computing@tanet.edu.te.ua
www.tanet.edu.te.ua/computing

ISSN 1727-6209
International Scientific
Journal of Computing

A Robust Speech Recognition System Using a General Regression Neural Network

Abderrahmane Amrouche [1], Abdelmalik Taleb-Ahmed [2], Jean Michel Rouvaen [3], Mustapha C. E. Yagoub [4]

[1] Faculty of Electronics and Computer Science, USTHB, P.O. Box 32, Bab Ezzouar, Algiers, Algeria.
namrouche@ usthb.dz.
[2] LAMIH (UMR CNRS 8530)
taleb@univ-valenciennes.fr
[3] OAE-IEMN (UMR CNRS 8520),
rouvaen@univ-valenciennes.fr
[2,3] Valenciennes University, P.O. Box 304, Le Mont Houy, 59 313, France.
[4] SITE, University of Ottawa, 800 King Edward, Ottawa, ON, K1N 6N5, Canada.
myagoub@site.uottawa.ca

Abstract: *General Regression Neural Networks (GRNN) have been applied to phoneme identification and isolated word recognition in clean speech. In this paper, the authors extended this approach to Arabic spoken word recognition in adverse conditions. In fact, noise robustness is one of the most challenging problems in Automatic Speech Recognition (ASR) and most of the existing recognition methods, which have shown to be highly efficient under noise free conditions, drastically fail drastically in noisy environments. The proposed system has been tested for Arabic digit recognition at different Signal-to-Noise Ratio (SNR) levels and under four noisy conditions: multitalker babble background, car production hall (factory), military vehicle (leopard tank) and fighter jet cockpit (buccaneer) issued from NOISEX-92 data base. The proposed scheme was successfully compared to similar recognizers based on the Multilayer Perceptron (MLP), the Elman Recurrent Neural Network (RNN) and the discrete Hidden Markov Model (HMM). The experimental results show that the use of nonparametric regression with an appropriate smoothing factor (spread) improved the generalization power of the neural network and the global performance of the speech recognizer in noisy environments.*

Keywords: *Arabic Digits, General Regression Neural Network, Hidden Markov Model, Non Parametric Density Estimation, Speech Recognition, Noisy Environment.*

1. INTRODUCTION

In the past two decades, various systems have been tested in Automatic Speech Recognition (ASR). They are usually based on the stochastic approach using the Hidden Markov Model (HMM) that provides a mathematically rigorous approach to the development of statistical speech models [1], [2]. The HMM-based methods are suitable for acoustic modelling but suffer from intrinsic limitations, mainly due to their arbitrary parametric assumption and the complexity to estimate those parameters.

A promising technique for speech recognition is the hybrid based approach, which combines the function capabilities of Artificial Neural Networks (ANNs) with the modelling power of HMM. ANNs have been integrated into hybrid HMM/ANN models to compute emission posterior state probabilities [3],[4]. The best known approach is the one proposed by Bourlard [5], [6]. Rigoll [7] has also proposed another hybrid approach where the ANN has been used as a vector quantizer for discrete HMM. If for continuous speech recognition, these methods were essential for the isolated word recognition (e.g., digits which represent shorter units), other directions need to be explored. For this purpose, neural networks, which have great discrimination ability, can be particularly adapted to spoken words recognition.

Speech recognition modelling by ANNs does not require a prior knowledge of the speech process. Neural networks, such as the multilayer feed-forward networks (MLPs) or the Recurrent Neural Networks (RNN), can be trained to associate unknown input data to learned words. As recognizers, ANNs have been shown to yield better performance than HMM on short isolated speech

units [8].

The neural network recognizer based on a static network, such as MLP, and a dynamic network like RNN [9] or Time Delay Neural Network (TDNN) [10], uses parametric representation of the activation function. This assumption can be relaxed by introducing a nonparametric technique. Nonparametric techniques in pattern recognition can be used when no functional form for the density function is assumed. The density of estimates is driven by the data without making any assumption on the form of the distribution [11]. In [12], [13], an adaptation scheme based on nonparametric regression using GRNN is presented. GRNN has been used in a variety of applications, including prediction, plant process modelling and control, and general mapping problems [14]. Some comparative studies have been also published to demonstrate the modelling capability of the GRNN with respect to the other types of neural networks. Works have been reported in the speech field [15], [16], on the detection of the human emotion in speech [17], and on the speech music classification [18]. Moreover, GRNNs have been not extensively used in speech recognition, particularly for Arabic language.

Arabic is currently one of the most widely spoken languages used in the world with an estimated number of 300 million speakers and covers a large geographical area. The major works related to speech recognition in Arabic language deal with the morphological structure [19], [20] or the phonetic features in order to identify the particular Arabic phonemes (pharyngeal, geminate and emphatic consonants) [21], [22] and discuss their further implication in a larger vocabulary speech system. This opened a very interesting way for researchers, but the applications in term of implementation of a recognition system dedicated to spoken isolated words or continuous speech are not extensively conducted and only few examples have be discussed. For Arabic language, Shoaib & al. [23] have developed a derivative scheme, named the Concurrent GRNN. The GRNN has been implemented for accurate Arabic phonemes identification in order to automate the intensity- and formants-based feature extraction. The validation tests expressed in terms of recognition rate obtained with clean speech were up to 93.37 %. El Otaibi [9] has developed an isolated word speech recognizer using the RNN. The word accuracy was over 94.5 % in term of recognition rate in speaker independent mode and 99.5% in speaker dependent mode. The spoken data set used was limited to 17 speakers for both training and testing process. Saeed and Namous [24] have proposed an heuristic method of Arabic digit recognition, using the Probabilistic Neural Network (PNN). The recognition rate, obtained in speaker dependent mode with twenty people, was over 98%. In [25], a hybrid method has been applied to Arabic digits recognition. The Fuzzy C-Means method has been added to the traditional ANN/HMM speech recognizer using RASTA-PLP features vectors. The Word Error Rate (WER) was over 14.4%. With the same approach, a method using data fusion gave a WER of 0.8%. However, this method was tested only on one personal corpus and the authors indicated that the obtained improvement required the use of three neural networks working in parallel. Thus, the recognition step would need more time to be achieved compared to the traditional ANN/HMM method. Another alternative hybrid method has been proposed in [26] where the Support Vector Machine (SVM) and the K nearest neighbor (KNN) were substituted to the ANN in the traditional hybrid system. The training phase was made by only 10 persons by gender and the best results, expressed in term of recognition rate, did not exceed 92.72 % for KNN/HMM and 90,62 % for SVM/HMM. In previous work, we have already shown the superiority of the GRNN speech recognizer over the MLP [27] and the HMM [28] in quiet environment.

