Hybrid SVM/HMM Model for the Arab Phonemes Recognition

Elyes Zarrouk and Yassine Benayed
Multimedia Information System and Advanced Computing Laboratory, Sfax University, Tunisia

Abstract: *Hidden Markov Models (HMM) are currently widely used in Automatic Speech Recognition (ASR) as being the most effective models. Yet, they sometimes pose some problems of discrimination. The hybridization of Artificial Neural Networks (ANN) in particular Multi Layer Perceptions (MLP) with HMM is a promising technique to overcome these limitations. In order to, ameliorate results of recognition system, we use Support Vector Machines (SVM) witch characterized by a high predictive power and discrimination. The incorporation of SVM with HMM brings into existence of the new system of ASR. So, by using 2800 occurrences of Arabic phonemes, this work arises a comparative study of our acknowledgment system of it as the following: The use of especially the HMM standards lead to a recognition rate of 66.98%. Also, with the hybrid system MLP/HMM we succeed in achieving the value of 73.78%. Moreover, our proposed system SVM/HMM realizes the best performances, whereby, we achieve 75.8% as a recognition frequency.*

Keywords: *ASR, Hybrid System, HMM, MLP, SVM.*

Received July 24, 2012; accepted September 27, 2012; published online October 29, 2015

1. Introduction

Speech is neither over 10 milliseconds stationary intervals, nor a sequence of segments. Indeed, speech signal has a great amount of variability and articulator moves asynchronously. That's why the Automatic Speech Recognition (ASR) poses several problems and difficulties to researchers since 50s.

After forty years of the launch of the first draft of the speech recognition and despite the technological and scientific progress, we are still far from achieving communications' systems between humans and machines by speech. This fact illustrates the central interest of our work which involves the automatic recognition of Arabic speech.

Hidden Markov Models (HMM), introduced in the late 60s and early 70s, became the perfect solution to the problems of ASR. Indeed, these models are rich in mathematical structures and therefore can be used in a wide range of applications. Despite the enormous progress made by the HMM, they suffer from their lack of discrimination capability, specifically the learning phase of the HMM which requires a large amount of data to end to approach the conditional probabilities [24].

Despite the progress that these modals try to achieve, they still suffering from an absence of discriminatory power. Actually, to come near to this ability demands a very great number of examples to implement learning this system. Furthermore, the start using of these modals requires strict assumption with the cost in competition, time and memory [11].

Several approaches were proposed to integrate the Artificial Neural Networks (ANN) in particular Multi Layer Perceptions (MLP) which own high generalization ability from incomplete data when the volume of data is limited with HMM, to estimate HMM posterior probabilities. The integration of the MLP which has got a power class' discrimination with HMM is useful for identifying a temporal sequence and bringing the benefits presented on [10, 12, 22].

Our approach is based on Support Vector Machines (SVM) which have proved that they can solve multiple complex classification problems in many areas and have some strict properties in terms of discrimination and prediction. Briefly, the formalism of SVM embodies the principle of structural risk minimization which has shown that it is better than the empirical risk minimization, traditionally used by the ANN. The structural risk minimization decreases the upper bound of expected risk, which is opposed to empirical risk minimization which tends to the error on the training set. This is the reason behind the SVM's have large capacity of generalization which will be our focus of static learning [9]. In order to prove the effectiveness and efficiency of the hybrid system proposed below, we present a comparative study between recognition systems commonly used namely HMMs standards, the hybrid model MLP/HMM and the hybrid model proposed SVM/HMM.

In this paper we present how the system SVM/HMM works. In the beginning, we present a part of state of the art which contains a brief overview of the related work carried out in ASR of Arab phonemes then on the second part and third respectively, we present the operating procedure of both HMM and MLP/HMM. To describe how our system works SVM/HMM, we first describe the basic concepts of using SVM in section 5, then, in section 6, we describe the procedure of integration of SVM for estimating a posteriori probabilities with HMM. In section 7, we present the experimental results and interpretation is made in section 8. Finally, we end by presenting the conclusion of this work and a note of future works.

2. Related Works

In this part, we present a brief overview of the evolution of Arabic speech recognition systems. It provides a literature survey of Arabic speech recognition systems.

A number of researchers investigated the use of neural networks for Arabic phonemes and digits recognition [6, 17, 35]. For example, El-Ramly *et al.* [17] studied recognition of Arabic phonemes using an ANN. Alimi and Ben Jemaa [3] proposed the use of a fuzzy neural network for recognition of isolated words. Shoaib *et al.* [35] investigated a hybrid of neural networks and HMMs for Neural Networks/HMM for speech recognition or using the fuzzy rule [28]. Alotaibi [4] reported achieving high performance Arabic digits recognition using recurrent networks. Essa *et al.* [19] proposed different combined classifier architectures based on NN by varying the initial weights, architecture, type, and training data to recognize Arabic isolated words. Emami and Mangu [18] studied the use of Neural Network Language Models (NNLMs) for Arabic broadcast news and broadcast conversations speech recognition. El-Obaid *et al.* [16] applied an MLP network for Arabic phonemes recognition with KAPD data base. Ghassaq and Abduladhem [23] show that it is possible to use the hierarchical structure to recognize phonemes using Neural Fuzzy Petri net (NFPN).

