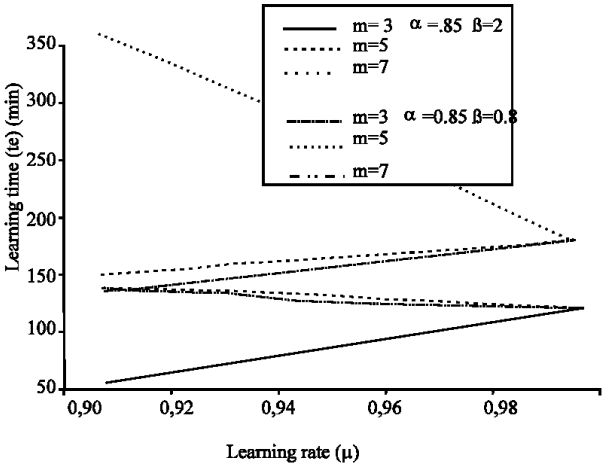

Fig. 7: The convergence duration $t_e = f(\mu)$ in the one hidden layer network.

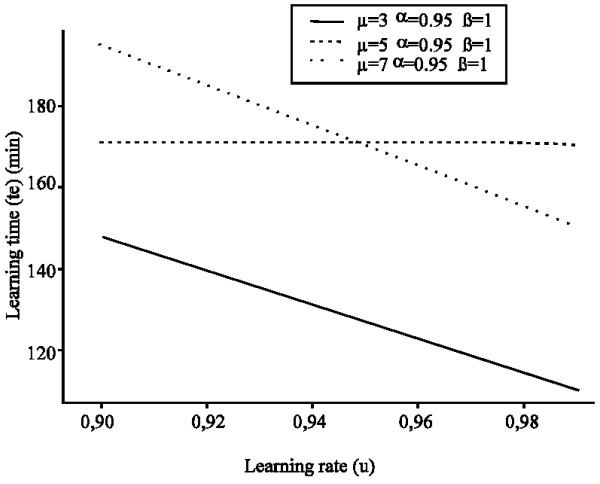

Fig. 8: The convergence duration $t_e = f(\mu)$ in the one hidden layer networks

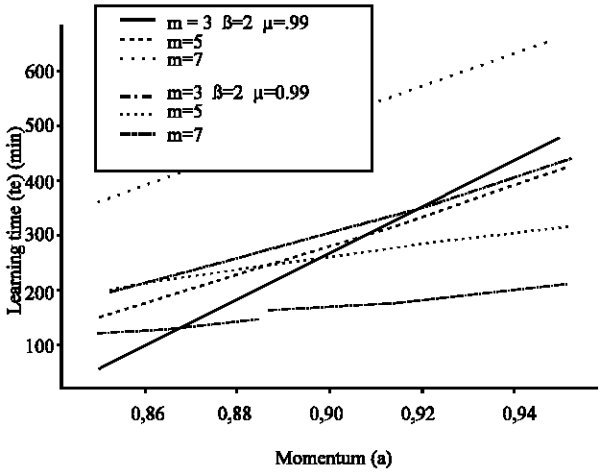

Fig. 9: The convergence duration t_e= f(α) in the one hidden layer networks

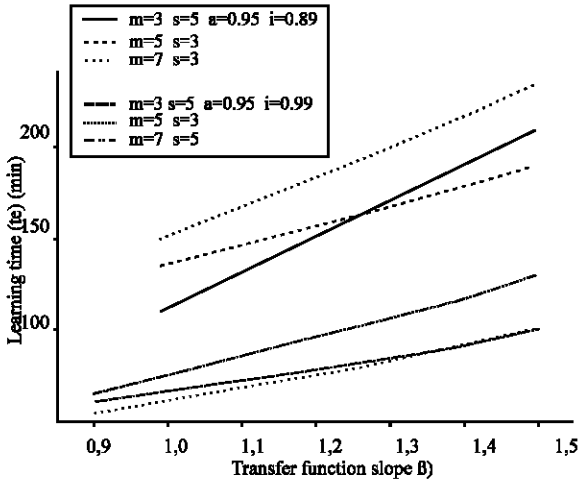

Fig. 10: The convergence duration t_e= f(β) in the two hidden layer networks.

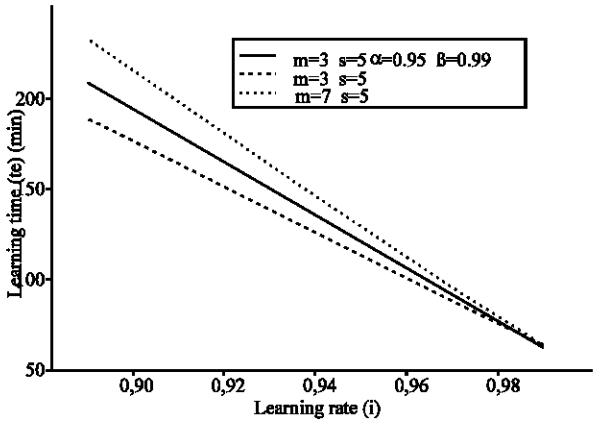

Fig. 11: The convergence duration t_e= f(μ) in the two hidden layer networks.