The main motivation of this work is to develop an isolated word recognition system based on the GRNN, which is a statistical neural network, and to compare the robustness of our speech recognition system with the discrete HMM, the MLP and the RNN recognizers in adverse conditions. The speech data used in this work are the Arabic digits, which are polysyllabic words.

This paper is organised as follows: in section 2, the basic concept of the nonparametric regression and the neural network implementation are recalled. Section 3 describes our proposed adaptation scheme based on GRNN. The experimental results obtained in quiet and adverse conditions are presented in section 4 and discussed in section 5.

2. NONPARAMETRIC REGRESSION

2.1 THEORETICAL FUNDATIONS

Let $f(x,y)$ be the joint continuous probability density function of a vector random variable x, and a scalar random variable y. Let X be a particular measured value of the vector random vector x of elements x_i (i = 1,…,p). The regression of y given X, is given by the conditional expectation of y on X [12]:

$$E[y|X] = \frac{\int_{-\infty}^{+\infty} y.f(X,y)dy}{\int_{-\infty}^{+\infty} f(X,y)dy} \quad (1)$$

In nonparametric density estimation, no fixed parametrically-defined shape for the estimated density is assumed. Then, the probability density function must be estimated empirically from a sample of observations (data points) of x and y.

The general form of the estimator is given by the following equation [11]:

$$f_n(x) = \frac{1}{n\lambda}\sum_{i=1}^{n}\varphi\left(\frac{x - x_i}{\sigma}\right) \quad (2)$$

where the x_i are independent, identically distributed random variables with absolutely continuous distribution function.

One useful shape of the weighting function φ is the Kernel density function (Gaussian). Parzen has shown that these estimators are consistent [11]. They asymptotically converge to the underlying distribution at the sample point when it is smooth and continuous. Parzen's results have also been extended to the multivariate distribution case [11]. Based upon sample values X^i and Y^i of the random variables x and y, a good choice for the probability estimator, as in [12], [13] is given by:

$$\hat{f}(X,Y) = \frac{1}{(2\pi)^{(p+1)/2}\sigma^{p+1}}\cdot\frac{1}{n}\sum_{i=1}^{n}\exp\left[-\frac{(X - X^i)^T(X - X^i)}{2\sigma^2}\right].\exp\left[-\frac{(Y - Y^i)^2}{2\sigma^2}\right] \quad (3)$$

where p is the dimension of the random vector x, n the number of observations (pattern sample), σ the smoothing factor (spread) of the estimating kernel factor, and Y^i the desired scalar output given the observed input X^i.

Let us define the scalar function D_i^2 as

$$D_i^2 = (X - X^i)^T(X - X^i) \quad (4)$$

Combining (3) and (4) and interchanging the order of integration and summation, yields the desired conditional mean, expressed as

$$\hat{Y}(X) = \frac{\sum_{i=1}^{N} Y^i \exp\left(-\frac{D_i^2}{2\sigma^2}\right)}{\sum_{i=1}^{N}\exp\left(-\frac{D_i^2}{2\sigma^2}\right)} \quad (5)$$

The resulting regression (5), known also as the Nadaraya-Watson kernel regression estimator [11] – [13], is directly applicable to problems involving numerical data. The estimate $\hat{Y}(X)$ can be considered as a weighted average of all the observed values, Y^i, where each observed value is weighted exponentially according to its Euclidean distance from X [12].

2.2 NEURAL IMPLEMENTATION

General regression neural network implementation was firstly proposed by D. Specht [12], [13]. Let w_{ij} be the target output corresponding to the input training vector x_i and the j^{th} output. (5) can be expressed as

$$y_j = \frac{\sum_{i=1}^{n} w_{ij}.h_i}{\sum_{i=1}^{n} h_i} \quad (6)$$

$$\text{with } h_i = \exp\left(-\frac{D_i^2}{2\sigma^2}\right) \quad (7)$$

According to (6) and (7), the topology of a GRNN described in Fig. 1 consists on

- An input layer (input cells), which is fully connected to the pattern layer
- A pattern layer which contains one neuron for each pattern. It computes the pattern functions $h_i(\sigma, C_i)$ expressed in (7) using the centres C_i.
- A summation layer which has two units: N and D. The first unit, which has input weights equal to X^i, computes the numerator N by summing the exponential terms multiplied by the Y^i associated with X^i. The second unit has input weights equal to 1. Thus, the denominator D is the summation of the exponential terms only.
- Finally, the output unit divides N by D to provide the prediction result.

The choice of the smoothing factor is very important. When σ is small, only few samples play a significant role. If σ is large, even distant neighbours can affect the estimate at X.