Several other researchers have been developed for the automatic recognition but the results are far from being too close to people skills. The difference between the techniques used and the databases limit the possibility of learning to compare the results of this work. Among the best work done, we focus on HMMs, the hybrid systems combining neural networks and HMMs. As we will show in section 5 the efficiency and performance of SVM compared to neural networks in fact minimization of structural risk.

3. Hidden Markov Models

The HMM can be defined as a probabilistic automaton. It consists of a set of states linked by transitions with probabilities denoted a_{ij} forming the matrix of state transition i to j states also commented O_k probability for each state in which b_k every moment each state generates an observation. As it is described on [31], HMM can be interpreted as the set M=(N, A, B, π) with:

- N: The number of states of the model.
- $A=\{a_{ij}\}=P(q_{t=1}|q_{t-1=i})$ is the matrix of transition probabilities on the set of states of the model.
- $B=\{b_k(O_t)\}=P(O_t|q_{t=k})$ is the matrix of emission probabilities of the observations O_t for the state q_k.
- π is the initial distribution of states, $P(q_{i=0})$.

The learning phase is to estimate the model parameters M= (N, A, B, π) given sequence of observations O. Each phoneme is presented by an HMM. The representation of posterior probabilities is done by mixtures of Gaussians [15]. Several research studies have been developed by the HMM for automatic recognition of the Arabic language in different contexts [1, 2, 7, 34]. To evaluate the approach of hybridization and the hybrid system proposed, we will apply the HMM for the recognition of phonemes by the Arab HTK tool to perform the comparative study.

One of the major problems of modeling a Markovian model lies in the choice of the initial model, which is generally selected randomly with same probabilities, so it becomes quite clear that a better initialization of the Markovian model paves the way to get a better rate of recognition. It is necessary to evaluate the optimum number of Gaussians mixtures and the number of iterations for the learning algorithm since most HMMs is using Gaussian distributions'.

As a result of the difficulties found in the application of the HMM to speech recognition, mostly motivated by the temporal variability of the speech instances corresponding to the same class, a variety of different architectures and novel training algorithms that combined both HMM with ANNs were proposed in the late80's and 90's. For a comprehensive survey of these techniques [36]. In our study, we have focused on those that employ ANNs to estimate the HMM state posterior probabilities proposed by Bourlard and Morgan [13, 27].

4. Hybrid Model MLP/HMM

ANN structures have been used to classify inputs which are high dimensional and temporally correlated by the inclusion of neighboring frames as context. Hybrid connectionist HMM-ANN systems were developed as an alternative to the HMM-GMM, taking advantage of the MLP's model accuracy, context sensitivity and parsimonious use of parameters [20].

In recent years, neural networks have played an increasingly important role in the world of research, particularly since Rumelhart showed the different possibilities of neural networks with multi-layers [12]. They are particularly used as a statistical estimator. Networks commonly used in speech recognition are the MLP. One hidden layer is generally used [10, 30]. The addition of this hidden layer allows the network to model complex decision functions and nonlinear space between any input and output. Indeed, ANN is useful for the classification of static forms while being low in the treatment of temporality of the speech signal. Thus, it seems worthwhile to try to combine the respective capabilities of HMM and ANN to produce new hybrid models performing. However, this combination is not easy to achieve. Several studies have shown that a MLP trained in appropriate conditions is asymptotically equivalent to an estimator of posterior probability of belonging to a class [12]. Those probabilities will be the input of the Viterbi algorithm for decoding optimal sequence of observations.

4.1. Emission Probabilities with MLP

The problem of recognition of phonemes is that it has a suite of acoustic vectors $O=\{o_1, o_2, \ldots, o_N\}$ which must be associated phoneme following the most likely phonetic model for $Q=\{q_1, q_2, \ldots, q_L\}$ of L given word. L is the total number of phonemes. The goal is to find:

$$q_{optimal}=argmax_j \; P(q_j|O) \tag{1}$$

The posterior probability $P(qi|O)$ is not easily calculated so, the use of ANN that can solve problems of a great complexity and estimate posterior probabilities, it is necessary however to adapt the basic theory of HMM so that, it will be able to deal with these posterior probabilities. Learning the criterion for optimal learning and recognition HMM is based on the posterior probabilities of models M_λ given an acoustic sequence O and a set of parameters.

The hybrid system of speech recognition MLP/HMM is performed on the same principle as the HMM recognition. As shown in Figure 1, the neural network is merely used as an estimator of local probabilities for HMM.

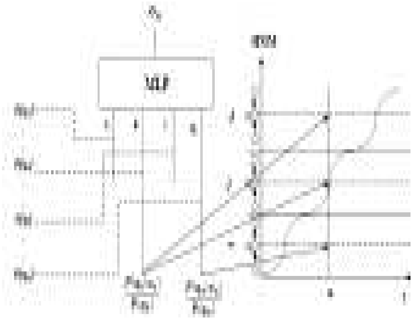

Figure 1. Architecture of MLP/HMM hybrid model.

