American Journal of Intelligent Systems 2012, 2(1): 1-8
DOI: 10.5923/j.ajis.20120201.01

A Comparitive Survey of ANN and Hybrid HMM/ANN Architectures for Robust Speech Recognition

Mondher Frikha*, Ahmed Ben Hamida

Advanced Technologies for Medical and Signals (ATMS) Research Unit, National School of Engineering of Sfax, B.P.W, Sfax ,Tunisia

Abstract This paper proposes two hybrid connectionist structural acoustical models for robust context independent phone like and word like units for speaker-independent recognition system. Such structure combines strength of Hidden Markov Models (HMM) in modeling stochastic sequences and the non-linear classification capability of Artificial Neural Networks (ANN). Two kinds of Neural Networks (NN) are investigated: Multilayer Perceptron (MLP) and Elman Recurrent Neural Networks (RNN). The hybrid connectionist-HMM systems use discriminatively trained NN to estimate the a posteriori probability distribution among subword units given the acoustic observations. We efficiently tested the performance of the conceived systems using the TIMIT database in clean and noisy environments with two perceptually motivated features: MFCC and PLP. Finally, the robustness of the systems is evaluated by using a new preprocessing stage for denoising based on wavelet transform. A significant improvement in performance is obtained with the proposed method.

Keywords Speech Recognition, HMM, ANN, MLP, RNN, Hybrid Sys

1. Introduction

Most of the current state of the art automatic speech recognition (ASR) systems are probably based on the use of a continuous density hidden Markov models (HMM) of which functionality is based on a rigorous probability theory[16],[17]. This was basically due to the efficiency with which HMM model the variation in the statistical properties of speech, both in the time and the frequency domains[15]. Also, the major advantages on the use of such models rely on their relatively fast estimation of their parameters from training data. Dynamic programming could be then used effectively to reduce computational complexity. Furthermore, their performance in terms of recognition accuracy is very high, even when computational efficiency requirements are very strict. One drawback of HMM is the various conditional independence assumptions imposed by the Markov Model. These assumptions essentially state that each speech frame is independent of its neighbours.

However, over the last few years, several attempts have been undergone to evaluate the HMM deficiencies. Artificial Neural Networks (ANN) and more specifically multilayer perceptrons (MLP) appeared to be a promising alternative in this respect to replace or help HMM in the classification mode. So, a number of ANN approaches have been suggested and used to improve the state of the art of ASR systems[20],[24]. The fundamental advantage of such approach is that it introduces a discriminative training[18]. The two main drawbacks of NN systems is their increased training computational requirements as well as their incapacity of accommodating time sequences of speech.

A plethora of results indicated that NN can be trained as a probability estimator[10],[21]. This important research finding eased their integration with the HMM current state of the art recognition system technology[2]. This fact led to the possibility of unifying HMM and ANN within unifying novel models[5].

In this study, we combined the advantages of the HMM and the ANN paradigms within a single hybrid system to overcome the limitations of any approach operating in isolation. The goal in this hybrid system for ASR is to take the advantage from the properties of both HMM and ANN improving its flexibility and recognition performance. An hybrid HMM/ANN recognizer that combines efficient discriminative learning capabilities of NN[5] and the superior time warping and decoding techniques associated with the HMM approach was therefore developed[1]. ANN were trained to estimate HMM emission probabilities required in HMM based only on the acoustic information in a limited number of local speech frames[26]. Those probabilities were then used by a Viterbi decoding process for recognition[8],[9]. Two kinds of ANN were investigated: Multilayer perceptrons (MLP) and recurrent neural networks (RNN), to compute posterior probabilities of classes that should be fed into the HMM decoder. The robustness of the constructed

* Corresponding author:
mondher_frikha05@yahoo.fr (Mondher Frikha)
Published online at http://journal.sapub.org/ajis

hybrid ASR system operating under noisy environments was also evaluated and a new preprocessing denoising algorithm based on wavelet transform was proposed.

The remainder of this paper is organized as follows. The state of the art of speech recognition process which is composed of the acoustic analysis and classification modules is reviewed in section 2. In order to study the robustness of the constructed hybrid recognition system, a new preprocessing speech enhancement approach based on wavelet transform is described in section 3. Finally, experiments and results obtained in both clean and noisy environments are presented and discussed followed by some conclusions.

2. Speech Recognition Process

Speech recognition systems have a wide range of applications from isolated-word recognition as in name-dialling and voice-control of machines to continuous natural speech recognition as in auto-dictation or broadcast-news transcription. Most practical speech recognition systems consist of two modules: the front end feature module and back end classification module. Figure 1 shows a general scheme of a speech recognition system.