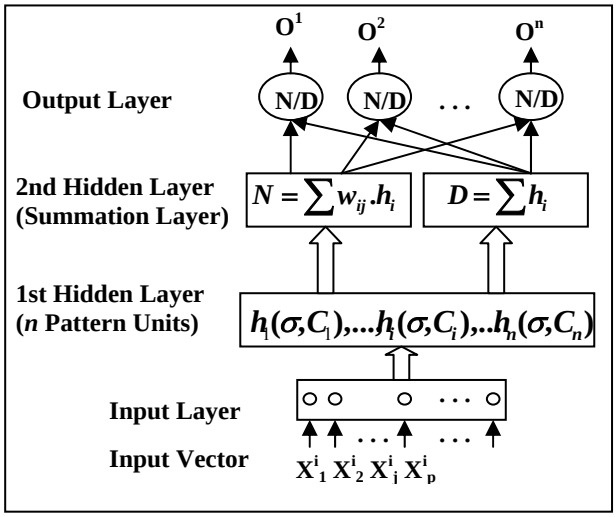

Fig.1 – Neural network implementation based on the nonparametric density estimation method.

3. GRNN-BASED SYSTEM FOR SPEECH RECOGNITION

The general scheme of the proposed speech recognizer is depicted in Fig. 2. It contains two parts: the preprocessing step and the recognition step based on the learning and recognition tasks.

3.1 PREPROCESSING STEP

The speech signal is firstly digitized and end pointed. In order to flatten the signal, the digitized speech signal, pre-emphased by a first-order digital filter, is given as

$$\overset{\Lambda}{s}(n) = s(n) - \mu.s(n-1) \qquad (8)$$

where the pre-emphasis parameter μ is equal to 0.96, and where $s(n)$ and $\overset{\Lambda}{s}(n)$ are the n^{th} speech sample before and after pre-emphasis, respectively.

Pre-emphasis ensures that, in the frequency domain, all the formants of the speech signal have similar amplitudes so that they get equal importance in subsequent processing stages. Then, the signal is fragmented into frames by using a Hamming window (256 points with half covering). In order to reduce the amount of the information in the speech signal, the frame features are extracted using the Mel Frequency Cepstrum Coefficients (*MFCC*), the most popular features for ASR. For each frame, the features extraction consists of 12 *MFCC* along with their first and second derivatives (dynamic features) and the log(energy). The j^{th} frame of the word W_i is represented by an acoustic vector S_{ij}

$$S_{ij} = \{MFCC, \Delta MFCC, \Delta(\Delta MFCC), log\ (Energy)\} \qquad (9)$$

3.2 RECOGNITION STEP

3.2.1 LEARNING TASK

The feature vectors represent the inputs of the GRNN used as recognizer, as shown earlier in Fig. 1. The input vector corresponding to the first word in the learning set is used to compute first the pattern function $h(\sigma,C_i)$ expressed in (7) and then, the output of the neural network expressed in (6). Finally, this first pattern is memorized. For the following words in the learning phase, which is a sequentially process, only the new patterns are memorized.

3.2.2 RECOGNITION TASK

When presented with features of unknown word, the distance between the unknown word and each pattern memorized in the hidden layer is computed and passed through a kernel function. The output of such kernel function is an estimate of how likely the unknown pattern of a word belongs to the pattern distribution stepped in the hidden layer.

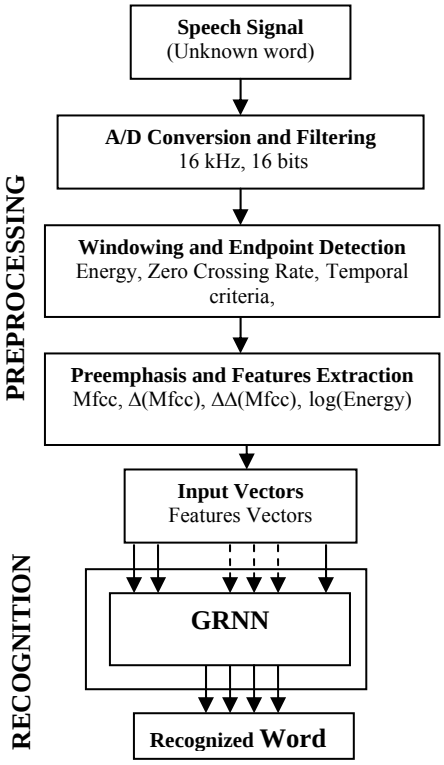

Fig.2 – General scheme of the isolated word recognition system using GRNN.

4. EXPERIMENTAL RESULTS

4.1 SPEECH DATABASE

The used spoken words are the Arabic digits. Standard Arabic has basically 34 phonemes, of which six are vowels and 28 are consonants. It has three long and three short vowels, while there are twelve vowels for American English and at least

twelve oral and four nasal vowels for French language. Arabic utterances can only start with a consonant. The allowed syllables in Arabic languages are: CV, CV, CVC, and CVCC, where V indicates a long or short vowel and C indicates a consonant. The particularity of this language is that the Arabic phonemes contain two distinctive classes, namely the pharyngeal and emphatic phonemes, which can be found only in Semitic languages like Hebrew, Persian and Urdu. All Arabic digits are polysyllabic words, excepted for the digit zero which is a monosyllabic. The Arabic spoken digits, subscribed in API code (International Phonetic Alphabet), are reported in Table 1, second column.

The database used for training and testing the recognition system is a locally Arabic speech data base collected from 150 Algerian natives aged from 18 to 50. This database, named ARADIGITS, has been recorded in a large auditory room, which was very quiet, at 22.050 kHz and down-sampled at 16 kHz.