To facilitate the choice of the parameters, we have represented different curves for two types of networks; we show the influence of μ (learning rate), α (momentum) and β (transfer function slope) on the convergence time.

The one hidden layer networks (1hl): We have proposed three networks with one hidden layer, which have respectively 3, 5 and 7 neurons. We present the learning time t_e (convergence time), versus β for some values of α and μ (Fig. 6). Figures 7 and 8 illustrate the learning time t_e versus μ for some values of α and β. We noticed that the slope is always negative for the seven hidden neurons networks.

We also show in (Fig. 9) the variation of t_e versus α for some values of μ and β. Here the slopes have the same positive variation.

We noticed that t_e doesn't necessary change with the neurons number, but as shown in the presented curves, each case gives a particular result. Sometimes, we have the same duration for two different networks and sometimes, we obtain the decreasing of t_e when we expect its increasing. This is a normal phenomenon because the used parameters play the role of accelerator and for some precise values the inverse role.

We also obtain the same time t_e for all networks for some particular values of μ and β. Looking the obtained results, we can conclude that the choice of the parameters is purely experimental and differ from one case to the other.

The two hidden layers networks (2hl): The also proposed networks are with two hidden layers with respectively m and s as neurons numbers in the first and the second hidden layers. The used couples (m, s) are (3, 5), (5, 3) and (7, 5). Figures 10 and 11 represent the learning time t_e respectively versus β for some values of α and μ and versus μ for some values of α and β.

Practically, we can conclude that the choice of the parameters is purely experimental. So, there are no absolute rules for the different applications that we could imagine.

Outputs and errors networks calculation: In (Table 2) we show the ten network outputs values, the corresponding error and the total error after the formula of [8-11] for the learning of one pronounced Arabic number 0 (sifr).

The neural networks results interpretation: We have met the local minimum problem for the one and two hidden

Table 2: The convergence of the number 0 (sifr)

The outputs number	The desired outputs	The networks outputs	The corresponding errors
1	1	9.98000039766339e-0001	1.99996023366111e-0003
2	0	1.99231315324511e-0003	1.99231315324511e-0003
3	0	1.99668224578176e-0003	1.99668224578176e-0003
4	0	1.99538507840202e-0003	1.99538507840202e-0003
5	0	1.99312214493830e-0003	1.99312214493830e-0003
6	0	1.99524399790008e-0003	1.99524399790008e-0003
7	0	1.99319549692234e-0003	1.99319549692234e-0003
8	0	1.99567137412870e-0003	1.99567137412870e-0003
9	0	1.99355941033019e-0003	1.99355941033019e-0003
10	0	1.99305115049953e-0003	1.99305115049953e-0003
		The total error Et=1.98965266339202e-05	

layers networks that we have overcome by doing many tests. We have noticed that the iterations number in the networks convergence is lower in the sigmoid activation function compared to the hyperbolic tangent activationfunction. We explain this by the greater slope value and the lower and higher boundaries of the hyperbolic tangent activation function.

For example, we have 16059 iterations in a time of 1h 35 min for the learning of the number 7 (sebaa) and 25340 iterations and a time of 2 h 40 min for the learning of the number 9 (tissaa).

We have also noticed that for a relative great momentum α value, we obtain the acceleration of the convergence, but without exceeding some limits, otherwise the momentum α becomes the more important value, which slows down the convergence of the networks. The momentum α has an influence on the networks convergence rapidity.

The learning rateμ accelerate the convergence, when it has a relative great value, but in certain limit, otherwise we have the slowing down of the convergence. Theμ factor has also an influence on the networks convergence rapidity.

After many tests, we noticed that the parameters, which have given the convergence, are in the following limits:

- For the 1 hl networks: $0.8 \leq \alpha \leq 0.85$; $0.8 \leq \beta \leq$; $0. \leq \mu \leq 0.99$
- For the 2 hl networks: $0.8 \leq \alpha \leq 0.85$; $0.99 \leq \beta \leq 1.5$; $0.89 \leq \mu \leq 0.99$

We can conclude that the α, β andμ values and the hidden neurons numbers are purely experimental and differ from one case to the other.

We can say that we have not obtained confusion in our application. This method is efficient as probabilistic methods like the Hidden Markov Model and the Hybrid model[1-5].