The first part of the experiments is to determine the optimal configuration parameters of the model hybrid MLP/HMM. Indeed, to obtain the most representative possible measures, it is necessary to develop an hybrid system with a good performance. We should base in mind that the training time of a neural network can vary from several hours to several days depending on the size of the training vectors, the number of hidden layers and the number of neurons in each hidden layer. In the MLP and with any activation function nonlinear hidden layer with a large number of hidden units are sufficient to solve the more complex approximation [25]. But there is still no theory that can limit the number of neurons in the hidden layer. So, the network of the MLP used consists of 3 layers: an input layer, a hidden layer and an output layer.

The MFCC feature extraction produces vectors with 39 as coefficients for each phoneme, which will normally be the input values of the MLP reached. However, several studies [5, 30] have shown that the use of an acoustic context improves the high performance. In fact the assumption that at time n the dependence on O is restricted to a certain acoustic context centered on.

$$O_{(n-c)}^{(n+c)}=\{o_{n-c},\ldots,o_n,\ldots,o_{n+c}\} \tag{2}$$

The approach used is inspired from this design, so, in order to find best number of acoustic vectors to introduce the neural network, we resorted to test the ASR system with multiple contexts.

The recognition phase is to apply the decoding algorithm, the Viterbi HMM in order to classify the vector suitable candidate for the appropriate class. Given a sequence of observations O, the problem is to find the sequence representing the observations. The Viterbi algorithm is the best solution to the problem of estimating a posteriori probability $P(Q|O)$. The sequence of states most likely at time t depends only on t and the sequence most likely to t-1. This algorithm finds the optimal solution for each input element. So, the class is designated with the highest score.

5. Supports Vectors Machines

In the last 15 years or so, a novel breakthrough for ANN has been achieved in the field of pattern recognition and classification within the framework of kernel-based machine learning. They have gained wide popularity owing to the theoretical guarantees regarding performance and low computational complexity in nonlinear algorithms. Pioneered by Vapnik's [37] SVM for classification and regression, kernel-based methods, are nonlinear algorithms that can be adapted to an extensive class of nonlinearities. The main problem of the static learning theory is the generalization ability, in other words when does a low training error cause a low real error?

The solution given by the learning theory of Vapnik [37] is the theory of SVM. Similarly to neural networks, SVM proved their efficiency in several applications of prediction and classification that motivate researchers to better adapt this technology to optimize current systems especially systems of ASR [9]. The support vectors machines were introduced in the late 70s [38]. According to Vapnik, the general model of learning consists of the following three central components which are a generator of input vectors, a supervisor and a learning machine.

The problem of static learning from labeled examples, called supervised learning is to seek the function f which approaches the best answers supervisor in all functions $F=\{f(O, \alpha), \alpha\in A\}$where A is the set of parameters of f and $O=\{o_1, o_2, \ldots, o_m\}$ contains all acoustic input vectors, achievable by the chosen model M. We only have knowledge of the whole learning $D=\{(o_1, y_1), (o_2, y_2), (o_m, y_m)\}$ consisting of m independent input vectors, identically distributed according to $P(O, Y)$. $P(O_n, Y_n)$ is the distribution probability of the input vector O_n and desired output y_n known from the learning model.

The classification of data depends on the nature of data separation. There are cases of both linearly separable data and nonlinearly ones. With SVM a discriminative hyper lane with maximal border is searched when classes are separated linearly. With a constant intra class's variation classification, confidence grows with increasing interclass distance. The former are the simplest SVM because they can easily find a linear separation.

5.1. Linear SVM

Consider the problem of separating a set S of m vectors linearly separable $S=\{(o_1, y_1), (o_2, y_2), (o_m, y_m)\}$, where O_i is a feature vector ™*IRn* and y_i ™{-1, 1} a class label. We apply the transformation φ to obtain the set S of feature space $SFS=(\varphi(o_i), \varphi(y_i))$, $i=1, \ldots, m$ in the feature space [8, 32].

Each hyper plane H in the FS should satisfy the following conditions:

$$(w \, . \, \phi(o_i))+b >= +1 \;\; if \;\; y_i=1 \tag{3}$$

$$(w \, . \, \phi(o_i))+b <= +1 \;\; if \;\; y_i=-1 \tag{4}$$

5.2. Non-Linearly SVM

In this case, the set of the training vectors of both classes are non-linearly separable. To solve this problem, Vapnik [37] introduce non-negative variables, ξi>= 0, which measure the miss-classification errors. The optimization problem is now treated as a classification error minimization one [8]. The separating hyper plane must satisfy the following inequalities:

$$(w \, . \, \phi(o_i))+b >= +1 \; \xi_i \;\; if \;\; y_i=1 \tag{5}$$

$$(w \, . \, \phi(o_i))+b <= +1- \xi_i \;\; if \;\; y_i=-1 \tag{6}$$

There is therefore a possible shift of a nonlinear separation problem in the input space into a linear separation problem in a feature space of higher dimension. The transformation of features O ™IR^n into higher-dimensional space IR^m is done by:

$$\phi(O): IR^n \rightarrow IR^m$$

In both cases of data, the classification function $class(O)$ is written as follows:

$$class(O) = sign\,[(\Sigma_{o_i \in SV} \; y_i \;\; \alpha_i \;\; \phi(o_i \;) \; \phi(o_j)) + b^0] \tag{7}$$

To solve the problem of cases of non-linearly separable classes, the idea of SVM is to change the data space. Thus, the principle of SVM is to project input vectors into a feature space of a larger dimension so that, the optimal hyper plane constructed on this space is general, regardless of the size of the latter [26].