In the training phase, a total of 1800 utterances pronounced by 90 speakers of both sex (equally distributed) were used. In the testing phase, 1000 utterances pronounced by 50 others speakers of both gender (25 males and 25 females) were used. The experiments were conducted in speaker independent mode; therefore, the data of the testing set do not intersect with those of the training set.

4.2 EXPERIMENTAL FRAMEWORK

A set of experiments have been conducted to test the accuracy by measuring the ASR performance. All recognition results are given in term of WER, defined as:

$$WER = \frac{N - R}{N} \times 100\% \qquad (10)$$

where N is the total number of words in the test set, and R the total number of words correctly recognized in the test set.

The clean speech material is used to train the speech recognizer system. In order to compare the performance of the GRNN speech recognizer, we have interchanged the GRNN classifier by the MLP, the Elman Recurrent Neural Network (RNN) and the HMM. The HMM baseline system achieved is the five states left-to-right model. The recognition experiments were conducted in both clean speech and adverse conditions in order to detect the intrinsic robustness of the recognizer. The optimization of the smoothing factor is critical to the GRNN performance and is usually found through iterative adjustment and cross-validation procedure. In previous works, we have shown that the suitable interval for speech recognition is $14 < \sigma < 20$, and that $\sigma = 15$ is a convenient value [27], [28]. The results obtained in quiet environment are reported on table I for both genders.

Table 1. Comparative study using HMM, MLP, RNN and GRNN speech recognizers.

		Word Error Rate (%)			
	Digit	HMM	MLP	RNN	GRNN
0	şifr	7	6	7	4.0
1	wa:ħid	5	1	1	0.0
2	?iθna:n	44	4	3	1.0
3	θala:θa	24	7	9	4.0
4	?arbaça	10	1	3	0.0
5	χamsa	8	0	3	0.0
6	sitta	2	0	0	1.0
7	Sabça	24	0	0	1.0
8	θama:nija	27	8	3	3.0
9	Tisça	15	0	0	0.0
	Overall	16.6	2.7	2.9	1.4

With the GRNN speech recognizer, the digits 1, 4, 5 and 9 are correctly recognized (WER = 0%), as shown in table 1. The digits 2, 6 and 7 are recognized with 1% WER, and digits 0, 3 and 8 present a WER of 4%, 4% and 3%, respectively. The global WER is 1.4 % for both genders, 0.8% for the female speakers and 2% for the male speakers.

For the MLP, the spoken digits 5, 6, 7 and 9 are correctly recognized; the digits 1 and 4 are recognized with a WER of 1%. For the remainders, WER ranges between 4% and 8%. The global performance is 2.7% WER, 1.6% for the female speakers and 3.8% for the male speakers.

For the RNN, only the spoken digits 6 and 7 are recognized without error. The digit 1 was recognized with 1% WER, the digits 2, 4, 5 and 8 are recognized with a 3% WER, whereas 0 has 7% WER and 3 has 9% WER. The global performance is 2.9% WER for the RNN, 1.4% for the female speakers and 4.4% for the male speakers.

For the HMM, only the digit 6 is recognized with 2% WER, whereas for the other digits the error rate varies from 5% to 44%. For the HMM based speech recognizer, the global performance is 16.6% WER, 15.4% for the female speakers and 17.8% for the male speakers.

4.3 SENSITIVITY TO NOISE

For real word applications, a speech recognition system must operate in situations where it is not possible to control the acoustic environments. This may result in a serious mismatch between the training and test conditions. Differences in the acoustic environments may result from additive noise (background noise: car noise, babble …), convolutive distortion (such as transmission channel distortion, room reverberation, microphone distortion …) or any combination of them. Both

classes of noise have been found to degrade seriously the speech recognition performance. The problem of minimizing the degradation in performance is the problem of robustness. The approaches that may be used to enhance the robustness of an ASR system are usually classified into two types: the noise reduction techniques [29], [30] and the acoustic model adaptation techniques [31]. The problem discussed here is the sensitivity to noise of the speech recognizers, or in other words, the intrinsic robustness to noise (inherently robustness).

The observed signal corrupted by additive noise can be represented as

$$x(t) = s(t) + n(t) \quad (11)$$

where $x(t)$, $s(t)$ and $n(t)$ denote observed noisy speech, clean speech signal and additive signal noise, respectively.

The additive noise is classified into two types, i.e., the stationary noise that has time constant characteristics of spectral features and signal energy, and the non-stationary noise that has time variable characteristics of spectral features and signal energy. An example of the degradation of an original signal by a non-stationary noise is given in Fig. 3. Fig. 3a shows the temporal form of the original clean signal of the word "/sitta/" (Arabic digit 6). Fig. 3b shows the temporal form of the same spoken word corrupted by a noise signal recorded in a factory (car production hall) at SNR = 5dB. The transitions from silence to speech did not only vary but also some speech parts might be masked by noise. It is noted that the stopped plosive /t/ and the fricative /s/ are completely masked by the noise signal.

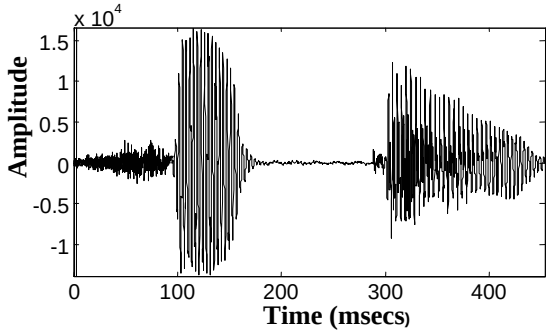

Fig. 3a – Speech signal of the word "/sitta/" (Arabic digit 6), recorded in quiet environment.

