Hybrid System of Optimal Self Organizing Maps and Hidden Markov Model for Arabic Digits Recognition

[1]ZAKARIAE EN-NAIMANI, [2]MOHAMED LAZAAR, [3]MOHAMED ETTAOUIL
[1,3]Modelling and Scientific Computing Laboratory,
Faculty of Science and Technology,
University Sidi Mohammed ben Abdellah, Fez, MOROCCO
[2]National School of Applied Sciences (ENSA)
Abdelmalek Essaadi University, Tetouan, MOROCCO
[1]z.ennaimani@gmail.com [2]lazaarmd@gmail.com, [3]mohamedettaouil@yahoo.fr

Abstract: - Thanks to Automatic Speech Recognition (ASR), a lot of machines can nowadays emulate human being ability to understand and speak natural language. However, ASR problematic could be as interesting as it is difficult. Its difficulty is precisely due to the complexity of speech processing, which takes into consideration many aspects: acoustic, phonetic, syntactic, etc. Thus, the most commonly used technology, in the context of speech recognition, is based on statistical models. Especially, the Hidden Markov Models which are capable of simultaneously modeling frequency and temporal characteristics of the speech signal. There is also the alternative of using Neuronal Networks. But another interesting framework applied in ASR is indeed the hybrid Artificial Neural Network (ANN) and Hidden Markov Model (HMM) speech recognizer that improves the accuracy of the two models. In the present work, we propose an Arabic digits recognition system based on hybrid Optimal Artificial Neural Network and Hidden Markov Model (OANN/HMM). The main innovation in this work is the use of an optimal neural network to determine the optimal groups, unlike in classical Kohonen approach. The numerical results are powerful and show the practical interest of our approach.

Key-Words: - Automatic Speech Recognition, Hidden Markov Models, Self Organizing Maps, Vector Quantization.

1 Introduction

Speech recognition is very useful in many applications, in our daily life, including command and control, dictation, transcription of recorded speech, searching audio documents and interactive spoken dialogues [3]. It has interested many scientists since the early 1950's. In fact because speech is the most essential medium of human being communication, he always tries to replicate this ability on machines. However, the processing of speech signal has to consider its different aspects, which complicates the task. Indeed, the signal information could be analyzed from different sides: acoustic, phonetic, phonologic, morphologic, syntactic, semantic and pragmatic [8] [29].

Artificial Neural Network (ANN) often called as Neural Network (NN), is a computational model or mathematical model based on biological neural networks.

The Artificial Neural Networks (ANN) are a very powerful tool to deal with many applications, and they have proved their effectiveness in several research areas such as analysis and image compression, recognition of writing, speech recognition , speech compression, video compression, signal analysis, process control, robotics, and research on the Web [25][20][28].

The Hidden Markov Model has become one of the most powerful statistical methods for modeling speech signals [15] [26] [27]. Its principles have been successfully used in automatic speech recognition, formant and pitch tracking, speech enhancement, speech synthesis, statistical language modeling, part-of-speech tagging, spoken language understanding, and machine translation. The Hidden Markov Model allows the temporal variation, with local spectral variability modeled using flexible distributions such as mixtures of Gaussian densities [12]. New techniques have emerged to increase the system performance speech recognition. These are based on the hybridization of Artificial Neural Network and Hidden Markov Model (ANN/HMM). The approaches ANN/HMM represent a competitive alternative to standard systems [4] [11] [16].

We describe in this paper an original and successful use of unsupervised ANNs and Hidden Markov modeling for speech recognition systems. Although it's conceptually weaker than continuous supervised systems like hybrid systems, the proposed approach in this paper offers performance and learning speedup advantages when compared to other discrete systems. In this paper, we use unsupervised ANNs called Self-Organizing Map (SOM) of Kohonen[20]. Since the training stage is very important in the Self-Organizing Maps (SOM) performance, the selection of the architecture of Kohonen network, associated with a given problem, is one of the most important research problems in the neural network research [13] [14]. More precisely, the choice of neurons number and the initial weights has a great impact on the convergence of learning methods. The optimization of the artificial neural networks architectures, particularly Kohonen networks, is a recent problem [13][14].In this work, we discuss a new method of Automatic Speech Recognition using a new hybrid ANN/HMM model, the main innovation is to use the optimal Self Organizing Maps (OSOM), which allows determining the optimal codebook and reducing the coding time unlike in classical approach.

This paper is organized as follows: The section 2 presents the speech recognition process. In section 3 the vector quantization by Kohonen networks is described. The Hidden Markov Model is presented in section 4. In section 5, we introduce the Hybrid System SOM/HMM for the recognition. In section 6 we give a presentation of the proposed approach, which consists on determining the hybrid system OSOM/HMM for the recognition based on optimal architecture SOM hybridized with the HMM. For more explanation, we propose a mathematic model optimizing the SOM architecture maps. This model, composed of an objective nonlinear function and a mixed variable, is subject to constraints. It is solved via a meta-optimization method, in this case the Genetic Algorithm. After this step, we construct an optimal dictionary, which constitutes the set of symbols used in the input of the HMM. And before concluding, experimental results are given in the section 7.

2 Speech Recognition Process

A recognition system is essentially composed of three modules:

- Acoustic analysis model;
- Learning model;
- Recognition model.

In the acoustic analysis step, the speech signal which is captured by a microphone, is converted to a digital signal and then analyzed in the acoustic analysis step. Once the signal analyzed, it is represented by vectors of suitable coefficients for word modeling [29].

The principal components of a large acoustic analysis are illustrated in Fig.1, the input audio waveform from a microphone is converted into a sequence of fixed size acoustic vectors O in a process called feature extraction.