For the nonlinear SVM, we are in front of very high dimension of the feature space IR^m. So, $\varphi(o_i)$ $\varphi(o_i)$ must not be calculated explicitly, but can rather be expressed with reduced complexity with kernel functions.

$$K(o_i, \; o_j)) = \phi(o_i) \; \phi(o_j) \tag{8}$$

It is actually useless to know how the new feature space IR^m looks like [32]. All that we need to specify is kernel function as a measure of similarity [39]. The kernel is related to the transform $\varphi(o_i)$ by Equation [9]. The value of the kernel function is twofold: The calculation is done in the original space; this is much less expensive than a scalar product in large size. The transformation φ need not be known explicitly, only the kernel function involved in the calculations. This may lead to complex transformations and even space of infinite dimension redescription.

Among the most common kernel functions used in SVM. We quote the frequently kernel functions used in many applications:

- Polynomial-Kernel:

$$K(o_i, o_j) = [(o_i * o_j) + 1]^d \tag{9}$$

- Linear Kernel:

$$K(o_i, o_j) = o_i * o_j \tag{10}$$

- Sigmoid-Kernel:

$$K(o_i, o_j) = tanh(\beta_1 o_i * o_j + \beta_2) \tag{11}$$

- Radial Basis Function Kernel:

$$K(o_i, o_j) = exp(-\gamma \;\; |o_i - o_j|^2) \tag{12}$$

Where d, β_1, β_2 and γ are parameters that will be determinate empirically.

6. Hybrid Model SVM/HMM

To overcome the problems of discrimination of HMM, several researches are proposing to integrate the ANN, especially MLP, with HMM to estimate posterior probabilities. Our approach to combine SVM with HMM is based on those researches and applications.

The results obtained by the hybrid model MLP/HMM are challenging us to further deepen on the issue of the hybridization. For that we choose to replace the MLP by SVM.

Osowski *et al.* [29] SVM are better than MLP. The capacity of prediction and minimization of error of SVM prove our choice for the estimation of emission probabilities of the states of HMM for a sequence of observations. SVM take the task of estimating probabilities of issuing statements of HMM which will subsequently reformulated likelihoods to generate the optimal sequence by the decoding algorithm used by HMM.

6.1. Posterior Probabilities with SVM

SVM differ radically from MLP in that SVM training always finds a global minimum. The main difference between MLP and SVM is the principle of risk minimization. In case of SVM, structural risk minimization principle is applied by minimizing an upper bound on the expected risk whereas in MLP, traditional empirical risk minimization is used minimizing the error on the training data. The difference in risk minimization is to improve the generalization performance of SVM compared to MLP [33].

Here, we have used SVM to estimate posterior probabilities in the training phase and the recognition phase. First, we train one SVM for every sub-task signal which means one versus all. Every phoneme is a separate class. The function $f(O_i)$ that describes the separation plane measures the distance of the element O_i to the margin. The inclusion of the element O_i on one of the classes depends of sign $f(O_i)$. Also, the distance is far from the margin it has a higher probability of belonging to the class [14].

After choosing and applying the kernel function the conditional probability $P(o_i|class_j)$ is generating when a general model is summarized by minimizing the number of support vectors and supports the maximum data [27]. We need to calculate the likelihoods that the input vector O_i is given the *classj* of the appropriate phoneme j $P(class_j|o_i)$. We apply the Bayes rule to obtain those HMM emission probabilities:

$$P(class_j|o_i)=\frac{P(o_i|class_j)\,P(class_j)}{P(o_i)} \tag{13}$$

Where $P(class_j|o_i)$: Is the likelihood of the input vector O_i is given the class of the phoneme j, $P(class_j)$: Is the prior probability of the phoneme j, and $P(o_i)$: Is the priori probability of acoustic vector O_i.

6.2. Classification

For each phoneme we attribute a HMM with three state. We consider that all the states are combined on one HMM because the first and the last state for each HMM don't have any transition to another state. Given a sequence of observations $O=\{o_1, o_2, \ldots, o_N\}$ and a HMM M with N states (Number of phonemes), we wish to find the maximum probability state path $Q=\{q_1, q_2, \ldots, q_L\}$. This can be done recursively using the Viterbi algorithm.