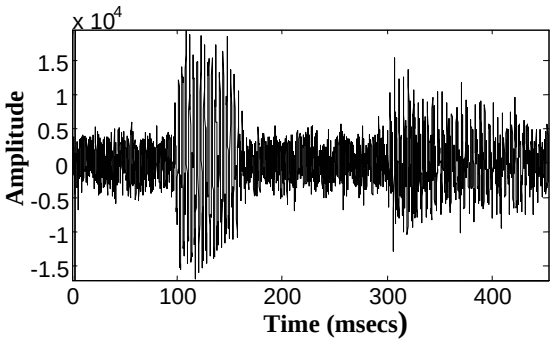

Fig. 3b – Noisy speech signal of the word " /sitta/" corrupted with the noise signal recorded in a car production hall at SNR= 5dB.

To evaluate the robustness of the speech recognizer in various kinds of noise, stationary as well as non stationary noises, issued from the NOISEX-92 database, were added to the testing database in a range of SNRs from 0dB to 20dB, step size 5 dB. We were particularly interested by four kinds of noises: two stationary and two non stationary. For the stationary noise, we considered the military vehicle noise acquired by recording noise signal from the leopard vehicle moving at 70 km/h, and the cockpit noise acquired from the fighter jet (buccaneer moving at a speed of 450knots). For the non stationary noise, we considered the speech babble acquired from 100 people speaking in canteen, and the factory floor noise recorded in a car production hall. The power spectrum densities of each noise signal are presented in Figs. 4 to 7.

The comparative performance of the four types of recognizers in these various adverse conditions is given in Figs. 8 to 11. The first observation is that the speech recognizers alone are not able to adapt themselves with environmental noisy conditions if the SNR is less than 15 dB, because the rise of WER curves increases strongly after this threshold level. The spectral features of the background noise are a significant element for the speech recognition, much more than the SNR level. For example, the degradation caused by the fighter jet noise, characterized by a broadband spectrum, at 20 dB SNR exceeds the degradations caused by the military vehicle noise, located in the low frequency narrow band, at 0 dB SNR level. Figures 8 to 11 show that the less harmful is the vehicle noise, which have a narrowband spectrum. The most significant degradations are caused by the fighter jet noise whose spectral broadband can cover the whole speech spectrum, masking completely the speech, even on high level SNR.

As shown in Fig. 8, the transition from a quiet condition to a noisy environment, with a speech babble background at SNR=15 dB, is followed by a

loss of accuracy of 13.6% for the GRNN, 15.3% for the MLP, 20.1% for the RNN and more than 28% for the HMM. At 10dB SNR, the accuracy drops by 22.6%, 23.3%, and 34.1% and more than 49%, respectively. Fig. 8 shows that the GRNN is the least sensitive to the environmental condition degradation; the MLP exhibits a performance close to the GRNN and better than of the RNN. Finally, the HMM has the most degraded performance with this type of noise.

For the noise factory, which is also a non-stationary noise, the GRNN still gives the best results as shown in Fig. 9. The loss of effectiveness when one passes from quiet environment into a factory environment producing a noise at 15dB SNR, is 11.6 % for the GRNN, 14.3% for the MLP, 14.1% for the RNN and 41% for the HMM. This loss of effectiveness reaches respectively 19.6%, 25.3%, 25.1% and 48% for the four speech recognizers, if one passes from quiet environment into a factory environment producing noise at 10dB SNR. It is noticed that the ANN based recognizer working in adverse conditions caused by factory noise, in particular the GRNN, exhibits a performance at SNR = 15dB similar to the HMM based recognizer working in quiet environment.

Fig. 10 shows that the fighter jet noise, which is a stationary type, is the most harmful of the four noise conditions studied in this paper, because of its spectral broadband. Indeed, the loss of effectiveness at SNR=15dB reaches to 35.6% for the GRNN, 41.3% for the MLP, 48.1% for the RNN and 66% for the HMM. The degradation is very significant, even if the GRNN presents the best results.

Finally, the military vehicle noise, which is a stationary noise with a low frequency limited bandwidth, did not affect seriously the speech signal, particularly the fricative and pharyngeal phonemes, which are the most present consonants in Arabic digits.

5. DISCUSSION

The experimental results show that the global performance of the GRNN-based speech recognizer is 1.4% WER. This error rate is lower than that obtained by the MLP-, the RNN- and the HMM-based speech recognizers. In fact, the improvement is 1.3%, 1.5% and 15.2%, respectively, relatively to the MLP (WER 2.7%), the RNN (WER 2.9) % and the HMM (WER 16.6%).

The improvement is significant (over 13%) compared to the ANN/HMM hybrid method used by Lazli and Sellami [25]. The obtained results are also better than those obtained in [26], where two others hybrid methods were used namely, KNN/HMM and SVM/HMM, showing an improvement of 6% and 8% respectively. Compared to the RNN based speech recognition system proposed by Alotaibi [9], the improvement is over 6%. Notice that the speech recognition system that we have achieved for comparative study performs better than those presented in [9].

It appears clearly that the GRNN-based speech recognizer has better performance than the other neural networks based recognizers. This improvement of the effectiveness results from the use of a statistical nonparametric function in the GRNN. Furthermore, this results exceed those obtained with the HMM baseline system. The discrete HMM is unable to adapt itself to the variability of the words, in particular the longest words like the spoken digits 2, 3, 7 and 8, which undergo a syntactic modification according to the geographical origin of the speakers. This is due to the fact that the alignment frame/state or the state probability estimate and the transitions between states are made very complex.