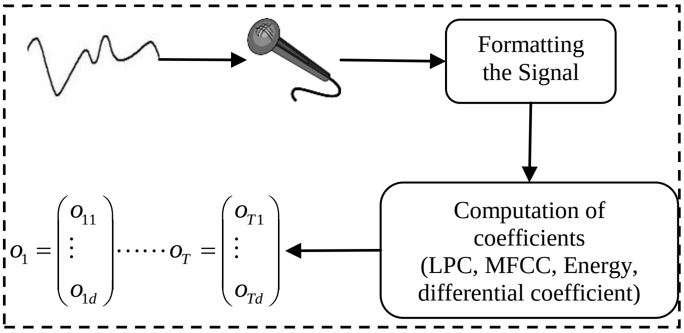
Formatting the Signal
Computation of coefficients (LPC, MFCC, Energy, differential coefficient)
$o_1 = \begin{pmatrix} o_{11} \\ \vdots \\ o_{1d} \end{pmatrix} \cdots\cdots o_T = \begin{pmatrix} o_{T1} \\ \vdots \\ o_{Td} \end{pmatrix}$

Fig.1, Acoustic analysis step

Feature vectors O are typically computed every 10 ms using an overlapping analysis window of around 25ms. In this work, we use one of the simplest and most widely used encoding schemes is based on mel frequency cepstral coefficients (MFCCs) [24].

The following sections describe the learning model and recognition model.

3 Vector Quantization

Vector quantization (VQ) is a process of mapping vectors from a vector space to a finite number of regions in that space [22]. These regions are called clusters and represented by their central vectors or centroids. A set of centroids, which represents the whole vector space, is called a codebook. In Speaker identification, VQ is applied on the set of feature vectors extracted from the speech sample and as a result, the speaker codebook is generated [10] [12] [19].

Mathematically a VQ task is defined as follows: given a set of feature vectors, find a partitioning of the feature vector space into the predefined K number of regions, which do not overlap with each other and added together form the whole feature vector space. Every vector inside such region is represented by the corresponding centroid [19].

In addition, vector quantization is considered as a data compression technique in the speech coding

[13] [14]. Vector quantization has also been increasingly applied to reduce complexity problem like pattern recognition. The quantization method using the Artificial Neural Network, particularly in Self Organizing Maps, is more suitable in this case than the statistical distribution of the original data changes with time, since it supports the adaptive data learning [14]. Also, the neural network has a huge parallel structure and the possibility for high-speed processing.

The Self Organizing Map (Kohonen neural network) is probably the closest of all artificial neural network architectures and learning schemes to the biological neuronal network. The aim of Kohonen learning is to map similar signals to similar neuron positions.

The Kohonen network has one single layer of neurons arranged in a two-dimensional plan; let's call it the output layer Fig.2, the additional input layer just distributes the inputs to output layer. The number of neurons on input layer is equal to the dimension of input vector. A defined topology means that each neuron has a defined number of neurons as nearest neighbors, second-nearest neighbors, etc. The neighborhood of a neuron is usually arranged either in squares, which means that each neuron has either four nearest neighbors. Kohonen has proposed various alternatives for the automatic classification, and presented the Kohonen topological map; these models belong to the category of unsupervised learning artificial neural networks without human intervention, the little information is necessary to respect the characteristics of input data. This model simply calculates Euclidean distance between input and weights.

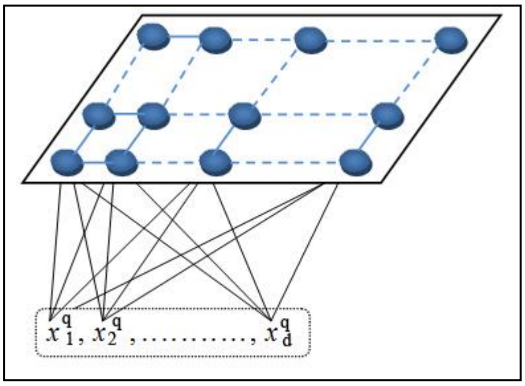

Fig.2, Kohonen topological map

To look at the training process more formally, let us consider the input data as consisting of n-dimensional vectors $X = \{x^1, x^2, \ldots, x^n\}$.

Meanwhile, each of N neurons has an associated reference vector $w^j = (w_1^j, w_2^j, \ldots, w_d^j)$.

During training, one x at a time is compared with all w^j to find the reference vector w^kthat satisfies a minimum distance or maximum similarity criterion. Though a number of measures are possible, the Euclidean distance is by far the most common:

$$k = \arg min_{j=1}^{N} \left\|x - w^j\right\| \qquad (1)$$

The best-matching unit (BMU) and neurons within its neighborhood are then activated and modified:

$$w^i(t+1) = w^i(t) + \beta_{k,i}(t)\left\|x - w^i\right\| \qquad (2)$$

One of the main parameters influencing the training process is the neighborhood function ($\beta_{k,i}(t)$), which defines a distance-weighted model for adjusting neuron vectors. It is defined by the following relation:

$$\beta_{k,i}(t) = \exp\left(\frac{-d_{k,i}}{2\sigma_i^2(t)}\right) \qquad (3)$$

One can see that the neighborhood function is dependent on both the distance between the BMU and the respective neuron ($d_{k,i}$) and on the time step reached in the overall training process (t). In the Gaussian model, that neighborhood's size appears as kernel width (σ) and is not a fixed parameter. The neighborhood radius is used to set the kernel width with which training will start. One typically starts with a neighborhood spanning most of the SOM, in order to achieve a rough global ordering, but kernel width then decreases during later training cycles [13].