Let $\delta_j(t)$ be the p of the most probable path ending in state j at time t:

$$\delta_j(t)=\underbrace{max}_{q_1,q_2,\ldots,q_{t-1}}\; P(q_1,q_2,\ldots,q_{t-1},q_{t=j},o_1,o_2,\ldots,o_t|M) \tag{14}$$

So, $\delta_j(t)=P(q_i|o_t)$ which is the probability estimated by the SVM kernel function from the observation o_t. We have to determinate finally:

$$q_{optimal}=\underbrace{max}_{1\le j\le N}[\delta_j(L)] \tag{15}$$

At the end we choose the highest probability endpoint, and then we backtrack from there to find the highest probability path [20]. We obtain a sequence of states that represents the observations' sequence O. Every state is according to a specific phoneme. This state is the only hidden state for the HMM of this phoneme. Thus, every phoneme is representing by a three state HMM with start state, hidden state and end state. Figure 2 shows the general architecture of the hybrid model SVM/HMM for the recognition of Arabic phonemes.

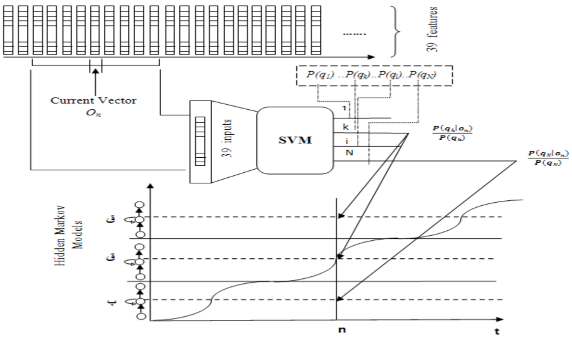
39 features
Current Vector
O_n
$P(q_1)$ ..$P(q_k)$..$P(q_i)$..$P(q_N)$
39 inputs
SVM
1
k
i
N
$\frac{P(q_k|o_n)}{P(q_k)}$
$\frac{P(q_N|o_n)}{P(q_N)}$
Hidden Markov Models
ق
ق
ب
n
t

Figure 2. Architecture of SVM/HMM hybrid model.

7. Experimental Results

To evaluate our proposed hybrid SVM/HMM system, we performed three recognition system based on: The standards HMM, a hybrid MLP/HMM and a hybrid SVM/HMM. Then, we make a comparative etude using the recognition rate of the Arabic phonemes between those three systems [40].

7.1. Database and Parametrisation

All experiments are carried out with database of "Aljazeera" for ASR especially Arabic phoneme. Full Corpus was extracted from Aljazeera emissions (spontaneous speech). In fact the recordings were performed in a noisy environment: Microphones, background noise, hesitations, breathing. This corpus was segmented in phonemes using PRAAT. We choose different utterances of phonemes from different speakers. For each phoneme, we choose 100 utterances. Finally, we have 2800 ones, (100 occurrences of each one of the 28 Arabic phonemes). This corpus was divided into two parts, 80% for training and 20% for recognition. Then, the database used for training contains 2240 utterances and for the test we use a set of 560 utterances.

From a speech signal, the first treatment is to extract the characteristic parameters. Among the factors most commonly used and which best represent the speech signal in speech recognition, we find the cepstrum or cepstrum coefficients. For spectral analysis, we used the mel frequency cepstral

Coefficients MFCC. In our work, we have chosen to use Mel scale. The number of filters used empirically:

Zwicker offers 24 filters [6, 21]. Before any calculation, it is necessary to perform some operations to prepare the speech signal. First, the signal is filtering. Second, it will be sampled at a regular frequency and the high frequency will be removed. Finally, the signal will be segmented into frames. Each frame is composed of a fixed number of speech samples. In our case, the duration of each frame is 25 ms during which the speech is stationary. The phase of filtering speech signal brings several problems and in order to avoid them, we use the weighting windows. Among the most common windows, we apply the Hamming window. In our experiments, we will apply 12 MFCC with first and second derivatives plus energy to obtain a vector of 39 coefficients. To extract those features, we have used HMM Toolkit (HTK 3.4).

7.1 The HMM Standard System

HTK is used here in the state of modeling the phoneme. For each Arab phoneme we assigned a left-right HMM with 5 states. The emission probabilities are modeled by mixtures of Gaussians. To obtain the best results of recognition, we have done some test to select the best parameters for the application of the HMM.

The number of Gaussians for a representation of data and the number of iterations of the training algorithm are both the primordial parameters which depend on the recognition. 64 was the number of Gaussians chosen, and 40 times is the maximum of iterating the learning algorithm.

The following table presents the results obtained for the recognition of Arab phonemes by HMMs.

7.2. The Hybrid MLP/HMM Based System

Like any neural architecture, the MLP has parameters depending on the nature of the application. You should know that the training time of neural network can vary from several hours to several days depending on the size of the training vectors, the number of hidden layers and number of neurons in each hidden layer. At first, we choose a one hidden layer having 420 neurons. The results depend on the input context [13]. To adopt the best parameters of the MLP on our database, we have done some tests. After the tests, 351 coefficients MFCC were obtained as the best size of the input vector, i.e. it combines 9 MFCC vectors. The back-propagation algorithm applied to the learning of the MLP was iterating 100000 times to achieve the good results of the Arab phonemes recognition.

The following table presents the results obtained for the recognition of Arab phonemes by MLP/HMM.