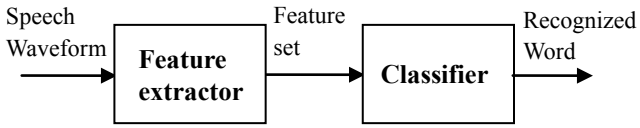

Figure 1. General Scheme of a Speech Recognition System

2.1. Feature Extractor

The design of the front end feature extraction module is a relevant aspect for the performance of the speech recognizer because this module is intended to extract the discriminative information utilized by the classification module to perform recognition. Front end design has been an area of active research in the last few decades. The two front end dominant approaches in speech recognition are based on Mel frequency cepstral coefficient (MFCC)[19] and perceptual linear prediction (PLP)[11]. They are the most widely used acoustic features in current ASR systems. The steps followed in computing those features are detailed in figure 2.

In the case of the speech signal, the feature extractor will first have to deal with the long-term non stationary. For this reason, the speech signal is usually cut into frames of about 10-30ms and feature extraction is performed on each piece of the waveform. Secondly, the feature extraction algorithm has to cope with the short-term redundancy so that reduced and relevant acoustic information is extracted. For this purpose, the representation of the waveform is generally swapped from the temporal domain to the frequency domain, in which the short-term temporal periodicity is represented by higher energy values at the frequency corresponding to the period. Thirdly, feature extraction should smooth out possible degradations incurred by the signal when transmitted on the communication channel. Finally, feature extraction should map the speech representation into a form which is compatible with the classification tools in the remainder of the processing chain. Some classification algorithms will, for example, require a decorrelation of the features.

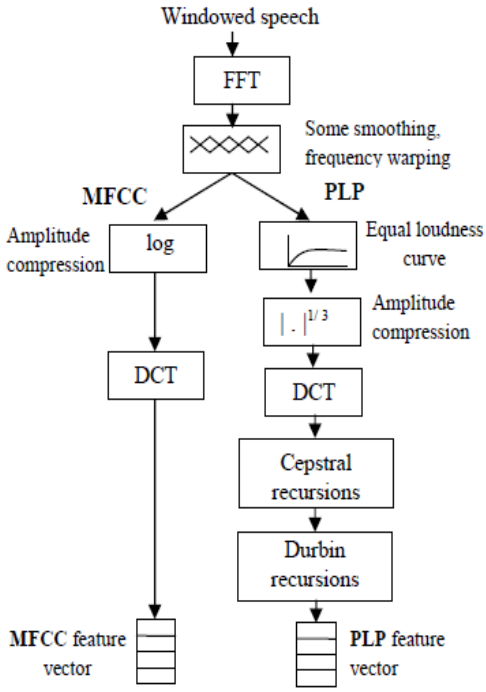

Figure 2. Steps Followed to Compute Mel Frequency (MFCC) and Perceptual Linear Prediction (PLP) Features

2.2. Classification Module

We are interested in this section in three kinds of classifiers. The statistical classifier based on Hidden Markov Models constitutes actually the predominant approach in speech recognition. The connectionist models or artificial neural network (ANN) proposed in recent years as an alternative potential approach to speech recognition systems because of their impressive ability to decorrelate the input features and therefore ameliorate the interclass discrimination. The hybrid connectionist-HMM approach which combines the temporal modeling structure of HMMs with pattern classification capabilities of ANNs. We give a brief overview of these approaches in the next subsections.

2.2.1. HMM Speech Recognition

Hidden Markov modeling of speech, assumes that speech is a piecewise stationary process. That is an utterance is modelled as a succession of discrete stationary states, with transitions. HMM are “hidden” because the state of the model, q, is not observed whereas the output of the stochastic process attached to that state is observed. This is described by a probability distribution P(x/q), where x is the acoustic evidence emitted by state q. The other set of pertinent probabilities are the instantaneous transition probabilities distribution, $a_{ij}=P(q_i/q_j)$, between state i and state j. Figure 3 illustrates a simple Bakis HMM topology.

Essentially, a HMM is a stochastic automaton, with a stochastic output process attached to each state. Thus, we have two concurrent processes: a Markov process modelling the temporal structure of speech and a set of state output processes modelling the instantaneous character of the speech signal. We have around 60 basic phone HMM (for English), and from these we construct word models. For any given sentence, we may write down the corresponding HMM; each state in that HMM is contributed by a constituent phone HMM.

