Contributions to HMM-based Speech Recognition Systems

Hazmoune Samira[1], Bougamouza Fateh[1], Mazouzi Smaine[1], Benmohammed Mohamed[2]
[1]Department of Computer Science
20 Août 1955 University, Skikda, Algeria
[2]Department of Computer Science,
Mentouri University, Constantine, Algeria
{ hazsamira, bougamfateh, mazouzi_smaine}@yahoo.fr, Algeria ben_moh123@yahoo.com

ABSTRACT: *In this paper, we propose a new approach based on multiple modeling by Hidden Markov Models (HMM) for isolated word recognition, which aims to maximize word recognition rate by combining several models coming from different start points. Our approach operates on 2 steps; first we create a large set of candidate markovian models for every vocabulary word by changing in their initial models in the Expectation-Maximization (EM) training algorithm, and then we select the best subset of models. The selection of appropriate models to be combined is one of the most difficult but important factors in our approach. For this purpose, we have experimented in our previous work [1] three methods: In the first one, we selected models whose give best individual recognition rates. Secondly, we selected random models. Thirdly, we used a Genetic Algorithm (GA) to select the optimal set of models by maximizing the recognition rate of the group and minimizing the number of selected models. In this paper we propose another new selection method, in which we select models whose maximize the difference between the average of likelihoods of the current class and the average of likelihoods of all others. The performance of the proposed approach will be evaluated by comparing its effectiveness against classical markovian approach.*

Keywords: Speech Recognition, HMM, EM, Initial Models, Multiple Modeling, Models Selection

Received: 7 December 2012, Revised 18 January 2013, Accepted 24 January 2013

1. Introduction

One most attractive features of the HMM framework is that its parameters can be learned automatically from the training data. The parameters are usually learned by the maximum likelihood (ML) criterion based on the Expectation-Maximization (EM) algorithm. However, the most serious problem that we face with this algorithm is the sensitivity to the initial parameters.

To overcome this shortage of the EM algorithm, several refinements had been recently proposed, in [2] authors present several ways to initialize and train HMMs for gesture recognition. These include using a single initial model for training, multiple random initial models to generate a single final model, and initial models directly computed from physical considerations. Each of the initial models is trained on multiple observation sequences using both Baum-Welch and the Viterbi Path. In [3] several candidate HMMs are created by applying EM on multiple initial models. A single HMM is chosen from the candidate HMMs which has highest value for likelihood function. Authors in [4] proposed a hybrid algorithm, Simulated Annealing Stochastic version of EM (SASEM), combining Simulated Annealing (SA) with EM that reformulates the HMM estimation process using a stochastic step

between the EM steps and the SA. In this field, we investigate some contributions to HMM-based automatic speech recognition systems that can be summarized in the following:

• The 1st contribution is the exploitation of the sensitivity of Hidden Markov models recognizers to the initial point in Expectation-Maximization training; in order to generate a set of different models with different accuracies for each class. So, in this new approach we used multiple modeling instead of the use of only one model per class. This approach aims to improve the accuracy and to come closer to the global optimum by using several models which are distributed in the space.

• The 2nd contribution is the proposition of two new model selection methods. In the first one, we use a Genetic Algorithm to select the subset of models to be combined by maximizing the accuracy of the ensemble and minimizing the number of selected models, in order to attain a compromise between performance and response time. In the second method, we select models which maximize the average of likelihoods of the current class and minimize the average of likelihoods of all the others. The goal of this method is to enhance the inter-classes separability.

The remainder of this paper is organized as follows. First, we give a short introduction to HMMs in Section II. The description of HMMs is relatively short because it is widely used and detailed descriptions can easily be obtained from [5]. Then we discuss the effect of initial model choice in HMM training in Section III. In Section IV, we describe the proposed approach. Experimental results are reported in Section V. Finally, the paper is summarized in Section VI.