4 Hidden Markov Models (HMM)

4.1 Introduction

The Hidden Markov Model is a stochastic automaton with a stochastic output process, where the output states are attached to each other. Thus we have two concurrent stochastic processes: an underlying (hidden) Markov process modeling the temporal structure of speech and a set of state output processes modeling the stationary character of the speech signal. For more detailed information about HMMs used in speech recognition, several texts can be recommended such as [26] [27].

A HMM is a finite set of states, each of which is associated with a probability distribution. Transitions among the states are governed by a set of probabilities called transition probabilities.

More precisely, an HMM is defined by:

$$\lambda = (Q, O, P, B, \pi) \quad (4)$$

Where:

- Q: set of finite states of the model of cardinal N;
- O: set of finite observations of the model of cardinal M;
- $P = (p_{ij})_{\substack{1\le i\le N \\ 1\le j\le N}}$: State transition probabilities matrix. Transition probability is the probability to chose the transition p_{ij} to access to the state q_i;
- $B = (b_j(o_t))_{\substack{1\le i\le N \\ 1\le t\le |T|}}$: probability matrix to emit observation o_t in the state q_j;
- $\pi = (\pi_1, \pi_2, \dots, \pi_N)$: initial distribution of states $\pi_j = P(q_0 = j), \forall j \in [1, N]$, where q_0 represents the initial state of the model.

4.2 Properties of the HMM used in ASR

In the ASR, some HMM properties hypotheses are commonly emitted to reduce the computing time. Thus,

- A Markov Model is stationary, so that:

$$P(q_t = s_i \mid q_{t-1} = s_j) = P(q_{t+k} = s_i \mid q_{t+k-1} = s_j) \quad \forall t, k \quad (5)$$

- The observations are considered as independent, i.e.:

$$P(o_t \mid q_1 q_2 ... q_t, o_1 o_2 ... o_{t-1}) = P(o_t \mid q_1 q_2 ... q_t) \quad (6)$$

- The probability of emitting an observation depends only of the courant state, i.e.:

$$P(o_t \mid q_1 q_2 ... q_t) = P(o_t \mid q_t) \quad (7)$$

However, in most of the current ASR systems, acoustic models are represented with HMM's of the 1st order; which means that the probability of being in a state s_i at a moment t knowing that $t-1$ states have been visited, is equal to the probability of being in the state s_i, knowing the state s_j previously visited. Literally:

$$P(q_t = s_i \mid q_1 q_2 ... q_{t-1} = s_j) = P(q_t = s_i \mid q_{t-1} = s_j) \quad (8)$$

4.3 Basic problems for HMM

The aim of this model is finding the probability of observing a sequence O. Therefore, the model has to provide a maximal probability. That's the reason why we are using HMM with its three fundamental problems: evaluation problem, decoding problem and learning problem [2] [26] [27].

4.3.1 Evaluation problem

Given the observation sequence $O = o_1 o_2 ... o_T$, and a model $\lambda = (\mathrm{P}, \mathrm{B}, \pi)$, how do we efficiently compute $P(O/\lambda)$ the probability of the observation sequence, given model?

To answer this question we will go through several stages:

Firstly, we will calculate the probability of observing the sequence O for a sequence of states Q is equal to:

$$P(O \mid Q, \lambda_c) = b_{q_1}(o_1).b_{q_2}(o_2)...b_{q_T}(o_T) \quad (9)$$

However, the probability of the sequence can be written as:

$$P(Q \mid \lambda) = \pi_{q_1} p_{q_1 q_2} p_{q_2 q_3} \cdots p_{q_{T-1} q_T} \quad (10)$$

Then,

$$P(O, Q \mid \lambda) = P(Q \mid \lambda).P(O \mid Q, \lambda) \quad (11)$$

Whence

$$P(O \mid \lambda) = \sum_{q_1,\dots,q_T} \pi_{q_1} b_{q_1}(o_1) p_{q_1 q_2} b_{q_2}(o_2) ... p_{q_{T-1} q_T} b_{q_T}(o_T) \quad (12)$$

The direct calculation of this formula is too complex and impossible to implement, Fortunately, there is a fast and efficient algorithm known as forward-backward to give a solution to effectively carry out this calculation [2] [26] [27].

4.3.2 Decoding problem

Given the observation sequence $O = o_1 o_2 ... o_T$, and the model λ, how do we choose a corresponding state sequence $O = q_1 q_2 ... q_T$ which is optimal in some meaningful sense (best explain the observation)?

To resolve this, we should apply the Baum Welch algorithm [2] [26] [27].

5 Hybrid System ANN/HMM For the Recognition

In the standard HMM system, each word is represented by a separate HMM [5]. In the learning stage, each pronunciation is converted to spectral domain (MFCC, energy and second order coefficients) that constitutes an observation sequence for the evaluation of HMM parameters associated to the word. The evaluation consists of the optimization of learning data probability corresponding to each word in the vocabulary.

Typically, the algorithm of Baum-Welch is used for the optimization.

In the recognition stage, the observation sequence representing the word to be recognized is used in the computation of probabilities, for all the possible models. The recognized word corresponds to the model with the highest probability, in this stage, the Viterbi algorithm is used.

Markov models are known for their lack of discrimination. Hence, to correct this deficiency, the main proposed solutions intervene in the models learning stage. Another alternative consists of introducing the discrimination locally in the models definition. One of the proposed methods is the use of the SOM as discriminant probability estimator. However, this technique of strengthening the discrimination by SOM presents difficulties in its implementation and the learning process doesn't have a convergence condition.