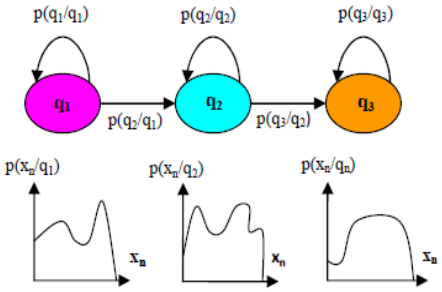

Figure 3. A Schematic of a Three State, Left to Right HMM

The basic problem of speech recognition is to be able to output the correct word corresponding to a spoken utterance. A general approach to this problem is to output the most probable sentence (W) given the acoustic data (X). Thus we must choose word, for which the probability P(W/X), is a maximum. If we choose to use hidden Markov models, then a sentence is represented by a particular state sequence, $Q=q_1q_2\ldots q_n$, and the probability we require is P(Q/X). It is not obvious how to estimate this probability directly. However we may re-express this probability using Bayes rule as follows[15]:

$$P(Q/X)=\frac{P(X/Q).P(Q)}{P(X)} \quad (1)$$

This separates the probability estimation process into two parts: acoustic modeling, in which the data dependent probability density P(X/Q) are estimated; and language modelling in which the *a priori* probabilities of state sequences, P(Q) are estimated. Thus, using the HMM assumptions, we are able to treat acoustic modeling and language modeling independently, using the data dependent and a priori probability estimates.

There are 3 problems to be solved in the HMM: the evaluation of probabilities, the training and the recognition. The solution for the first problem is the use of the Forward Backward algorithm or the Viterbi algorithm[14]. The second one is the use of the Baum-Welch training or the Viterbi training and the third one is the use of the Viterbi algorithm[12] that is going to determine the likeliest hidden state sequence that produces a given sequence of observation.

Finally, we can simplify these problems with maximizing the *a posteriori* probability as follows:

$$W^* = \arg\max_W \left[P(W/X)\right] = \arg\max_W \frac{P(X/W).P(W)}{P(X)} \quad (2)$$

Viterbi Search Algorithm

We have outlined how trained HMM are used to recognize speech. HMM are generative stochastic models of speech. To recognize speech we take an input speech signal, and compute the most probable sequence of models that have generated the speech signal. An efficient algorithm for computing this state sequence is a dynamic programming algorithm known as Viterbi decoding[6]. The Viterbi algorithm essentially traces the minimum cost (or maximum probability) path through a time-state lattice subject to the constraints imposed by the acoustic and language models.

The Viterbi algorithm may also be used in training. In this case a Viterbi alignment is performed for a known word model sequence to obtain the optimal state segmentation. Given this optimal segmentation the output probability distribution function pdf parameters (e.g. means and variances of Gaussians, weights of a MLP, etc…) may be re- estimated[14].

Prior Probabilities

The combination of phone models to form word models is constrained by a phone-structured lexicon that details the allowed pronunciations for each word. In a statistical speech recognition system, the language model will thus assign a prior probability. Since sentences are composed of words, a prior probability is specified for each sentence. Using the allowable pronunciations for each word (which may be probabilistic), prior probabilities are also specified for each phone (and for each state of each phone model). So the specification of the language model, phone-structured lexicon and basic phone HMM, sets the prior probabilities for HMM states, phones, words and sentences. These prior probabilities are encoded in the topology and associated transition probabilities of the hidden Markov word models. It will be important later to distinguish these prior probability estimates from the prior probability estimates from the phone relative frequencies obtained from the training data. We generally do not wish to use the latter since a typical speech training database is much smaller than a typical textual corpus from which the language model is derived.

Connectionnist models for Recognition

Artificial Neural networks (ANN) is a computer system inspired from the organization of cells in the human brain. Multilayer perceptrons (MLP) are the best studied class of ANN frequently applied in speech recognition[20]. They have layered feedforward architecture with an input layer, zero or more hidden layers and an output layer, as shown in Figure 4.