2. Hidden Markov Models

HMMs are probabilistic models useful for modeling stochastic sequence with underlying finite state structure. Stochastic sequences in speech recognition are called observation $O = o_1 o_2 \ldots o_T$, where T is the length of the sequence [6]. An HMM with N states can be characterized by three sets of parameters $\pi = \{\pi\}$, $A = \{a_{ij}\}$ and $B = \{b_{jk}\}$, $1 \le i, j \le N$, $1 \le k \le M$, where π is the initial state probability distribution, A is the state transition probability distribution matrix and B is the observation symbol probability distribution matrix. Here M is the total number of distinct observation symbols per state. The elements of the matrices π, A and B always satisfy the following stochastic conditions:

$$\sum_{j=1}^{N} \pi_i = 1 \tag{1}$$

$$\sum_{j=1}^{N} a_{ij} = 1, \text{ where i} = 1, 2, \ldots N \tag{2}$$

$$\sum_{k=1}^{M} b_{jk} = 1, \text{ where j} = 1, 2, \ldots N \tag{3}$$

For the classical use of HMMs, three main issues need to be solved [5]:

• **Problem 1 (evaluation)**: Given the observation sequence $O = o_1 o_2 \ldots o_T$, and a model $\lambda = (A, B, \pi)$, how do we efficiently compute $P(O|\lambda)$, the probability of the observation sequence, given the model? This is usually done by the Forward-Backward algorithm.

• **Problem 2 (decoding)**: Given the observation sequence $O = o_1 o_2 \ldots o_T$, and the model λ, how do we choose a corresponding state sequence $Q = q_1 q_2 \ldots q_T$ which is optimal in some meaningful sense (i.e., best "*explains*" the observations)? This is efficiently determined by the Viterbi algorithm.

• **Problem 3 (training)**: How do we adjust the model parameters $\lambda = (A, B, \pi)$ to maximize $P(O|\lambda)$? This problem is commonly solved by Baum-Welch algorithm which is an application of the Expectation-Maximization algorithm (EM).

3. Effect of Initial Model Choice in the Expectation-Maximization Training

The power of HMMs is the automation of the training parameters $\lambda = (A, B, \pi)$, which is achieved using the algorithm of Baum-Welch or Viterbi. The idea of the application of these algorithms is to use the EM algorithm which is an iterative procedure to estimate gradually the model $\lambda^{\sim}$ according to the following steps:

Step 1: Propose an initial set of parameters λ_0,

Step 2: Calculate $\lambda^{\sim}$ from λ_0 using Baum-Welch (Forward-Backward) or Viterbi criterion,

Step 3: If $P(O|\lambda^{\sim}) \geq P(O|\lambda_0)$ was considerably improved by the updated model $\lambda^{\sim}$ with respect to λ_0 let $\lambda_0 \leftarrow \lambda^{\sim}$ and continue with step 2 Otherwise Stop!

Both Viterbi and Baum-Welch algorithms are very dependent on the initial models. They only guarantee convergence to a local maximum of likelihood [8]. Hence, the choice of initial parameters is an integral element of a development of a HMM system, and as such has been a subject of active research such as [2], [3], [4], [6] and [7].

4. Proposed Approach

Before describing the proposed approach, we will first present briefly the classical one based on single HMM modeling and applied in isolated word recognition. For each word of a *W* word vocabulary, a separate N-state HMM is designed. The speech signal of a given word is presented as a time sequence of feature vectors. For each vocabulary word, we have a training sequence consisting of a number of repetitions (utterances) by one or more talkers. The first task is to build individual word models. This task is commonly done by using the Baum-Welch algorithm to optimally estimate model parameters for each word model. Finally, once the set of *W* HMMs has been designed and optimized, for each unknown word which is to be recognized, the measurement of the observation sequence $O = o_1 o_2 \ldots o_T$, is performed. This latter is done via a feature analysis of the speech corresponding to the word, followed by calculation of model likelihoods for all possible models, $P(O|\lambda^w)$, $1 \leq w \leq W$; followed by selection of the word whose model likelihood is highest, i.e.,

$$w^* = argmax\,[P(O|\lambda^w)],\ 1 \leq w \leq W \qquad (4)$$

The probability computation is generally performed using the Viterbi algorithm (i.e., the maximum likelihood path is used), for more details refer to [5].

Figure 1 shows the principle of the classical markovian approach.

Unlike classical approach, our approach is based on multiple modeling; where several final models coming from different start points are generated and combined in order to increase the accuracy. The basic idea is that every time we change initial parameters a new HMM with a different accuracy is obtained. That's why the use of multiple HMMs leads to increase the accuracy of the group and to come closer to global maximum of likelihood. The figure 2 shows different steps of the proposed approach.