The SOM/HMM approach supposes that the global discrimination of Markov models could derive from a discrimination of the models learning sequence, by a transformation of the representation space using vector quantization. The idea is to generate reference prototypes for the vocabulary phonemes from the learning data, and then align all the learning data with the corresponding prototype. For this, we use the Kohonen network (SOM), which is the best prototype reference giving a significant difference between the training data.

To understand this approach we propose the following two schemes Fig.3, Fig.4.

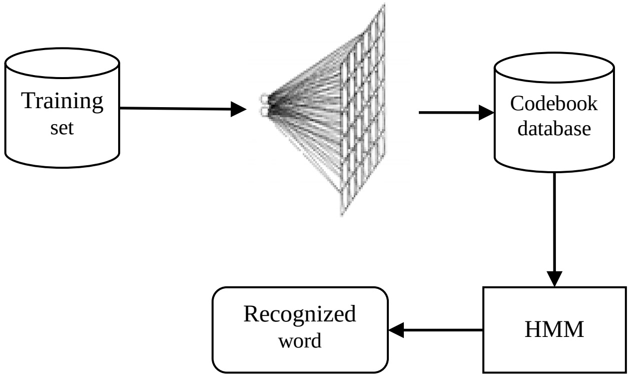

Fig.3, Recognition steps

After experiments, we found that this model presents many problems:

- Neurons number;
- Neurons weights initialization;
- HMM states number;
- HMM initial parameters definition.

In the following section we introduce a new model trying to solve these problems.

6 Hybrid System OSOM/HMM for the Recognition

These recent problems affect, indeed, the learning time of the model. In fact, the model as presented in Fig.4 needs a long time for learning. That's the reason why we propose an optimized model OSOM/HMM that reduces that time, by using Optimized SOM

6.1 Model Optimizing the Kohonen Architecture Maps

Since the size of the Kohonen topological map is randomly chosen, some neurons in the Kohonen topological map have a negative effect in the learning algorithm. To reduce this type of neurons, without losing of the learning quality, we propose the following proposed model.

6.1.1 Optimization model

In this section, we will describe the model construction steps of this model. The first one consists in integrating the special term which controls the size of the map. The second step gives the constraint which ensures the allocation of every data to only one neuron.

For modeling the problem of neural architecture optimization, we need to define some parameters as follows:

- n: Observation number of data base;
- p: Dimension of data base;
- N: Optimal number of artificial neurons in the Kohonen topological map;
- N_{min},: Minimal number of artificial neurons in the Kohonen topological map;
- N_{max},: Maximal number of artificial neurons in the Kohonen topological map;
- $X = \{x^1, x^2, ..., x^n\}$: Training base, where $x^k = (x_1^k, x_2^k, ..., x_p^k)$ for $k = 1, \ldots, n$.;
- $U = (u_{i,j})$: The binary variables for $i = 1, \ldots, n$ and $j = 1, \ldots, N_{max}$, $u_{i,j} = 1$ if the i^{th} example is assigned to j^{th} neuron, else $u_{i,j} = 0$.

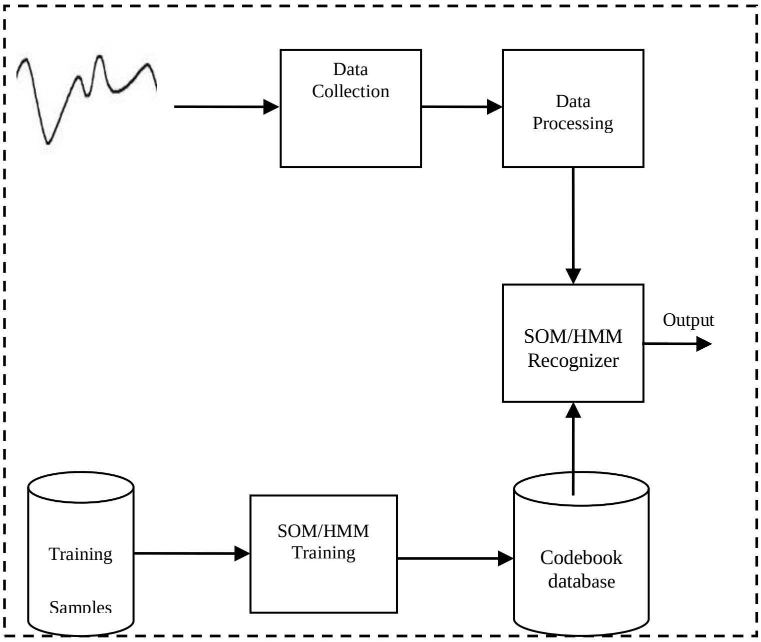

Fig.4, Training Step Hybrid Model SOM/HMM

Objective function

The objective function of nonlinear programming model is a summation of product of a decision variable of each neuron in the Kohonen map, the matrix of neighbors and the distances between the observations and the neurons, this function is defined by:

$E(U,W) = \sum_{i=1}^{n} \sum_{j=1}^{N_{max}} u_{i,j} K_j(\delta/T) \left\| x^i - w^j \right\|^2$ (13)

Where $K_j(\delta/T) = exp(-\delta/T)$, δ represents the distance on the map between the referent neuron for the observation x^i and his neighbors, $W = \{w^1, w^2, \dots, w^{N_{max}}\}$, $w^j \in \mathbb{R}^d$ and $j = 1, \dots, N_{max}$. If $u_{i,j} = 1$ then the i^{th} example x^i is assigned to the j^{th}neuron, and the corresponding error $\left\| x^i - w^j \right\|$ has been calculated on the objective function E.

Constraints

Each observation is assigned to a single neuron of the map; the constraint ensures that this assignment is:

$\sum_{j=1}^{N_{max}} u_{i,j} = 1 \quad \text{for} \quad i = 1, \dots, n$ (14)

To summarize what has been stated above, we introduce the following model.