7.3. The Hybrid SVM/HMM System

The first step on application of SVM is the choice of the kernel function. According to the RBF kernel is the kernel the most effective and suitable for signal processing applications, especially for prediction problems. For this reason, we have used the RBF kernel 5. The selection of the suitable parameters is generally made by empirical tests. So, we should maximize the value of the Gamma function and SVM empirical parameter *C* associated with the application of any kernel SVM. After the tests on our database, we obtained γ=0.7 and *C*=10 are the best values of parameters related to the application RBF kernel on Arab phonemes recognition.

Table 1 shows the results obtained for the recognition of Arab phonemes by our proposed system SVM/HMM.

Table 1. Recognition rate of arabic phonemes with three systems of ASR (HMM, MLP/HMM and SVM/HMM).

	HMM	MLP/HMM	SVM/HMM
3a ع	63.79 %	73.65 %	**83.65** %
Ga غ	70.37 %	**72.96** %	70.63 %
5a خ	62.07 %	73.79 %	**76.79** %
Wa و	66.67 %	72.26 %	**73.26** %
Thaa ظ	72.22 %	**73.96** %	57.12 %
Tha ث	72.22 %	**74.81** %	**74.81** %
Taa ط	74.07 %	**76.81** %	**76.81** %
Ta ت	62.96 %	69.61 %	**76.61** %
Shaaش	61.11 %	72.21 %	**74.43** %
Saa ص	72.22 %	73 %	**75.18** %
Sa س	70.37 %	68.6 8 %	**80.68** %
Ra ر	64.81 %	73.81 %	**87.81** %
Ma م	64.81 %	68.96 %	**79.96** %
La ل	63.16 %	**73.7** %	70.48 %
Ja ج	**70.37** %	69.61 %	69.61 %
Fa ف	57.89 %	68.7 %	**72.33** %
Dha ذ	59.65 %	68.7 %	**68.7** %
Da د	59.26 %	**73.7** %	71.7 %
A أ	62.96 %	72.21 %	**75.21** %
7a ح	66.67 %	73.81 %	**81.44** %
9a ق	63.16 %	**73.16** %	71.16 %
Ha ه	56.14 %	**68.7** %	66.7 %
Ba ب	62.96 %	**73.29** %	**73.29** %
Dhaaض	68.42 %	73.7 %	60.69 %
Na ن	63.16 %	**68.88** %	56.88 %
Ka ك	61.4 %	68.7 %	**71.27** %
Za ز	64.91 %	73.91 %	**89.1** %
Ya ي	61.4 %	73.89 %	**87.61** %

7.4. Discussion

By evaluating the recognition rate of each phoneme separately, we can judge all systems together. Here, as the previous tables mentioned, the recognition rates of Arabic phonemes are reached by the HMM system, but, standards are the lowest as 23 phonemes have the lowest rate among 28 awards. Only the phoneme "Ja" has the highest recognition frequency of 74.04%, whereas, the phoneme "Ha" gets the lowest one by 56.14%.

As it is seen the hybrid systems behave well, although, the hybrid system MLP/HMM seems more efficient than the last one, it obtains the best recognition's rate for 10 of 28 phonemes. Thus, comparing to the recognition rate of the last system, we perform that there is a big difference between them i.e., 26 of 28 phonemes are bigger than those obtained by the HMM standards which prove the effectiveness of the hybridization of the MLP with HMM.

Finally, looking at the hybrid model presented in this paper SVM/HMM, the recognition rates are presented as the best results. 17 among 28 phonemes have the best rate of recognition compared to the hybrid system MLP/HMM, the multiclass SVM and that of the HMM standards. Moreover, the best recognition rate obtained for all experiments is obtained by this hybrid system for the phoneme “Za” (89.1%). Those results confirm the good choice of estimation posterior probabilities by the SVM. However, three phonemes have the lowest results.

We already presented the results of automatic recognition of Arabic phonemes by three ASR systems. It showed that hybrid models are the most effective in realizing the best results especially SVM/HMM model. And by testing the reliability of these systems, we find that the vocal sequences recognition consists of an arbitrary mixture of acoustic vectors of all Arab phonemes. Furthermore; data from experiments are randomly mixed from MFCC corpus of data already mentioned in paragraph VIIA. Table 2 shows the obtained the results.

Table 2. Recognition rate of all systems of ASR (HMM, MLP/HMM and SVM/HMM).

	HMM	MLP/HMM	SVM/HMM
Recognition Rate	66,98 %	73,78 %	**75,80 %**

We may deduce from the results shown above that hybrid systems are most successful in obtaining the best recognition rate even for a sequence of acoustic vectors extracted from all the phonemes and mixed randomly namely 73.78% for system MLP/HMM and 75.8% for the most efficient hybrid SVM/HMM. The good result obtained by the system multiclass SVM can be classified near to the rank of the system MLP/HMM.

8. Conclusions

In this paper, we have performed a comparison of the recognition rate of Arab phonemes between three ASR systems using consecutively: Standards HMM, MLP/HMM and SVM/HMM which is our proposed work.