4.1 Feature Extraction

Feature extraction is one of the key modules for any recognition system. In our experiments, the Mel-Frequency Cepstrum Coefficients (MFCCs) are used as recognition features. Speech is segmented in frames of 256 samples, and the window analysis is shifted by 128 samples. Each frame is converted to 13 MFCCs plus a normalized energy parameter. The first and second derivatives (D's and DD's) of MFCCs and energy are estimated, resulting in 42 numbers, thus, for each frame the feature extraction module delivers 42 numbers regrouped in an acoustic vector. In the modeling stage, each speech signal is represented as a sequence of acoustic vectors $O = o_1 o_2 \ldots o_T$, where T (the number of vectors) depends in the temporal signal size. One should notice that the goal here is not speech compression but using features suitable for speech recognition.

4.2 Training

The training algorithm operates on two levels. Firstly, for each class, it performs initial model changing in order to generate a set of different final models for each word using Baum-Welch algorithm and then it chooses the best team of models.

4.2.1 Generation of a Set of Final Models for Each Word

In this step, several continuous left-to-right HMMs are created for each word. The number of states per HMM is equal to the number of phonemes in the word to be modeled. We have associated with each state i, a single Gaussian function in order to model the probability densities, where observation probabilities are completely defined by the mean vector μ and the covariance matrix σ. Let a particular observation sequence be described as $O = o_1 o_2 \ldots o_T$. The probability of a particular observation at a

particular time t for state i be described by: $b_i(o_t) = P(o_t|q_t=i)$,where q_t is a random variable with N possible values {1, 2, ...N}, representing the state at time t, here N is the number of states. The complete collection of parameters for all observation distributions is represented by $B=\{b_i(.)\}$. So in the case of a single Gaussian function, we can write:

$$b_i(o_t) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(o_t-\mu_i)^2}{2\sigma_i^2}} \quad (5)$$

The generation of the candidate HMMs can be summarized in the following pseudo code.

For each vocabulary word do
  Begin
    Number_of_candidate_HMMs := 0;
    While number_of_candidate_HMMs ≤ Max-iteration (*) do
      Begin
        1. Create random initial model $\lambda_0 = (A, B, \pi)$ in such a way that it satisfies the criteria shown in Equation1, Equation 2 and Equation 3, (**)
        2. Apply the EM algorithm (***) in order to generate a candidate model (HMM),
        3. Number_of_candidate_HMMs := Number_of_candidate_HMMs + 1;
      End While
End For