Optimization Model

$$(P) \begin{cases} E(U,W) = \sum_{i=1}^{n} \sum_{j=1}^{N_{max}} u_{i,j} K_j(\delta/T) \left\| x^i - w^j \right\|^2 \\ Subject\ to: \\ \sum_{j=1}^{N_{max}} u_{i,j} = 1 \qquad \text{for } i = 1, \dots, n \\ u_{i,j} \in \{0,1\} \qquad j = 1, \dots, N_{max}, i = 1, \dots, n \\ w^j \in \mathbb{R}^d \qquad j = 1, \dots, N_{max} \end{cases} \tag{15}$$

6.1.2 Solving the Optimization Model Using Genetic Algorithm

We use the Genetic Algorithm approach to solve this mathematical model.

Genetic algorithm

The Genetic Algorithm (GA) was introduced by J. HOLLAND to solve a large number of complex optimization problems [18][6]. Each solution represents an individual who is coded in one or several chromosomes. These chromosomes represent the problem's variables. First, an initial population composed by a fix number of individuals is generated, then, operators of reproduction are applied to a number of individuals selected switch their fitness. This procedure is repeated until the maximums number of iterations is attained. GA has been applied in a large number of optimization problems in several domains, telecommunication, routing, scheduling, and it proves it's efficiently to obtain a good solution [9]. We have formulated the problem as a nonlinear program with mixed variables.

The relevant steps of GA are:

1. Choose the initial population of individuals.
2. Evaluate the fitness of each individual in that population.
3. Repeat on this generation.
4. Select the best-fit individuals for reproduction
   a. Crossover and Mutation operations.
   b. Evaluate the individual fitness of new individuals.
   c. Replace least-fit population with new individuals until termination (time limit, fitness achieved, etc.)

Genetic algorithm for mathematical model

In this section, we will describe the genetic algorithms to solve the proposed model for Kohonen networks architecture optimization. To this end, we have coded individual by tree chromosomes; moreover, the fitness of each individual depends on the value of the objective function.

- Initial population

The first step in the functioning of a GA is, then, the generation of an initial population. Each member of this population encodes a possible solution to a problem.

The individual of the initial population are randomly generated, and $u_{i,j}$ take the value 0 or 1, and the weights matrix takes random values in space $[x_{min}, x_{max}]^p$where $x_{min} = min\{x_k^i\}$ and $x_{max} = max\{x_k^i\}$ where $k = 1, \dots, p$ and $i = 1, \dots, n$. Because all the observations are in the set $[x_{min}, x_{max}]^p$.

- Evaluating individuals

After creating the initial population, each individual is evaluated and assigned a fitness value according to the fitness function.

In this step, each individual is assigned a numerical value called fitness which corresponds to its performance; it depends essentially on the value of objective function in this individual. An individual who has a great fitness is the one who is the most adapted to the problem.

The fitness suggested in our work is the following function:

$$f(i) = \frac{1}{1+E(i)} \quad (16)$$

Minimize the value of the objective function is equivalent to maximize the value of the fitness function.

- Selection

The application of the fitness criterion to choose which individuals from a population will go on to reproduce.

Where:

$$P_i = \frac{f_i}{\sum_{j=1}^{n} f_j} \quad (17)$$

- Crossover

The crossover is a very important phase in the genetic algorithm, in this step, new individuals called children are created by individuals selected from the population called parents. Children are constructed as follows:

We fix a point of crossover, the parent are cut switch this point, the first part of parent 1 and the second of parent 2 go to child 1 and the rest go to child 2.

In the crossover that we adopted, we choose 2 different crossover points, the first for the matrix of weights and the second is for vector U.

- Mutation

The rule of mutation is to keep the diversity of solutions in order to avoid local optimums. It corresponds on changing the values of one (or several) value (s) of the individuals who are (s)

Based on the standard model SOM/HMM and the optimal architecture model SOM, we propose in the next section a new approach that we call OSOM-HMM

6.2 Hybrid System OSOM/HMM for the Recognition

The aim of the speech signal acoustic analysis is to extract the local parameters, which are time driven multidimensional vectors.

In the proposed hybrid system, these are the steps of the word recognition system modeling:

1. Construct the optimal model of the auto organizing card associated to the learning basis. This model is represented by a mix variables nonlinear mathematic program (P).
2. Use the genetic algorithm to solve (P), and then obtain the neurons optimal number used in the card.
3. Generate a learning basis optimal dictionary using the SOM algorithm. This dictionary is the set of symbols used in the input of the Hidden Markov Model. In this step, we attribute to each feature vector a symbol corresponding to a code word in the codebook created by our approach (OSOM).
4. Construct an associated HMM to each word of the learning basis.

The scheme Fig.5 illustrates these 4 steps.

The hybrid system OSOM/HMM presents two huge advantages compared to the standard model SOM/HMM. The first advantage is the short learning time, and the second concerns diminution of the storage space (Memory).