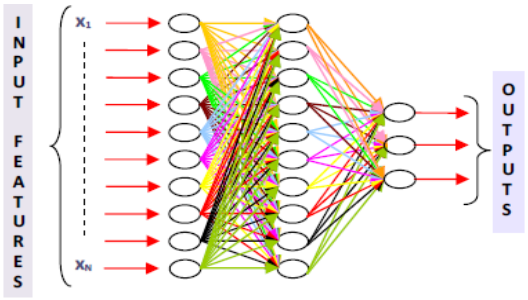

Figure 4. Feedforward MLP Architecture with One Hidden Layer

The obvious way of the use of MLP in speech recognition, is to present, all at once, the acoustic vectors of a speech unit (phoneme of word) at the input layer and to detect the most probable speech unit at the output layer by determining the output neuron with the highest activation. Each layer computes a set of discriminant functions (via a weight matrix) followed by a non linear function, which is often a sigmoid function:

$$f(x)=\frac{1}{1+\exp(-x)} \quad (3)$$

The learning algorithm can be the conventional back-propagation[18], or a more sophisticated variation of it[2]. In the learning phase, the desired output is 1 for the correct and 0 for all other speech units. In this way, not only the correct output is reinforced for the corresponding sequence of acoustic vectors, but simultaneously the wrong outputs are weakened.

Like the human brain, neural network can learn by experience and it has two different techniques of training: supervised and unsupervised. The backpropagation algorithm is intended to minimize the quadratic cost function 'E', according to equation 4:

$$E=\frac{1}{2}\sum_{k}(d_k - s_k)^2 \quad (4)$$

Where d_k is the target (desired output) and s_k is the output of the k^{th} neurone in the network[7].

Further details of the backpropagation algorithm can be found in[18],[20].

2.2.2. Hybrid Recognition System

The principle aim of an artificial neural network hybrid (HMM/ANN) system is to combine the efficient discriminative learning capabilities of neural networks and the superior time warping and decoding techniques associated with the HMM approach.

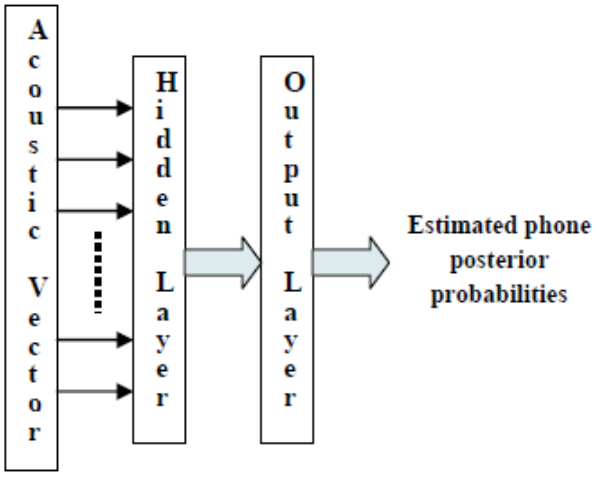

Figure 5. Generic MLP for Posterior Probabilities Estimation

The ANN is trained to estimate HMM emission probabilities which are then used by a decoder based on the well-known Viterbi algorithm. Among the advantages in using such an approach is that no assumption about the statistical distribution of the input features is necessary. Due to its classification procedure, an MLP has the ability to decorrelate the input features. Moreover, while in classical HMM based system, the parameters are trained according to a likelihood criterion, an MLP also penalizes the incorrect classes. Figures 5 and 6 show the general structure of a hybrid HMM/ANN based recognition system.

Figure 6. Schematic Recognition Process using HMM Viterbi Dynamic Programming Search

At every time n, the acoustic vector x_n is presented to the network. This generates local probabilities that are used, after division by priors, as local scaled likelihoods in a Viterbi dynamic programming algorithm[6].

Posterior Probability estimation

MLP may be used to estimate probabilities. Several authors proved that MLP appropriately trained may be used to estimate posterior probabilities of classes. That is, a MLP trained to perform classification is a class-conditional posterior probability estimator. If we associate each output neuron to a determined class C_i, then, the MLP output value, which is given an input X, will be an estimate of the posterior probability $P(C_i/X)$ of the corresponding class C_i given the input[24].

When a MLP is used for speech recognition tasks, the outputs neurons are associated to speech units, such as phones. Thus, if we consider each HMM phone model containing only one HMM state, a phone is equivalent to that HMM state q. Then, since the network outputs approximate Bayesian probabilities, the output of the k^{th} neuron is an estimate of the following posterior probability:

$$P(q_k|x_n)=\frac{P(x_n|q_k)P(q_k)}{P(x_n)} \quad (5)$$

Which contains the *a priori* class Probability $P(q_k)$ as a factor.Finally, scaled likelihoods $P(x_n|q_k)=\frac{P(q_k|x_n)}{P(q_k)}$, which will be used by the Viterbi decoder, are obtained by dividing the network outputs by the prior probabilities of class q_k, which are estimated by computing the relative frequency of the class q_k in the training set. During the recognition $P(x_n)$ is constant.