We have overcome the limitations of HMM emission distribution probability using SVM as posterior probabilities estimator.

The Arab phonemes recognition results that obtained by SVM/HMM hybrid model are compared with results obtained by HMM standards; as well as with those of the MLP/HMM hybrid system. Whereby; it showed a good effectiveness and performance.

Despite the fact that the results of Arab phonemes recognition were good; however, it can never be considered as theories or basis accepted by any environment. Indeed, all this depends on the corpus and nature of the data as well as on the parameters adapted for the presented system and the quality of audio recordings. On the other hands ,we can regard these results as an initiation for further researches by which other experiments continue speaking Arabic with this hybrid system especially the SVM/HMM. The main objective of this research is to use the SVM/HMM hybrid system for the recognition of Arabic continue speech. Finally, we propose to integrate it on an interactive born.

References

[1] Al-Zabibi M., “An Acoustic-Phonetic Approach in Automatic Arabic Speech Recognition,” *PhD Thesis*, Loughborough University Institutional Repository, 1990.

[2] Alotaibi A., “Investigating Spoken Arabic Digits in Speech Recognition Setting,” *Information Sciences*, vol .173,no. 1-3, pp. 173-115, 2005.

[3] Alimi A. and Ben Jemaa M., “Beta Fuzzy Neural Network Application in Recognition of Spoken Isolated Arabic Words,” *Control and Intelligent Systems*, vol. 30, no. 2, pp. 47-51, 2002

[4] Alotaibi Y., “Spoken Arabic Digits Recognizer using Recurrent Neural Networks,” *in Proceedings of the 4th IEEE International Symposiumon Signal Processing and Information Technology*, pp. 195-199.*2004*

[5] Aradilla G., Bourlard H., Magimai M., “Using KL-based Acoustic Models in a Large Vocabulary Recognition Task,” available at: http://publications.idiap.ch/index.php/publications/show/6, last visited 2008.

[6] Bahi H. and Sellami M., “A Hybrid Approach for Arabic Speech Recognition,” *in Proceedings of ACS/IEEE International Conference on Computer Systems and Applications*, Tunis, pp. 14-18, *2003*

[7] Baloul S., “*Développement d’un Système Automatique De Synthèse De La Parole à Partir du Texte Arabe Standard Voyellé*,” *PhD Thesis*, University Of Maine, 2003.

[8] Ben Ayed Y. and Jamoussi S., “A New Beta Function Based Kernel for SVMs: Application to Keyword Spotting,” *Journal of Computer Science and Engineering*, vol. 9, no. 2, 2011.

[9] Ben Ayed Y., Fohr D., Haton J., and Chollet G., “Confidence Measures for Key Word Spotting Using Support Vectors Machines,’’ *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 588-591, 2003.

[10] Bernadis G. and Bourlard H., “Confidence Measures in Hybrid HMM/ANN Speech Recognition,” *in Proceedings of the 1st workshop on Text, speech, Dialogue*, 1998.

[11] Bilmes J., "Natural Statistical Models for Automatic Speech Recognition," available at: ftp://ftp.icsi.berkeley.edu/pub/speech/papers/thesis-bilmes99.pdf, last visited 1999.

[12] Boite J., Bourlard H., D'hoore B., Accaino S., and Vantieghem J., "Task Independent and Dependent Training: Performance Comparison of HMM and Hybrid HMM/MLP Approaches," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Adelaide, pp. 617-620, 1994.

[13] Bourlard H. and Morgan N., *Connectionist Speech Recognition: A Hybrid Approach*, Norwell, Kluwer Academic, 1994.

[14] Castellani A., Botturi D., Bicego M., Fiorini P., "Hybrid HMM/SVM: Model for the Analysis and Segmentation of Teleoperation Tasks," *in Proceedings of IEEE International Conference on Robotics and Automation New Orleans*, pp. 2918-2923, 2004.

[15] Connel S., "A Comparison of Hidden Markov Model Features for the Recognition of Cursive Handwriting," *MS Thesis*, Computer Science Department, Michigan State University, 1996.

[16] El-Obaid Manal., Amer Al-Nassiri., and Imen Abul Maaly "Arabic Phoneme Recognition Using Neural Networks," *in Proceedings of the 5th WSEAS International Conference on Signal Processing*, Istanbul, Turkey, pp. 99-104, 2006.

[17] El-Ramly S., Abdel-Kader N., and El-Adawi R., "Neural Networks Used for Speech Recognition," *in Proceedings of the 19th National Radioscience Conference*, pp. 200-207, 2002.

[18] Emami A. and Mangu L., "Empirical Study of Neural Network Language Models for Arabic Speech Recognition," *in Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding,* Kyoto, pp. 147-152, 2007.

[19] Essa E., Tolba A., Elmougy S., "A Comparison of Combined Classifier Architectures for Arabic Speech Recognition," *in Proceedings of International Conference on Computer Engineering and Systems*, Cairo, pp. 149-153, 2008.

[20] Faria A., "An Investigation of Tandem MLP Features for ASR," available at: http://www.icsi.berkeley.edu/pubs/techreports/faria_icsitr.pdf, last visited 2007.