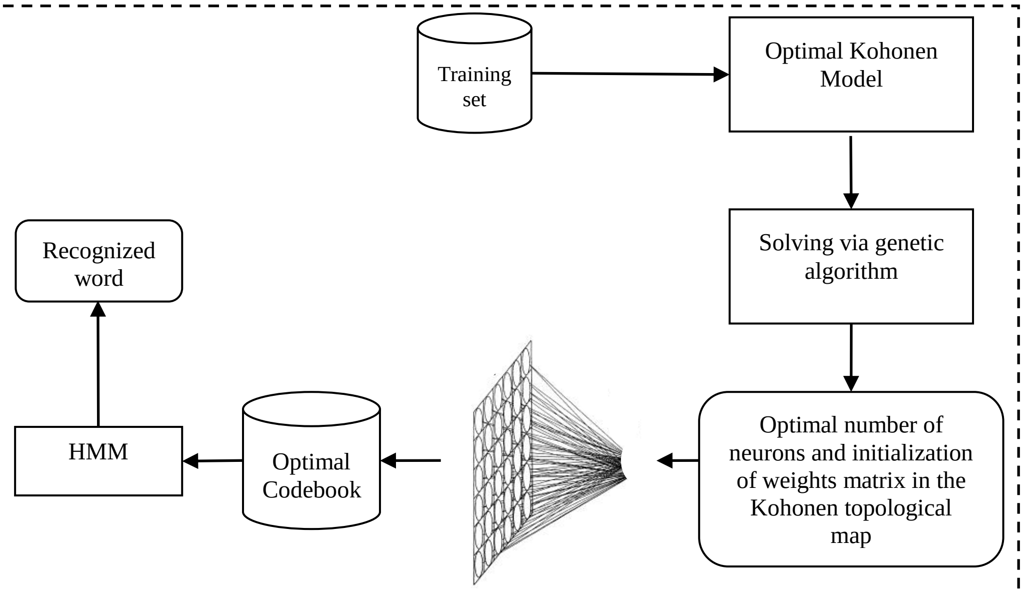
Training set
Optimal Kohonen Model
Solving via genetic algorithm
Optimal number of neurons and initialization of weights matrix in the Kohonen topological map
Optimal Codebook
HMM
Recognized word

Fig.5, Training hybrid Model OSOM/HMM

6.2.1 Iterative algorithm

We assume that given the I sets Ω_i, with N_i sequences each (repetition of the same word). The learning of all vocabulary bases:

$$\Omega=\bigcup_{i=1}^{I}\Omega_i \qquad \Omega_i\bigcap_{i\neq j}\Omega_j=\varnothing \qquad (18)$$

Where $\Omega_i=\left\{m_{i,1},\ldots,m_{i,N_i}\right\}$

- I is the number of words in the system vocabulary speech recognition.
- $m_{i,k}$ is representing the i repetition of word k.

Start

Step1:

- Calculating the optimal number of neurons N used via the optimization model.

Step2:

- Find codebook using Kohonen networks optimal $OSOM(\Omega,N)$.
- Outcome: codebook $O=\{o_1,\ldots,o_N\}$

Step3:

- Choose the number of states of the HMMs;
- $O=\{o_1,\ldots,o_N\}$ space of observation of the HMMs;
- An initial $HMM_i \quad 0\le i\le 9$.

Step4:

- Make learning HMM_i with Baum Welch;
- Outcome $HMM_i \quad 0\le i\le 9$.

7 Experiments Results

7.1 Dataset Description

The experiments were performed using the Arabic digit corpus collected by the laboratory of automatic and signals, University of Badji-Mokhtar - Annaba, Algeria. A number of 88 individuals (44 males and 44 females), Arabic native speakers were asked to utter all digits ten times [30]. Depending on this, the database consists of 8800 tokens (10 digits x 10 repetitions x 88 speakers). In this experiment, the data set is divided into two parts: a training set with 75% of the samples and test set with 25% of the samples. In this research, speaker-independent mode is considered.

Arabic	English	Symbol
صفر	ZERO	‘0’
واحد	ONE	‘1’
اثنان	TWO	‘2’
ثلاثة	THREE	‘3’
أربعه	FOUR	‘4’
خمسه	FIVE	‘5’
ستة	SIX	‘6’
سبعه	SEVEN	‘7’
ثمانية	EIGHT	‘8’
تسعه	NINE	‘9’

Table 1, Arabic Digits

Table 1 shows the Arabic digits, the first column presents the digits in Arabic language, the second column presents the digits in English language and the last column shows the symbol of each digit.

7.2 Numerical Results Obtained by Basic Hybridizing SOM/HMM

Numerical results obtained by applying the basic hybridization model SOM/HMM to the speech recognition are presented in the Table 2.

For each $HMMi\ 0 \le i \le 9$, we set the number of states to 48, which is the number of phonemes in the Arabic language. After, we chose the number of neurons (symbols of the HMM) randomly.

This table lists four sizes of neural architecture 8, 38, 100 and 420 neurons.

Digits	Size of state HMM = 48			
	Size of Neural network			
	N_{max} =8	N_{max} =38	N_{max} =100	N_{max} =420
0	0,65	0,77	0,86	0,79
1	0,93	0,90	0,94	0,77
2	0,92	0,97	0,92	0,69
3	0,84	0,89	0,87	0,69
4	0,87	0,82	0,87	0,87
5	0,87	0,89	0,92	0,67
6	0,80	0,89	0,91	0,60
7	0,85	0,83	0,84	0,80
8	0,88	0,90	0,77	0,79
9	0,89	0,90	0,9	0,85
mean	**0,85**	**0,87**	**0,88**	**0,75**

Table 2, Numerical results of classification

Fig.6 shows the comparison rate of recognition between the different sizes of neural maps (8, 38, 100 and 420). We remark from this graph that the good rate of recognition (mean 88%) through the number of groups equals 100 (100 neurons) and 48 states of HMM.