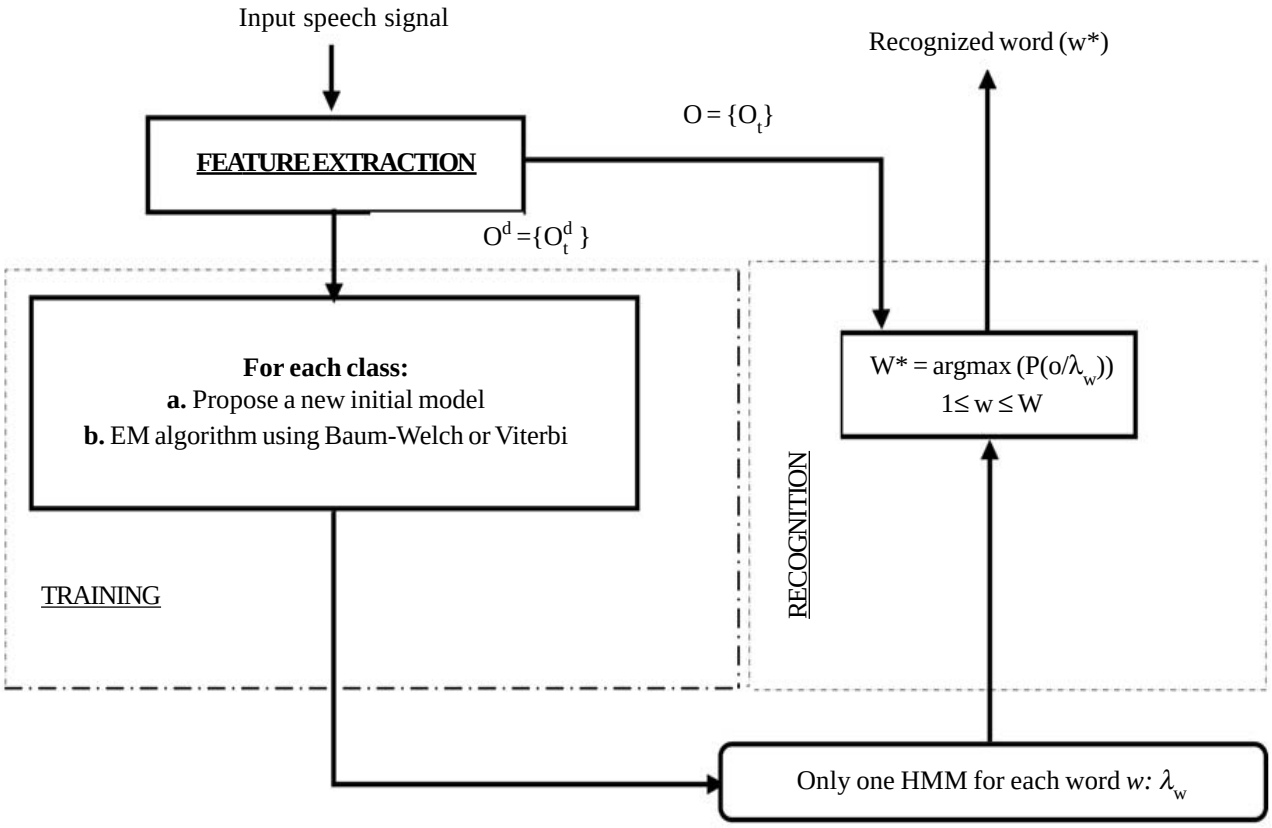

Figure 1. Block diagram of the classical markovian approach

(*) Max-iteration is fixed empirically. It corresponds to the number of candidate HMMs.

(**) We make initial models evenly distributed in the space. We expect that the group of final models should include several points near the global maximum. Here B is defined by μ_i and σ_i, where $1 \le i \le N$ (N is the number of states of the model).