[21] Garcia-Moral A., Solera-Urena R., C. Pelaez-Mor., and Diaz-de-Maria F., "Hybrid Models for Automatic Speech Recognition: A Comparison of Classical ANN and Kernel Based Methods," available at: http://www.eurasip.org/Proceedings/Ext/NOLISP2007/papers/p34.pdf, last visited 2007.

[22] Gemello R., Mana F., Scanzio S., Laface P., and De Mori R., "Adaptation of Hybrid ANN/HMM Models using Linear Hidden Transformations and Conservative Training," *in Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Toulouse, 2006.

[23] Ghassaq S. and Abduladhem A., "Arabic Phoneme Recognition using Hierarchical Neural Fuzzy Petri Net and LPC Feature Extraction," *Signal Processing: An International Journal*, vol. 3, no. 5, pp. 161-171, 2009.

[24] Gold B. and Morgan N., *Speech and Audio Signal Processing*: *Processing and Perception of Speech, and Music*. John Wiley and Sons Inc, 1999.

[25] Hornik K., "Some New Results on Neural Network Approximation," *Neural Networks*, vol. 6, no. 8, pp. 1069-1072, 1993.

[26] Märgner V., "Cours de Support Vector Machines," *Institute of Nachrichtentechnik (IfN) TU Braunschweig*, 2009.

[27] Morgan N. and Bourlard H., "Continuous Speech Recognition: An Introduction to the Hybrid Hmm/Connectionistapproach," available at: http://www1.icsi.berkeley.edu/ftp/pub/speech/papers/ieeespm95-hyb.pdf, last visited 1995.

[28] Muhammad M., "Recognition of Arabic Phonemes using Fuzzy Rule Base System," *in Proceedings of the 7th IEEE International Multi Topic Conference*, pp. 367-370, 2003.

[29] Osowski S., Siwek K., and Markiewicz T., "MLP and SVM-A Comparative Study," *in Proceedings of the 6th Nordic Signal Processing Symposium-NORSIG*, Espoo, Finland, pp. 37-40, 2004.

[30] Pujol P., Bourlard H., Pol S., Nadeu C., and Hagen A., "Comaparison and Combination of Features in a Hybrid HMM/MLP and a HMM/GMM Speech Recognition System," *IEEE Trans. on SAP EDICS: 1-RECO*, vol. 13, no. 1, pp. 14-22, 2003.

[31] Rabiner L. and Juang B., *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[32] Rafik D., Houcine B., and Amara K., "A Combination Approach of Gaussian Mixture Models and Support Vector Machines for Speaker Identification," *The International Arab Journal of Information Technology*, vol. 6, no. 5, pp. 490-497, 2009.

[33] Samanta B., Al-Balushi K., and Al-Araimi S., "Artificial Neural Networks and Support Vector Machines with Genetic Algorithm for Bearing Fault Detection," *Engineering Applications of Artificial Intelligence*, vol. 16, no. 7-8, pp. 657-665, 2003.

[34] Satori H., Hussein H., Harti M., and Chenfour N., "Investigation Arabic Speech Recognition Using CMU Sphinx System," *The International*

Arab Journal of Information Technology, vol. 6, no. 2, pp. 186-190, 2009.

[35] Shoaib M., Rasheed F., Akhtar J., Awais M., Masud S., Shamail S., "A Novel Approach to Increase the Robustness of Speaker Independent Arabic Speech Recognition," *in Proceedings of the 7th International Multi Topic Conference*, pp. 371-376, 2003.

[36] Trentin E. and Gori M., "A Survey of Hybrid ANN/HMM Models for Automatic Speech Recognition," *Neurocomputing*, vol. 37, no. 1-4, pp. 91-126, 2001.

[37] Vapnik V., *Estimation of Dependences Based an Empirical Data*, Springer Verlog, New York, 1979.

[38] Vapnik V., *The Nature of Statical Learning Theory*, Springer Verlag, New York, 1995.

[39] Von Luxburg U., Bousquet U., and Scholkopf O., "A Compression Approach to Support Vector Model Selection," *The Journal of Machine Learning Research*, vol. 5, pp. 293-323 2004.

[40] Zarrouk E. and Ben Ayed Y., "Automatic Speech Recognition with Hybrid Models," *in Proceedings of SPED Conference*, pp. 183-188, 2011.

Elyes Zarrouk received his MS degree in Computer Science at the Higher Institute of Computer of Monastir, Tunisia in 2007. He obtained his MS degree on New Information Technologies and Systems Dedicated from National School of Engineering in Sfax, Tunisia in 2010. Currently, he is a PhD student in MIRACL, Multimedia Information System and Advanced Computing Laboratory, university of Sfax, Tunisia, Focusing his research on speech recognition.

Yassine BenAyed Graduated in Electrical Engineering from National School of Engineering in Sfax, Tunisia in 1998. He obtained his PhD degree in Signal and Image from Telecom ParisTech in 2003. Curently, He is assistant professor in Electrical and Computer Engineering in the University of Sfax. He focuses his research on pattern recognition, artificial intelligence and speech recognition.