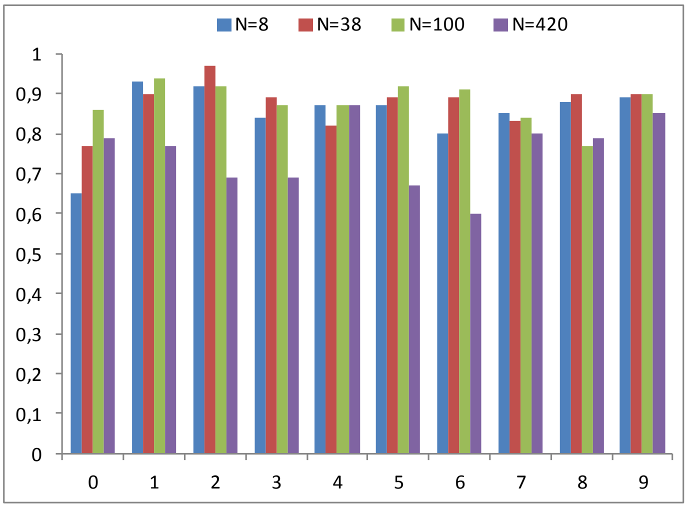

Fig.6, Rates of recognition with size of maps 8, 38, 100 and 420 neurons

7.3 Numerical Results Obtained by Optimal Hybridizing OSOM/HMM

To solve the optimization model (P), we propose a method using the genetic algorithm.

The most theoretical and logarithmical results are permit to determine the optimal number of neurons in the Kohonen topological map, and the good initial matrix of weights. The proposed approach for optimization of the Kohonen topological map is tested to realize the creating groups of our hybrid model OSOM/HMM. After the Table 2, we remark that the best result is achieved by the hybrid system which is based on a SOM of 100 neurons and HMM 48 states. We effectuate several tests with an initial architecture that contains 100 neurons to determine the optimal number of neurons.

Experimental results show that the mean number of neurons obtained by the proposed approach is 50 neurons.

	Size of Neural network			
Digits	N_{max}=100		N_{op}=50	
	T=48	T=10	T=48	T=10
0	0,86	0,78	0,85	0,73
1	0,94	0,92	0,95	0,93
2	0,92	0,72	0,89	0,8
3	0,87	0,8	0,9	0,78
4	0,87	0,86	0,88	0,8
5	0,92	0,83	0,9	0,83
6	0,91	0,9	0,89	0,85
7	0,84	0,8	0,87	0,75
8	0,77	0,82	0,81	0,83
9	0,9	0,9	0,9	0,86
mean	**0,88**	**0,83**	**0,884**	**0,822**

Table 3, Numerical results of classification

Numerical results obtained by applying the both models OSOM/HMM and SOM/HMM to dataset of Arabic digits are presented in the Table 3. This table lists the classification of the Arabic digits, for the respective sizes of maximal and optimal neural network N_{max}=100 and N_{op}=50 for both HMM number of states T=48 and T=10.

We remark that the result of the classification depends on the size of neural network and the number of states. With an optimal neural network of size 50 and 48 states of HMM, the recognition rate is equal to 88.4%. With 100 neurons it is equal to 88.0%.

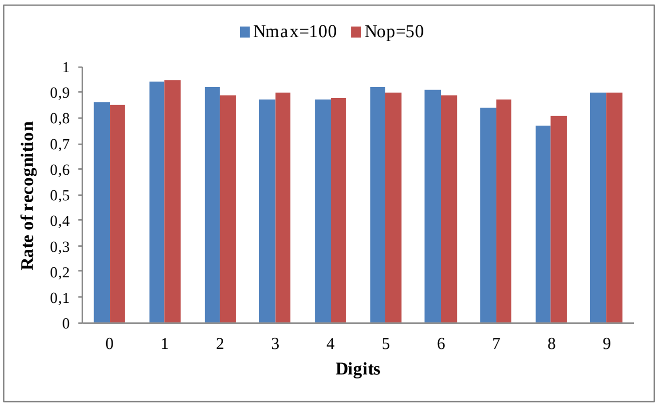

Fig.7, Rates of recognition with optimal size 50 neurons and 100 neurons with 48 states

From a numerical point of view, Fig.8 shows the importance of choosing the dictionary using neural network. This later plays an important role.

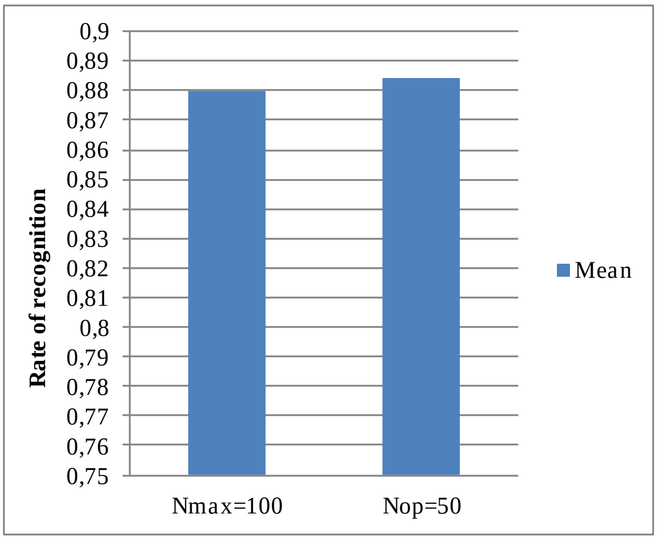

Fig.8, Mean rates of recognition with optimal OSOM/HMM and SOM/HMM

8 Conclusion

In this paper, we present a new approach to Speech recognition based on the construction of the optimal codebook by optimal Self Organizing Maps (OSOM) hybridized with the HMM.

As a first step we construct a mathematical model, and then we solve it via genetic algorithm. As a result we obtain the optimal number of neurons used in the map and the best initialization parameters of the OSOM that gives the optimal codebook. The codebook obtained constitutes the space symbol of the HMM of each Arabic digit.