The intrinsic robustness of the speech recognizer is studied in four different noisy conditions, using additive stationary and non-stationary noises. The spectral features, particularly the spectral broadband of the background noise, play a significant role in the performance degradation. For instance, the weakest performance was encountered with the fighter jet noise, which has a large spectral broadband. We can also deduce that the GRNN-based speech recognizer is the least sensitive to the background noise present in adverse conditions.

The major problem of the neural network classifiers is the static dimension of their input. The time alignment procedure used to normalize the acoustics vectors is not adequate for modelling the speech process and the neural network approach is unable to model the continuous speech process. Then, the proposed technique cannot be extended to the continuous speech recognition. Moreover, for isolated speech recognition this method is a successful alternative to the HMMs based techniques.

6. CONCLUSION

In this work we have proposed a GRNN adaptation scheme for spoken word recognition. The efficiency of our approach has been demonstrated through a comparative study with the MLP, the RNN and the discrete HMM speech recognizers. The use of a nonparametric density estimator with an appropriate smoothing factor improves the generalization capability of the neural network. Experimental results obtained with large corpora have shown that the proposed model present several advantageous characteristics such as (i) the training

process which is performed at one pass, (ii) the fast learning capability, (iii) the flexibility network size and (iv) the ability to adapt to speaker variability. The GRNN speech recognizer gives the best results in free noisy or quiet environment. The inherently robustness of the GRNN adapted scheme could significantly improve the recognition accuracy in adverse environments, including stationary and non stationary noises. The ANN-based speech recognizers confirm their discrimination capacity and remain a serious alternative to the HMM for the isolated word recognition. GRNN is a successful alternative to the other neural networks and to discrete HMM. It is therefore suitable to be applied in ASR systems.

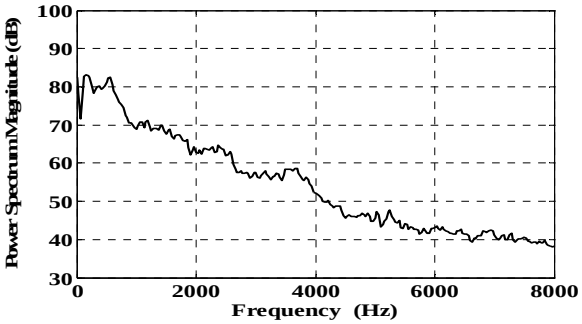

Fig 4 – Power spectrum of the multi talkers babble noise signal recorded in a canteen.

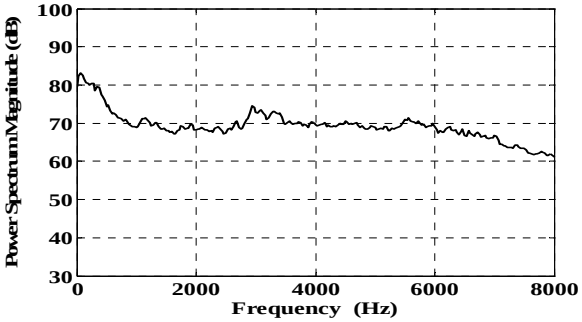

Fig 5 – Power spectrum of the fighter jet noise signal

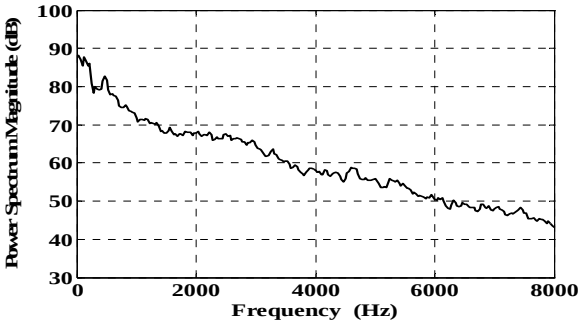

Fig. 6 – Power spectrum of the factory noise signal recorded in a car production hall.

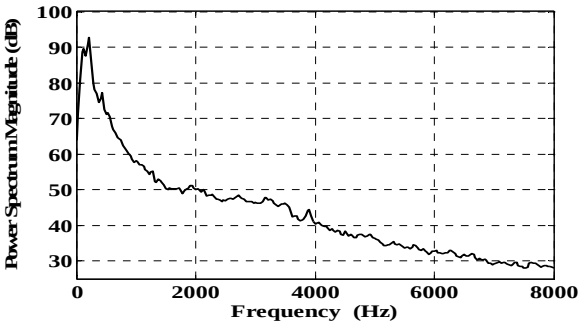

Fig. 7 – Power spectrum of noise signal from a military vehicle moving at 70 km/h.

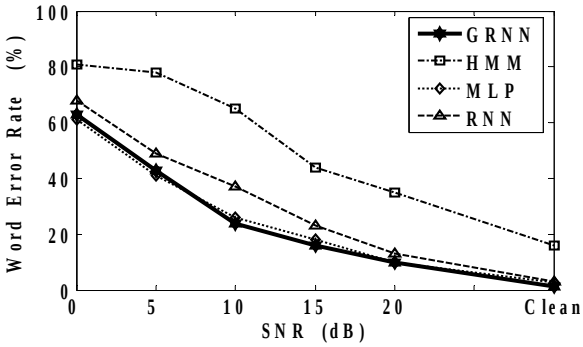

Fig. 8 – Comparative performance with the babble noise.

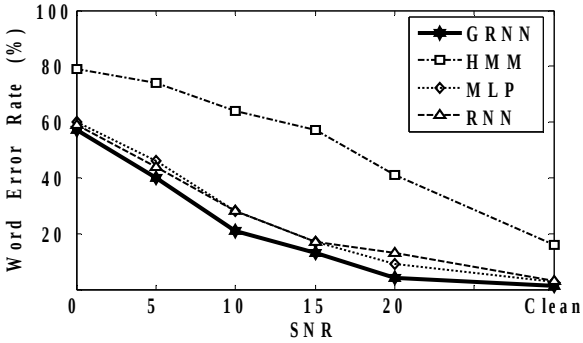

Fig. 9 – Comparative performance with the factory noise.