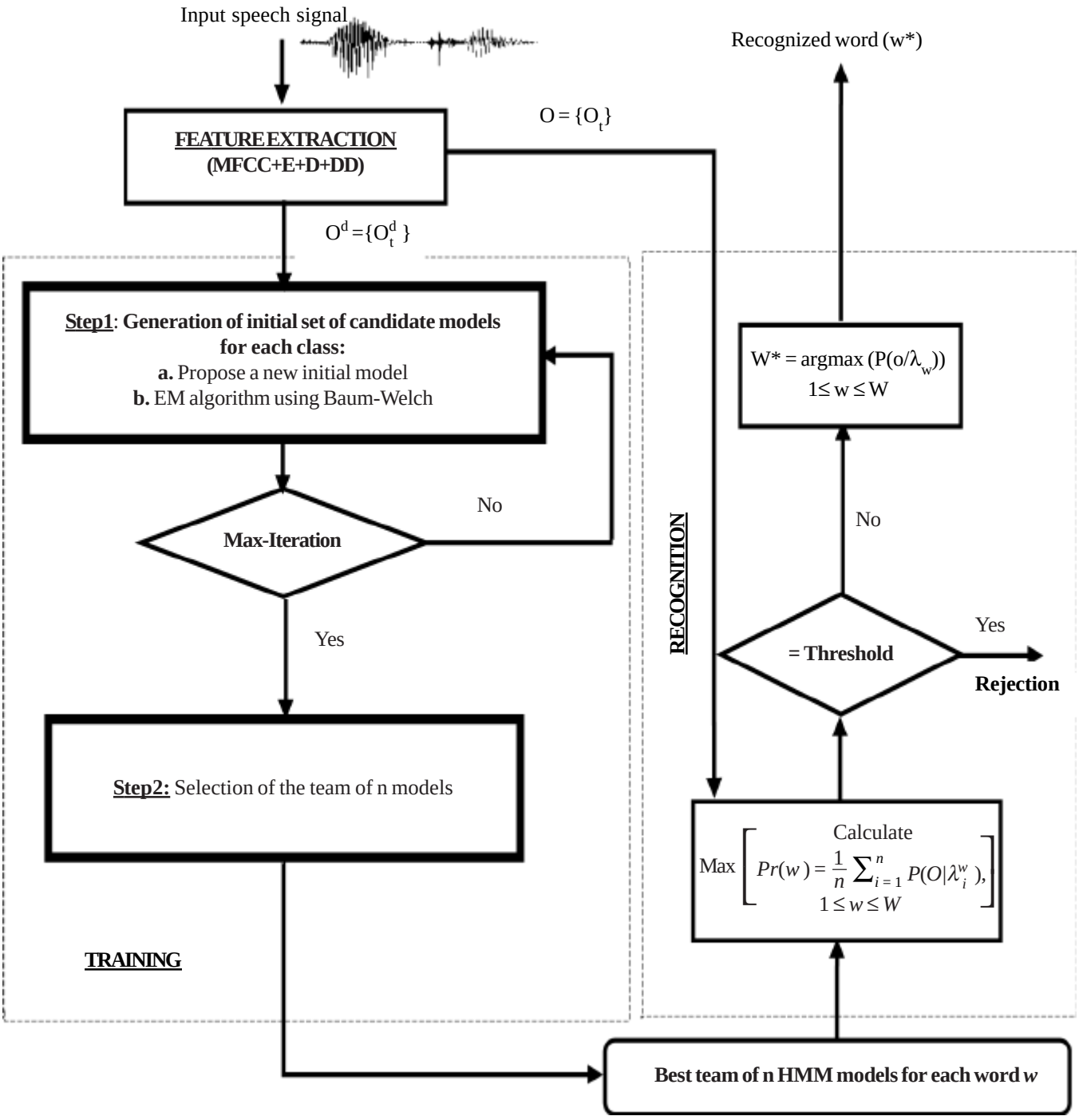

Figure 2. Block diagram of the proposed approach for isolated word recognition

(***) This step has been achieved by the well-known Baum-Welch algorithm [5].

In the following step, a selection method is applied to look for an optimal set of *n* HMMs among the various candidate HMMs. In order to reduce response time during the recognition phase, we assume that the number of selected models is $2 \leq n \leq 5$.

4.2.2 Selection of an Optimal Set of Models

In this step we have proposed two selection methods. The first one, is based on GA by maximizing the recognition rate of the

selected subset and minimizing the number of models in it. The second method is based on maximizing the difference between average likelihoods of the curent class and the average of likelihoods of the others (classes).

4.2.3 1st method: Selection by Genetic Algorithm

The genetic algorithm is a robust general purpose optimization technique, which evolves a population of solutions [9]. The canonical form of a GA is to generate a random population of individuals and while some criterion is not satisfied, evaluate and select the best individuals, mix them by crossover, and eventually mutate them to create a new population of individuals. Our GA operates according to this form.

a) Initial Population, Individuals Encoding and Fitness Evaluation

The initial population is a random set of 20 chromosome (*pop_size* = 20), each chromosome contains 100 genes coded in binary (each gene corresponds to a candidate HMM), where the presence of 1 means that the HMM corresponding to this gene is selected, and the presence of 0 means that it is not. In order to limit the response time, the number of selected HMMs in each chromosome must be between 2 and 5.

Once a generation is created the fitness value of each chromosome in the population is calculated. For our problem the fitness is the recognition rate.

b) Step 2: Reproduction

In this step, new individuals (offspring) are created by performing the following operations:

- Selection: The selection method is the elitist one. In this method, best chromosomes are selected.

- **Crossover:** After selection is over, one-point crossover is used, this point is chosen randomly between 2 and 99, the combination is done between the chromosome i ($1 \leq i \leq pop\_size/2$) and the chromosome j ($j = i + pop\_size/2$). After crossover, if we get offspring equivalent to their parents or offspring who do not check the condition ($2 \leq n \leq 5$), the cut point will be changed, and the crossing must be restarted again until the advent of new chromosomes. The crossover probability is $\rho\_c = 0.8$.

- **Mutation:** Contrary to the crossover, the mutation is done within the same chromosome and controlled by a low probability ($\rho\_m = 0.02$ in our case). All the chromosomes coming from the crossover are subject to mutation, where two positions in the chromosome are chosen randomly, and the corresponding genes are flipped, if these genes are identical then positions of mutation should be changed. This serves to ensure the evolution of populations and to exploit the research space.