CONCLUSION

In this study, we have modeled and determined the essential speech characteristics (which are the Arabic numbers in our case) to use them in an eventual phone number application.

Our system is valuable for many speakers. We have used twenty different speakers. All the speakers are men aged between 22 and 27 years.

We have used the LPC method, which has allowed to extract the different types of parameters characterizing the numbers like the a(i) prediction coefficients used in the extraction of the formants values using the factorization method. The formants are specific parameters for each number.

After the obtained results, we can conclude the efficiency of the neural networks in the Arabic pronunciation recognition of the number 0 (sifr) to 9 (tessea). We precise that we have respectively used the pronunciation of the two Arabic numbers 2 and 8 as ithnani and thamania but not as ithnan and thamaniya.

After many tests and to generate outputs values approximating the desired outputs values with a very small error, the choice of the α, β andμ values are very important to obtain a minimal error and an optimal learning time.

The obtained results allow confirming the efficiency of the back propagation algorithm in the Arabic number pronunciation recognition for an eventual vocal phone number.

The networks with one and two hidden layers illustrate the validity of the results using the BPN (back propagation networks).

We have also made many experiences with the variation of α, β,μ values and using different clusters of initial weights.

For the phone number, we can use a circuit, which record the number pronounced and recognized in the order of the phone numbering. This circuit can activate a

pulse generator which provides pulses for each number with a delay between the numbers which constitute the phone number (the number dial technique), or to active two particular sinusoidal signal generators corresponding to each number pronounced and recognized (the number Keyboard technique).

This work has allowed having a concrete vision on the use of the pronunciation of the number for the vocal phone number, which will be very beneficial and useful for some handicapped bodies and particularly for the blind bodies.

The results give that all the two methods (the HMM and the Neural Networks methods) possesses some limits in the real applications. So, we must have a good quality of the data acquisition: the appropriate frequency, a good recording conditions and good noise immunity.

REFERENCES

1. Ben Sassi, S., R. Braham and A. Belghi,t 2001. Neural speech synthesis system for arabic language using CELP algorithm. AICCSA., Beirut, Lebanon, pp: 119- 121.
2. Tsubata, Y., T. Kawahara and M. Dantsuji, 2002. Recognition and verification of English by japanese students for computer- assisted Language learning. School of Informatics Center of Information and Multimedia Studies, Kyoto University, pp: 301-307.
3. Dibazar, A.A., S. Naryanan and T.W. Berger, 2002. Feature analysis for automatic detection of pathological speech. Biomedical Eng. Dept. Elect. Eng. Dept, University of South California, pp: 405-411.
4. Istrate, D., 2003. Detection et reconnaissance des sons pour la surveillance médicale, CLIPS-IMAGdicale, These de Doctorat, CLIPS-IMAG, France, pp: 595-598
5. Bahi, H. and M. Sellami, 2001. Combination of Vector Quantization and Hidden Markov Models for Arabic Speech Recognition. ACS/IEEE International Conference on Computer system and Applications (AICCSA'01), Beirut, Lebanon, pp: 96-101.
6. Khabet, S., M.T. Laskri and Z. Zemirli, 2002. Reconnaissance de la parole à l'aide de Modèles de Markov Cachés : Application à la reconnaissance de chiffres Arabes isolés et d'un Système d'Appel Téléphonique Vocal, MCEAI'02, Annaba, pp: 89-94.
7. Pao, Y.H., 1989. Adaptive pattern recognition and neural networks. USA Addison- Wesley publishing company Inc.
8. Widrow, B., M.A. Lehr, 1990. 30 Years of Adaptive Neural Networks: Perceptron, Madaline and Back-propagation. New York. IEEE Vol.78, pp: 1415-1438
9. Simpson, PK., 1990. Artificial neural systems. Foundations, Paradigms, Applications and Implementations, New York: Pergamon Press, Inc, Maxwell House.
10. Caudill, M. and C. Buttler, 1992. Basic Networks and Advanced Networks. Understanding neural networks, Massachusetts, A Bradford Book. MIT press. 1: 169-196.
11. Vitela, J. E. and J. Reifman, 1997. Premature Saturation in Back propagation Networks: Mechanism and Necessary Conditions, Neural networks, Elsevier Science Ltd, Pergamon Express. 10, pp: 221-227.
12. Boukezzoula, N., 2004. Speaker Recognition using Neural Networks, CIGE'04 Congrés international sur le génie électrique, Setif (Algeria), pp: 109-112.
13. Boite, R., 2000. Traitement de la parole, Presses polytechniques et Universitaire
14. Markel, J.D. and A.H. Gray, Jr 1987. Linear prediction of speech. New York, Springer-verlag., Berlin Heidelberg, New york.
15. Kunt, M., 1981. Traitement Numérique Des Signaux, Edition Dunod, 3rd Edition.