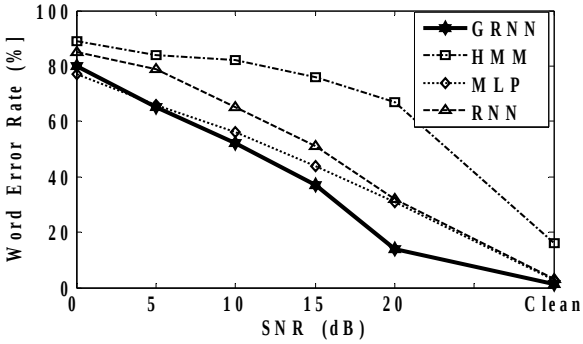

Fig. 10 – Comparative performance with the fighter jet noise (buccaneer).

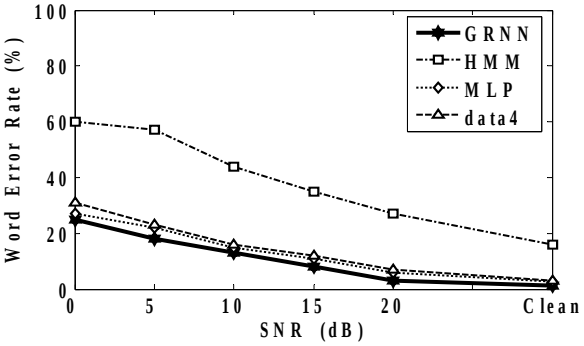

Fig. 11 – Comparative performance with the military vehicle noise (leopard tank).

7. REFERENCES

[1] L. Rabiner. A Tutorial on Hidden Markov Model and Selected Applications, *Proc. of the IEEE*, 77 (2), (1989), pp. 257-286.

[2] F. Jelinek. *Statistical Methods for Speech Recognition*, Cambridge, Massachusetts, MIT Press, 1997.

[3] S. Haykin. *Neural Networks: A Comprehensive Foundation*", 2nd ed., Cliffs, NJ, 1999.

[4] R. P. Lippman. Review of Neural Networks for Speech Recognition, *Neural Computation* 1 (1989), pp.1-38.

[5] H. Bourlard, N. Morgan. *Connexionist Speech Recognition: A Hybrid Approach*, Kluwer Academic Press, 1994.

[6] S. Renals, N. Morgan, H. Bourlard, M. Cohen H. Franco. Connectionist Probability Estimators in HMM Speech Recognition, *Proc. of the IEEE ICASSP'98*, 12-15 May 1998, Seattle, USA, pp. 9-12.

[7] G. Rigoll and C. Neukirchen. A New Approach to Hybrid HMM/ANN Speech Recognition Using Mutual Information Neural Networks, *Advances in Neural Information Processing Systems (NIPS-96).* 3-5 Dec. 1996, Denver, USA, pp. 772-778.

[8] H. Bourlard, N. Morgan. Connexionist Techniques, available at: http://cslu.cse.ogi. edu/ HLT survey/ch11node7.html, March 2003.

[9] Y. A. Alotaibi. Investigation of Spoken Arabic Digits in Speech Recognition Setting, *Informatics and Computer Sciences* 173(1-3), (2005), pp. 105-139.

[10] A. Waibel, T. Harazawa, G. Hinton, K. Shakano, K.J. Lang. Phoneme Recognition using Time-Delay Neural Networks. *IEEE Trans. on ASSP,* 37 (3), (1989), pp. 328–339.

[11] T. Cacoulos. Estimation of a Multivariate Density, *Ann. Inst. Math. Tokyo*, 18 (2), (1966), pp. 179–189.

[12] D. F. Specht. A General Regression Neural Networks, *IEEE Trans. on Neural Networks,* 2 (6), (1991), pp. 568–576.

[13] D.F. Specht, *Probabilistic Neural Networks and General Regression Neural Networks*, *Fuzzy Logic and Neural Network Handbook, Chap3*. Mac Graw Hill Inc. 1996.

[14] L. Rutkowski. Generalized Regression Neural Networks in Time Varying Environments. *IEEE Trans. on Neural Networks*, 15 (3), (2004), pp. 576-596.

[15] K. Chung and R. Tognieri. Extraction of Speech Signal in the Presence of Musical Note Signal Using the Generalized Regression Neural Networks, *Proc. of the Sixth Australian Conf. on Speech Sciences and Technology,* Dec.1996, Adelaide, Australia, pp. 133-137.

[16] T. Hoya and A. G. Constantinides, An Heuristic Pattern Correction Scheme for GRNNs and its Application to Speech Recognition, *Proc of the IEEE Workshop on NNSP'98*, , 31 Aug- 02 Sept. 1998, Cambridge, U.K., pp. 351-359.

[17] M. W. Bhatti, W. Yongin,G. Ling. A Neural Network Approach for Human Emotion Recognition in Speech, *Proc. of the ISCAS 2004,* (2), 23-26 May 2004, Vancouver, Canada, pp. 181-184.

[18] B. Bolat and O. Kucuk. Speech Music Classification by Using Statistical Neural Networks, *Proc. of the. 12th IEEE Signal Proc. and Com. Appl. Conf.* 28-30 April 2004, Kusadasi, Turkey, pp. 227-229.