3. Preprocessing Wavelet Denoising Stage

We propose in this paper, a new pre-processing stage in the speech recognition system to make it robust to four types of noise. The block diagram of the proposed scheme is shown in figure 7.

However, it is well known that, in Fourier based signal processing, the out of band noise can only be removed by applying a linear time invariant filtering approach. But, it

cannot be removed from the portions where it overlaps the signal spectrum. The denoising technique used in the wavelet analysis is based on an entirely different idea and assumes the amplitude rather than the location of the spectrum of the signal to be different from the noise. The localising property of the wavelet is helpful in thresholding and shrinking the wavelet coefficients that helps in separating the signal from noise[4]. The denoising by wavelet is quite different from traditional filtering approaches because it is non-linear, due to a thresholding step.

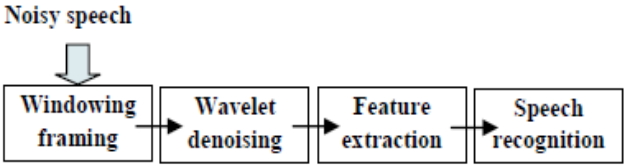

Figure 7. Block Diagram of the Recognition System with Denoising

Suppose that an original signal x(k) of length L is corrupted by a noise m(k) to give the noisy signal y(k) which is given by:

$$y(k) = x(k) + m(k) \quad (6)$$

Thresholding involves the following steps[22]:

- Perform the wavelet transform of the noisy data.
- Calculate the threshold δ depending upon the noise variance.
- Perform thresholding of the wavelet coefficients.
- The coefficients obtained from just the previous step are then padded with zeros to produce a legitimate wavelet transform and this is inverted back to obtain the signal estimate.

The threshold δ is calculated using the signal obtained from the high pass filter output (detailed coefficients) according to equation 7:

$$\delta = s\sqrt{2\log(n)} \quad (7)$$

Where 'n' is the size of the data used to calculate the threshold and 's' is the estimation of the noise done by using median absolute deviation[23].

Usually thresholding is applied on the detailed coefficients and the approximate coefficients (the low pass filter output) are left untouched. Mathematically, for the detailed coefficient d_{ij} , the thresholding is carried out as follows:

$$\tilde{d}_{ij} = \begin{cases} \text{sign}(d_{ij}).(d_{ij}-\delta) & \text{if} \quad d_{ij} > \delta \\ 0 & \text{if} \quad d_{ij} > \delta \end{cases} \quad (8)$$

Where sign(x) is +1 if x is positive and -1 if x is negative.

The technique of soft thresholding is also called wavelet shrinkage because all the wavelet coefficients are reduced. Shrinkage of the wavelet coefficients is more helpful in reducing the noise from the signal as compared to the hard thresholding method[4]. The extent of denoising depends upon the level of decomposition. For higher level of decomposition, denoising can be applied to all the detailed coefficients. It is possible that some of the signal information may also be lost during the denoising process and the loss increases with the increase in the level of decomposition. The mother wavelet chosen for denoising was Daubechies 4. The thresholding was applied to the detailed coefficients only. The signal after denoising is smoother which also causes the removal of some of the signal components. This may cause reduction in the recognition performance at higher signal to noise ratios for the phonemes having high frequency components (e.g. fricatives).

4. Experimental Results

4.1. Phonetic Recognition System

4.1.1. Results with Clean Speech

The phone recognizer was trained and tested with TIMIT database[20]. All trainings were carried out using 100 phonemes of clean data. A set of 50 phonemes was used for testing. In case of tests with noisy speech, those clean test speech files were contaminated by an additive noise extracted from Noisex-92 database[3]. The temporal average is 80 ms. MLP with only one hidden layer and N output units was used to estimate the *a posteriori* probabilities of the classes, given the acoustic input. Each output unit of the MLP was associated to each phone (as emission probabilities of the states of the models were tied, only one output unit was needed for each model). The acoustic input to the MLP was formed by the Mel Frequency Cepstral Coeficients (MFCC) feature vectors. The number of input neurons is therefore the product of number of MFCCs and the number of acoustic vectors. Different sizes of hidden layer neurons are tested. The training procedure of the MLP was performed using the backpropagation algorithm with a sigmoidal activated function. The criterion function was the mean squared error. Our first experiment is intended to build a connectionist and hybrid recognition systems capable to recognize 4, 5, 6, 7 and 8 clean phonemes using two kinds of acoustic features (MFCC and PLP). Tables 1 and 2 summarize the obtained results in term of recognition rate (RR) respectively achieved with the MFCC and PLP features.