To ensure the scientific contribution of the model, we compared the results of OSOM/HMM with those of SOM/HMM. As a conclusion we can tell that the results given by OSOM/HMM are promising and satisfactory in terms of memory usage and model convergence.

The main innovation is to use the optimal map of Kohonen network to generate the optimal codebook. The optimal codebook is used in the classification of Arabic digits using Hidden Markov Model. The robustness of the proposed method in the speech recognition is provided by the optimization of Kohonen architecture which determines the optimal codebook.

References:

[1] H. Bahi, M. Sellami, “A hybrid approach for arabic speech recognition” *ACS/IEEE international conference on computer systems and applications*,Tunisia,2003.

[2] L. Baum and T. Petrie, “statistical inference for probabilistic functions of finite state Markov chains”, *Proceedings of the IEEE*, 77(2), 1989, pp. 257-285.

[3] R. Boite and M. Kunt, “Traitement de la parole”, *presse polytechniques romandes*, 1987.

[4] H. Bourland, N. Morgan, "Hybrid HMM/ANN Systems for Speech Recognition : Overview and New Research Directions", in *Adaptive processing of sequences and Data Structures*, C.L Giles and M. Gori (Eds.), Lecture Notes in Artificial Intelligence (1387), Springer Verlag, pp. 389-417, 1998.

[5] E. Bourouba, M. Bedda and R. Djemili, “Isolated words recognition system based on hybrid approach DTW/GHMM”, *informatica* 30, pp.373-384, 2006.

[6] Camilleri M., Neri F., Papoutsidakis M. (2014). "An Algorithmic Approach to Parameter Selection in Machine Learning using Meta-Optimization Techniques". WSEAS Transactions on Systems,13, WSEAS Press (Athens, Greece), 202-213.

[7] A. Corradini and al. “A hybrid stochastic-connectionist approach to gesture recognition”, *international journal on artificial intelligence tools*, 9, pp. 177-204, 2000.

[8] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust.*, Speech, Signal Process., vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[9] J. Dréo, A. Pétrowski, P. Siarry and E. Taillard. "*Métaheuristiques pour l’optimisation difficile*". Eyrolles, 2003.

[10] H. Djellali, M. Laskri, “using vector quantization in automatic speaker verification”, *international conference on information technology and e-services*, 2012.

[11] R.Djemili, M. Bedda and Bourouba,”recognition of spoken arabic digits using neural predictive hidden Markov models”, *the international arab journal of information technology*.pp.226-233, 2004.

[12] R.Djemili, H. Bourouba and A.Korba,”A combination approach of Gaussian Mixture Models and support vector machines for speaker identification”, *international arab journal of information technology* vol 6 november 2009.

[13] M. Ettaouil, M. Lazaar, "Improved Self-Organizing Maps and Speech Compression", *International Journal of Computer Science Issues (IJCSI)*, Volume 9, Issue 2, No 1, pp. 197-205, 2012.

[14] M. Ettaouil, M. Lazaar, K. Elmoutaouakil, K. Haddouch, "A New Algorithm for Optimization of the Kohonen Network ArchitecturesUsing the Continuous Hopfield Networks", *WSEAS TRANSACTIONSon COMPUTERS, Issue 4, Volume 12, April 2013.*

[15] M. Gales and S. Young, “the application of hidden Markov models in speech recognition”, *foundations and tiends in signal processing* 2007.

[16] W. Gao and al. “A Chinese sign language recognition system based on SOFM/SRN/HMM”, *Pattern recognition* 37,pp. 2389-2402, 2004.

[17] N. Hammami, M. Bedda , "Improved Tree model for Arabic Speech Recognition", *Proc.IEEE, ICCSIT10*, 2010.

[18] J. Holland. "Adaptation in natural and artificial systems". Ann Arbor, *MI: University of Michigan Press*, 1992.

[19] A. Khan, N. Reddy and M. Rao,” Speaker recognition system using combined vector quantization and discrete hidden Markov model”, *international journal of computational engineering research*,pp. 692-696, 2012.

[20] T. Kohonen,”The self organizing maps”, *Springer*, 3th edition,2001.

[21] L. Lazli, M. Sellami,”Arabic speech clustering using a new algorithm”, *ACS/IEEE international conference on computer systems and applications*,2003.

[22] Y. Linde, A. Buzo, R.M. Gray,” An algorithm for vector quantization”, *IEEE trans COM-28*,pp.84-95, 1980.

[23] R. P. Lippmann, "Review of neural networks for speech recognition", *Neural Computing*, vol. 1, pp. 1- 38, 1989.

[24] L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques". *Journal Of Computing*, Volume 2, Issue 3, March 2010.

[25] M Papoutsidakis, D Piromalis, F Neri, M Camilleri (2014), "Intelligent Algorithms Based on Data Processing for Modular Robotic Vehicles Control", WSEAS Transactions on Systems, 13, WSEAS Press (Athens, Greece), 242-251.

[26] L. Rabiner, B.H. Juang. "Fundamentals of Speech Recognition", *edition Prentiee Hall PTR*, 1993.

[27] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE* 77 (2) ,257}286, 1989.

[28] S. Staines A., Neri F. (2014), "A Matrix Transition Oriented Net for Modeling Distributed Complex Computer and Communication Systems", WSEAS Transactions on Systems, 13, WSEAS Press (Athens, Greece), 12-22

[29] Q. Zhu, A. Alwan, “on the use of variable frame rate analisis in speech recognition”, *proc,IEEE ICASSP*, Turkey, vol 3,pp. 1783-1786, 2000.

[30] http://archive.ics.uci.edu/ml/datasets/Spoken+Arabic+Digit