[19] S. Datta, M. Al Zabibi, O. Farook. Exploitation of Morphological in Large Vocabulary Arabic Speech Recognition, *Int. Journal of Computer Processing of Oriental Language* 18(4), (2005), pp. 291-302.

[20] K. Kirschoff & al**.** Novel Approach to Arabic Speech Recognition, Final Report from the JHU Summer School Workshop, 2002, *Proc. of the Int. Conf. on ASSP (ICASSP'03)*, 6-10 April 2003, Hong Kong, pp. 344-347..

[21] M. Debyeche, J. P. Haton, A. Houacine. A New Vector Quantization Approach for Discrete HMM Speech Recognition System, *Int. Scientific Journal of Computing* 5 (1), (2006), pp. 72-78.

[22] S.A. Selouani, J. Caelen. Arabic Word Recognition by Classifiers and Context, *Journal of Computer Science and Technology* 20 (3), (2005), pp. 402-410.

[23] M. Shoaib, M. Awais, S. Masud, S. Shamail, J. Akhbar. Application of Concurrent Generalized Regression Neural Networks for Arabic Speech Recognition, *Proc. of the IASTED Int. NCI*

2004, 23-25 Feb. 2004, Grindelwald, Switzerland, pp. 206-210.

[24] K. Saeed. M. Nammous. Heuristic Method of Arabic Speech Recognition, *Proc. of the IEEE Int. Conf. on Digital Signal Processing and its Applications (IEEE DSPA'05),* Moscow, Russia, 2005, pp. 528-530

[25] L. Lazli. M. Sellami. Connectionist Probability Estimation in HMM Arabic Speech Recognition Using Fuzzy Logic, *Lectures Notes in LNCS (2734),* (2003), pp. 379-388.

[26] H. Bourouba, R. Djemili, M.Bedda, C. Snani. New Hybrid System (Supervised Classifier/HMM) for Isolated Arabic Speech Recognition. *Proc. of the 2nd. IEEE Int. Conf ICTTA'06,* 24-28 April 2006, Damascus, Syria, pp. 1264-1269.

[27] A. Amrouche. J.M. Rouvaen. Arabic Isolated Word Recognition Using General Regression Neural Network., *Proc. of the 46th. IEEE Int. MWSCAS'03*, 27-30 Dec. 2003. Cairo, Egypt, pp. 689-692.

[28] A. Amrouche. J.M. Rouvaen. On the Use of the Nonparametric Regression in Neural Network Based Approach Applied to Arabic Speech Recognition. *Proc. of the 9th. Int. Conf. Speech and Computer (SPECOM'2004),* 20-22 Sept. 2004, St Petersburg, Russia. pp. 276-281.

[29] X. Cui. A. Alwan. Noise Robust Speech Recognition Using Feature Compensation Based on Polynomial Regression of Utterance SNR, *IEEE Trans. on Speech and Audio Processing*, 13(6), (2005), pp. 1161-1172.

[30] U.H. Yapanel. J.H.L. Hansen. A New Perspective on Feature Extraction for Robust In-Vehicle Speech Recognition. *Proc. of the 8th (Eurospeech'03)*, 1-4 Sept. 2003, Geneva, Switzerland, pp. 1281-1284.

[31] C. Kermorvant. A. Morris. A Comparison of Two Strategies for ASR in Additive Noise: Missing data and Spectral Subtraction, *Proc. of the 6th European Conf. on Speech Communication and Technology (Eurospeech'99)*, 5-9 Sept. 1999, Budapest, Hungary, pp. 2841-2844.

Abderrahmane Amrouche *was born in Algeria. He received the "Diplôme d'Ingénieur" degree in Electronics from the Ecole Nationale Polytechnique, Algiers in 1980 and the "Magister" degree in 1995. Since 1982, he is with the University of Sciences and Technologies of Algiers-USTHB- as a senior lecturer and scientist researcher in Speech communication laboratory. His research interests include pattern recognition, speech processing, multilingual speech recognition, neural networks, prosodic modeling.*

Abdelmalik Taleb-Ahmed *was born in Roubaix, France, in 1962. He received a Post graduat degree and a Ph.D in Electronics and Microwaves from the Université des Sciences et Technologies de Lille 1, France, in 1989 and 1992, respectively. From 1992 to 2004, he was an Associate Professor at the Université du Littoral Cote d'Opale, Calais, France. He is currently a Professor at the Université de Valenciennes et du Hainaut Cambresis-UVHC-, and does his research at the LAMIH UMR CNRS 8530. His research interests includes signal and image processing.*

Jean Michel Rouvaen *was born in 1947 in France. He received M.S degree in 1968 and his PhD in 1971 from the Université de Valenciennes et du Hainaut Cambresis-UVHC, (France). He is now Professor of electronics at ENSIAME, an engineering school of UVHC and he is the head of Radio- communications, Detection and Signal processing research group at OAE–IEMN Institute (France). His primary interests are in nonlinear phenomena, speech processing, and signal processing for communication systems…*

Mustapha C. E. Yagoub *received the "Diplôme d'Ingénieur" degree in Electronics and the Magister degree in Telecommunications, both from the Ecole Nationale Polytechnique, Algiers, Algeria, in 1979 and 1987 respectively, and the Ph.D. degree from the INP-Toulouse, France, in 1994. He joined the USTHB during 1983 as Assistant and then Assistant Professor during 1994-1999. In 2001, he joined the School of Information Technology and Engineering (SITE), University of Ottawa, Ottawa, ON, Canada, where he is currently an Associate Professor. His research interests include neural networks modeling, RF/microwave device/system CAD, planar antennas, and applied electromagnetics. Dr. Yagoub is an IEEE senior member and a member of the Professional Engineers of Ontario, Ontario, Canada.*