Table 1. Performance of Phone recognizer with MFCC Features

Number of phonemes	**4** { sh/ iy/ ae /jh}	**5** {sh/i y/ae /jh/ d}	**6** {sh/iy/ae/ jh/d/}	**7** {sh/iy /ae/ jh/d/k / r}	**8** {sh/iy/a e/jh/ d/ k/ r/ s}
MLP/HMM **RR(%)**	92	90	87	84.29	84
MLP **RR(%)**	91	87.4	85.33	80.29	79.5
Relative Improvement **RI(%)**	**1**	**2.9**	**1.9**	**4.8**	**5.4**

When comparing the performance of the two phone recognizers for the two kinds of features, we noticed that the hybrid (HMM/MLP) system outperform the connectionist (MLP) system in term of recognition rate (RR). However, an average relative improvement (RI) of 3.2% is obtained for the MFCC features and of 3.4% for the PLP features. Hence, the features obtained with PLP technique gave slightly better performance than those of MFCC.

Table 2. Performance of Phone recognizer with PLP Features

Number of phonemes	**4** { sh/iy/ ae /jh}	**5** {sh/iy /ae /jh/ d}	**6** {sh/iy/ae/jh/d/}	**7** {sh/iy/ ae/ jh/d/k/ r}	**8** {sh/iy/ae /jh/ d/ k/ r/ s}
MLP/HMM **RR(%)**	92.5	91.2	88.67	87.67	87.15
MLP **RR(%)**	90	88.8	87.33	82.33	83.25
Relative Improvement **RI(%)**	**2.7**	**2.7**	**1.5**	**5.9**	**4.3**

4.1.2. Results with Noisy Speech

Four types of noises, extracted from NOISEX-92 database, have been added to the clean signal at different signal to noise ratio (SNR) levels:

Volvo noise (the noise of car running at 120 Km/h)

Babble noise (the noise of 100 persons)

Helicopter noise

Pink noise

The performance of the phonetic recognition system, in term of recognition rate, for PLP features is depicted in figure 8.

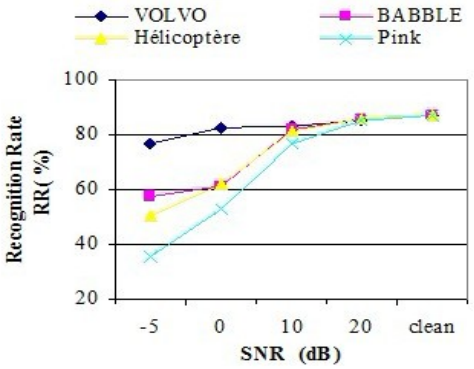

Figure 8. Performance of Hybrid Phonetic Recognition System in an Additive Noisy Environment with PLP Features

From the obtained results, we noticed the degradation of the performance of the phonetic system caused by the mismatch between training and testing conditions especially at low SNR levels. This degradation was more important when pink and the helicopter noises were considered. This is mainly due to the non stationary characteristics of such kind of noises[16].

4.2. Isolated Word Recognition System

4.2.1. Results with Clean Speech

The TIMIT database is phonetically transcribed using a set of 61 phones. We perform phonetic recognition on this database over a set of 39 classes that are commonly used[13]. Thus, each word in the English vocabulary could be composed by sequence of concatenating phones among the 39 phonemes. Therefore, the conceived isolated word recognition system should have a fixed input and 39 outputs which correspond to the number of output classes in the output layer of the neural network.

The vocabulary set used is composed of the 10 following words: “all”, “ask”, “carry ”, “greasy”, “had”, “like”, “rag”, “she”, “that” and “wash” . The phonetic transcription of each of these words is detailed in table 3.

Table 3. Phoneme’s Transcription of each Recognized Word Set

Word	ALL	ASK	CARRY	GREASY	HAD	LIKE	RAG	SHE	THAT	WASH
Phone set	**aa l**	**aa s k**	**k ae r iy**	**g r iy s iy**	**h ae d**	**l ay k**	**r ae g**	**sh iy**	**dh ae t**	**w a w s h**

In order to take into account the coarticulation effects-between phonemes in a word, besides MLP network, we tested Elman recurrent neural network. The Elman network is a simple recurrent network with feedback from each hidden node to all hidden nodes. The advantage of lman networks over fully recurrent networks is that back propagation is used to train the network while this is not possible with other recurrent networks where the training algorithms are more complex and therefore slower. Besides that, RNN may be used as an alternative posterior probabilities estimator.