- **Evaluation of the new population:** The evaluation is done by calculating the recognition rate for each chromosome of the new population.

- **Replacement (elitism)**: Fathers and sons are sorted in descending order of their fitness and the first (*pop_size*) chromosomes are chosen to be the individuals of the new population.

c) Step 3: Stopping Criteria

The stopping criterion that we have chosen is the number of generations be determined by experimentation, the convergence to a satisfactory solution is found after 20 generations. So the genetic algorithm stops when it reaches this value, and the n models having the highest fitness value are selected to be combined. If there are two sets having the same value of fitness, we choose which contains less of models, which reduces the response time of the system during the recognition phase. The result of this step is a team of n HMMs for each word.

Parameters (ρ_c, ρ_m, *pop_size*, *stop criteria*) and reproduction methods are all chosen after several experiments.

4.2.4 2nd method: Selection of models whose maximize the difference between likelihoods

In this method we have proposed to select models whose maximize the average of likelihoods of the current class, and minimize the average of likelihoods of the other classes. This method leads to enhance the separability inter-classes (margins between classes). Consequently, for each class w_{curent} we select the first n models according to the following equation.

$$\lambda^*_{w_{curent}} = argmax\,[mean\,(P(O^u_{w_{curent}} \mid \lambda^i_{w_{curent}})) - mean(P(O^u_{w_{curent}} \mid \lambda^i_{w_{curent}}))] \quad (6)$$

$1 \le i \le$ number of candidate HMM

$\lambda^{*}_{w_{curent}}$: the selected model for the current class

$O^{u}_{w_{curent}}$: the u^{th} utterence of w_{curent}

$1 \le u \le$ number of utterence of w_{curent}

w_{other} : the other classes

$1 \le v \le$ number of utterence of all the other classes

$O^{v}_{w_{other}}$: the v^{th} utterence of w_{other}

This method will be called HDLS and it means: Higher Difference of Likelihoods Selection.

4.3 Recognition

Once the training is done in order to design n different models for each *W* vocabulary word, recognition of unknown word defined by the sequence $O = o_1 o_2 \ldots o_T$ is done using Viterbi algorithm, but the different against single modeling approach is that instead of the selection of the word *w* whose model shows the highest likelihood $P(O|\lambda^{w})], 1 \le w \le W$, we select the word w whose team of its models λ^{i}_{w} maximize the quantity:

$$Pr(w) = \frac{1}{n} \sum_{i=1}^{n} P(O|\lambda^{i}_{w})], 1 \le w \le W \quad (7)$$

Where *n* is the number of models per word, λ^{i}_{w} is the i^{th} model of the word *w*.

$$i.e., w^{*} = \underset{1 \le w \le W}{argmax} \left[\sum_{i=1}^{n} P(O|\lambda^{i}_{w}) \right], 1 \le w \le W \quad (8)$$

(9)

This, on condition that *Pr* (*w*) is greater than or equal to a threshold associated with the word *w*, if not then the recognition decision is rejection.

Thresholds have been calculated empirically from the training data as follow. For each utterance (sample) of the word *w* in the training data, calculate *Pr* (*w*) and take the minimum divided by 3 as a threshold.

$$Thd(w) = \frac{1}{3} Min(Pr(O^{u}_{w}))$$

Where O^{u}_{w} is the u^{th} utterance for the word *w*.

5. Experimental Results and Discussion

The effectiveness of multi modeling approach and the proposed selection methods is evaluated on a one Arabic digit database containing the 10 digits (from 0 to 9) pronounced by 30 speakers (15 women and 15men); each digit is repeated 10 times by each speaker, so we obtain 300 utterances for each digit, which means 3000 speech files.

The data set is divided on 2 parts: 2000 utterances for training and 1000 utterances for the test. The utterances in the test data set are spoken by speakers who were not involved in training. So, the system is speaker independent.

In all experiments, all speech files are processed into Mel-Frequency Cepstrum Coefficients (MFCC), which include 13 cepstral coefficients, logarithm of energy and the first and second derivatives. More details are presented above in Section IV.A.

During our experiments, hundreds of tries are done for both classical and proposed approach. Results have shown that the recognition rate of the proposed approach where several models are combined in order to model each vocabulary word is always greater than the one obtained by classical markovian approach in which a single model is used for each word. Difference between models of the same word is created by changing in their initial models in the EM algorithm as explained in Section IV.B.