The obtained results by the two hybrid systems (HMM/MLP and HMM/ RNN) for MFCC features are gathered in table 4.

Table 4. Recognition of 10 Isolated Words using HMM/ MLP and Elman RNN/HMM Hybrid systems with MFCC Features

Technique	*MLP/HMM*	*RNN/HMM*
Number of input neurons	210	210
Number of output neurons	39	39
Number of input neurons	78	78
Number of hidden neurons	544	1047
Recognition rate (%)	79.5	85.2

From the obtained results, it can be noticed that the performance of RNN/HMM system surpasses that of MLP/HMM. In fact, a relative improvement of 6.7% is obtained.

4.2.2. Results with Noisy Speech: Denoising with Wavelet Preprocessing Stage

The pre-processing is based on the denoising using dicret wavelet transform and is carried out before the feature extraction phase. The recognition performance achieved by soft thresholding is evaluated and compared with a system without the proposed pre-processing. The features are based on the commonly used MFCC acoustic features. HMM based recogniser is implemented for the isolated word recognition task. Results, at SNR=-5dB, are gathered in table 5.

Table 5. Performance of HMM / MLP and HMM / RNN Hybrid System for Isolated Word Recognition task at SNR=-5dB

Noises / features	VOLVO	BABBLE	HELI-COPTER	PINK
Clean environment	84	84	84	84
Noisy environment	52.25	56.5	37,75	25.25
Denoising with wavelet	83	79.75	80	80
Relative Improvement (%)	37	29.2	52.8	68.5

As can be observed from figure 9, a significant improvement is obtained with the discret wavelet preprocessing stage. Although the contaminating additive noise level is very important (SNR =-5dB), a substantial average relative improvement rate in RR of 47% is obtained.

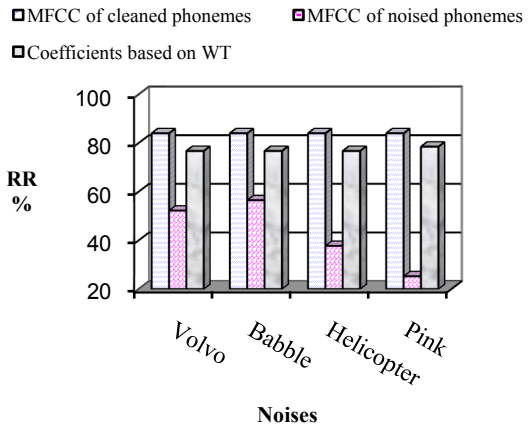

Figure 9. Recognition Rate (RR) in % at SNR=-5dB

5. Conclusions

In this research, we described three acoustical modeling approaches, HMM, ANN and hybrid HMM/ANN, used in state of the art speech recognition systems. Several experiments have been carried out in order to show the effectiveness of the hybrid approach especially when compared to the connectionist one. The focus of the first experiment was to study a hybrid phone recognition system where the connectionist architecture was based on MLP with only one hidden layer. Results showed the outperformance of the hybrid system over the MLP connectionist system. However, an average relative improvement in recognition rate of 3.4% was obtained using PLP acoustic features. The second experiment was intended to recognize 10 isolated words from the TIMIT database using two kinds of connectionist models: MLP and Elman RNN in the hybrid HMM/ANN recognition system. Results showed that Elman RNN enhanced the recognition rate when compared with MLP. In fact, a relative improvement in term of recognition rate of 6.7% was obtained. Finally, we investigated the robustness of the conceived hybrid systems when tested data were contaminated by various types of additive noise at different SNR values. So, we developed a pre-processing denoising stage based on wavelet transform. Results showed an important improvement in term of RR given by such technique. However, with additive noise level SNR =-5dB, a substantial average relative improvement of 47% was obtained.

REFERENCES

[1] A. Boubaker, M. Frikha, K. Ouni, A. Ben Hamida, 2006, "Une approche hybride neuro-markovienne pour la reconnaissance phonétique dans un milieu bruité, " 4th Int. Conf. JTEA 06, Hammamet, Tunisia.

[2] A. I. G-Moral, U. S-Urena, C. P-Moreno and F. D-Maria, 2011, "Data balancing for efficient training of hybrid ANN/HMM automatic speech recognition, " IEEE trans. on audio, speech and lang. proc., Vol 19, No. 3, 468-481.

[3] A. P. Varga et al., "The NOISEX-92 – Study on the effect of additive noise on an automatic speech recognition," Tech. Rep., DRA Speech Research Unit, 1992.

[4] D. L. Donoho, I. M. Johnston, 1995, "De-noising by soft-thresholding," IEEE trans. Information theory, Vol. 41, No.3, 613-627.

[5] E. Trentin, M. Gori, 2003, "Robust combination of neural networks and Hidden Markov Models for speech recognition," IEEE trans. on Neural net., Vol 14 , No. 6, 1519-1531.

[6] G. D. Formey, 1973, "The Viterbi Algorithm,"Proc. IEEE, (61), 7-13.

[7] H. Yi, P. C. Loizou, 2004, "Speech enhancement based on wavelet thresholding the multitaper spectrum," IEEE Trans. Speech and Audio Processing, (12), No. 1, 59-67.

[8] H. Bourlard, C. J. Wellekens, 1990, "Links Between Markov Models and Multilayer Perceptrons," IEEE Trans. on Pattern Analysis and Machine Intelligence, No. 12, 1167-1178.

[9] H. Bourlard, N. Morgan , 1993, "Continuous Speech Recognition by Connectionist Statistical Methods," IEEE Trans. on Neural Networks, (4), No. 6, 893-909.

[10] H. Ketadbar and H. Bourlard, 2010, "Enhanced phone post-

oriors for improving speech recognition," IEEE Trans. on audio, speech and lang. process. , Vol. 18, No. 6, 1094-1106.

[11] H. Hermansky, 1990, "Perceptual linear predictive_PLP. analysis of speech," J. Acoust. Soc. Amer., *(87)*, No. 4, 1738–1752.

[12] J. A. Bilmes, 1998, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixtures and hidden Markov models, " Tech. report 97-21, univ. of Berkely, USA.

[13] K. LEE and H. Hon, 1989, "Speaker-Independent Phone Recognition Using Hidden Markov Models," IEE*E* Trans. on ASSP, (31), No. 11.

[14] L.R. Rabiner, 1989, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,'' Proc. IEEE, (77), No. 2, 257–285.

[15] L. R. Rabiner and B. Juang, Fundamentals of Speech Recognition," Prent. Hall, Engl. Cliffs, New Jersey, USA, 1993.

[16] M. Frikha, "Approche Markovienne pour une reconnaissance robuste de mots isolés dans un environnement acoustique variable, '' PhD thesis, National School of Engineering of Sfax, Feb. 2007, Tunisia.

[17] M. Frikha, A. Ben Hamida and M. Lahyani, 2011, "Hidden Markov Models (HMMs) isolated word recognizer with the optimization of acoustical analysis and modeling techniques,'' Int. Jou. of Physical Sciences,Vol. 6(22), 5064-5074

[18] R. P. Lippmann, 1987, "An introduction to computing with neural nets," IEEE ASSP Magazine, 4-22.

[19] S.B. Davis and P. Mermelstein, 1980, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust. Speech Signal Process., (28), No. 4, 357– 366.

[20] S. Masmoudi, M. Frikha, A. Ben Hamida and M. Chtourou, 2010, "Efficient MLP constructive algorithm using neuron recruiting approach for isolated word recognition, " Int. Jour. of Speech Technol., Vol. 14, No 1,1-10.

[21] S. Renals, N. Morgan, H. Bourlard. M. Cohen, H. Franco, 1994, "Connectionist Probability estimators in HMM Speech Recognition ," IEEE Trans. on Speech and Audio Process., (2), No. 1, Part II, 161-174.

[22] O. Farooq, S. Datta, 2001, "Robust features for speech recognition based on admissible wavelet packet," Electronics letters, (37), No. 5, 1554-1556.

[23] O. Farooq, S. Datta, 2003, "Wavelet-based denoising for robust feature extraction for speech recognition, " IEEE Electronics letters, (39), No. 1, 163-165.

[24] P. Pujol, S. Pol, C. Nadeu, A. Hagen, H. Bourlard, 2005, "Comparison and Combination of Features in a Hybrid HMM/MLP and a HMM/GMM Speech Recognition System," IEEE Trans. on Speech and Audio Processing, (13), No. 1, 14-22.

[25] The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) Training and Test Data and Speech Header Software NIST Speech Disc CD1-1.1 October 1990.

[26] Z. Valsan, I. Gavat, B. Sabac, O. Cula, 2002, "Statistical and Hybrid Methods for Speech Recognition in Romanian," Int. Journal of Speech Technology, No. 5, 259-268.