BIRS: Best Individual recognition Rate Selection.
RS: Random Selection.
GAS: Genetic Algorithm Selection.
HDLS: Higher Difference of Likelihoods Selection

		Single Modeling system (SM)		Multi modeling system (MM)
		Recognition rate: 1st single model system (%)	*Recognition rate: 2nd single model system (%)*	*Recognition rate: combination of the 2 models (%)*
BIRS	*Try A1*	89.0000	87.3333	91.7778
	Try A2	87.3333	87.8889	89.5556
	Try A3	86.2222	86.6667	88.4444
RS	*Try B1*	86.7778	83.4444	91.2222
	Try B2	86.7778	87.8889	91.7778
	Try B3	80.6667	83.4444	90.1111
GAS	*Try C*	80.6667	86.6667	95.0000
HDLS	*Try D*	89.400	90.000	96.4000

Table 1. Recognition Rates In Single Modeling Systems And Multi-Modeling Systems (N = 2 Models Per Word)

		Single Modeling system (SM)			Multi modeling system (MM)
		Recognition rate: 1st single model system (%)	*Recognition rate: 2nd single model system (%)*	*Recognition rate: 3rd single model system (%)*	*Recognition rate: combination of the 3 models (%)*
BIRS	Try 1	85.1111	89.0000	87.3333	94.0000
	Try A2	87.8889	87.3333	87.8889	90.1111
	Try A3	86.2222	85.6667	83.4444	91.2222
RS	Try B1	83.4444	82.8889	89.0000	94.5556
	Try B2	79.5556	85.1111	89.0000	92.8889
	Try B3	80.6667	81.7778	81.7778	95.1111
GAS	Try C	83.4444	83.4444	89.0000	96.3333
HDLS	Try D	89.200	89.400	90.000	97.0000

Table 2. Recognition Rates In Single Modeling Systems And Multi-Modeling Systems (N = 3 Models Per Word)

The choice of appropriate models to be combined is one of the most difficult but important factors in our approach. For this purpose, we have experimented four methods. In the first one, we selected models whose give best individual recognition rates. Secondly, we selected random models. Thirdly, we used a Genetic Algorithm to select the best team of models by maximizing the recognition rate of the ensemble and minimizing the number of models. In the fourth method, we selected models whose give the higher difference between the average of likelihoods of the current class utterances and the average of likelihoods of the other classes utterances.

In the following, we present the results of the classical system and the proposed one in 3 tries for each system. In table1, we take n = 2 models. In table2, we take n = 3 and in table3, we take n = 4. In all tables we evaluate the four selection methods.

In each try we have used the whole test dataset. The difference from a try to another is the initial set of candidate models.

		Single Modeling system (SM)				**Multi modeling system (MM)**
		Recognition rate: 1st single model system (%)	***Recognition rate: 2nd single model system (%)***	***Recognition rate: 3rd single model system (%)***	***Recognition rate: 4th single model system (%)***	***Recognition rate: combination of the 4 models (%)***
BIRS	**Try 1**	85.1111	89.0000	87.3333	85.6667	95.1111
	Try A2	87.3333	83.4444	84.0000	89.0000	92.8889
	Try A3	85.1111	86.7778	85.6667	85.1111	93.4444
RS	**Try B1**	86.7778	81.2222	85.1111	87.8889	95.6667
	Try B2	81.2222	87.8889	85.6667	81.2222	94.0000
	Try B3	79.0000	85.6667	82.8889	89.0000	94.5556
GAS	**Try C**	87.8889	87.3333	79.0000	85.6667	96.8889
HDLS	**Try D**	89.200	89.400	90.000	90.000	98.3333

Table 3. Recognition Rates In Single Modeling Systems And Multi-Modeling Systems (N = 4 Models Per Word)

From tables we remark that:

- The performance is very sensitive to the initial models; every time we change the initial model in training we obtain a new final model (difference in recognition rate from a single modeling system to another).
- The recognition rate of the MM approach where several models are combined in order to model each vocabulary word (columns highlighted in tables) is always greater than the one obtained by classical approach in which a single model is used for each word. This demonstrates the effectiveness of the multiple modeling approach.
- Models selected by BIRS method are not necessarily the best for combination (try A1, try A2 and try A3 vs try B1, try B2 and try B3).
- Genetic Algorithm selection (GAS) method is better than RS and BIRS selection (try C).
- The best results are obtained using HDLS method (try D).
- Three models per word seem a good compromise between the recognition rate and the response time of the system.
- Our proposed selection method HDLS can achieve significant improvement of recognition accuracy over both the Multiple Modeling and the Single Modeling system (try D vs try A, B and C).

In summary, the reported results show that the multiple modeling approach is better than classical one in all tries. The recognition rate has achieved the value of 96.88% using 4 models selected by GA for each vocabulary word and 98.33% using 4 models selected by HDLS method for each vocabulary word against (70% to 89%) in the single modeling approach. On the other hand, individual models selected by the HDLS method attain significant improvement of the recognition rate comparing against individual models selected using the other selection methods.

4. Conclusion

In this work, a novel speech recognition approach based on multiple markovian modeling is presented. When HMMs are used

on pattern recognition, the most important problem is how to select initial parameters. In fact, EM algorithm commonly used to estimate parameters is strongly dependent on the initial model. In the proposed approach, for each vocabulary word several HMMs coming from different random initializations are combined in order to increase the number of correctly recognized utterances. The proposed approach has been tested on one digit numbers dataset and produced significantly higher recognition rate in comparison with common markovian approach. It also, leads to improvement in the average of likelihoods for each unknown word which is to be recognized even if it is not correctly recognized. On the other hand, the proposed selection method leads to enhance the inter-classes separability by maximizing the margin (difference in likelihoods) between models of different classes.

To the best of our knowledge, use of the sensitivity to the initial models in the EM algorithm to create diversity between models in multiple modeling systems, and the enhancement of the inter-classes separability by maximizing the difference between likelihoods is a novel contribution. The proposed techniques are application independent and can be easily applied in any classification problem such as handwriting, face and gesture recognition.

References

[1] Hazmoune, S., Bougamouza, F., Mazouzi, S., Benmohammed, M. (2013). A Novel Speech Recognition Approach based on Multiple Modeling by Hidden Markov Models. *In*: Proc. of the IEEE International Conference on Computer Technology. ICCAT'2013. Sousse, Tunisia.

[2] Liu, N., Davis, R-I. A., Lovell, B-C., Kootsookos, P-J (2004). Effect of Initial HMM Choices in Multiple Sequence Training for Gesture Recognition. *In*: Proc. of the International Conference on Information Technology: Coding and Computing. International Conference on Information Technology (ITCC), Las Vegas, Nevada, U.S.A., p.608-613.

[3] Shamsul, H., Ghosh, R., Yearwood, J. (2006). A Variable Initialization Approach to the EM Algorithm for Better Estimation of the Parameters of Hidden Markov Model Based Acoustic Modeling of Speech Signals. Springer-Verlag, P. Perner (Ed.): ICDM, LNAI 4065, p. 416 – 430.

[4] Shamsul, H., Yearwood, J., Togneri, R. (2009). A stochastic version of Expectation Maximization algorithm for better estimation of Hidden Markov Model. Elsevier B.V, *Pattern Recognition Letters*, 30, p.1301–1309

[5] Rabiner, L. R (1989). A Tutorial on Hidden Markov Models and Select Application in Speesh Recognition. *In*: Proc. of the IEEE, 77 (2) 257-286.

[6] Korayem, M., Badr , A., Farag, I. (2007). Optimizing Hidden Markov Models using Genetic Algorithms and Artificial Immune systems. *Computing and Information Systems Journal*, University of PAISLEY, Great Britain 11(2) 44-50.

[7] Nathan, K., Senior, A., Subrahmonia, J. (1996). Initialization of hidden Markov models for unconstrained on line handwriting recognition. *In*: Acoustics, Speech, and Signal Processing. Atlanta.

[8] Peinado, A.M., Lopez, J. M., Sanchez, V. E., Segura, J. C., Rubio Ayuso, A. J. (1991). Improvements in HMM-based isolated word recognition system. *In*: Proc IEEE, 138 (3) 201-206.

[9] Goldberg, D. E. (1989). Genetic Algorithm in Search, Optimization & machine learning, Addison-Wesley.