The Egyptian International Journal of Engineering Sciences and Technology

Vol. 29 (2020) 61–67

http://www.eijest.zu.edu.eg

# Improving the Recognition Rate of Phonetic Arabic Letters Via Artificial Intelligent

Zahraa M. I. Aly[a*], Ehab R. Mohamed[b] and Ibrahim Zedan[c]

[a, b] *Faculty of Computers and Informatics, zagazig University*
[c] *Faculty of Engineering, zagazig University*

**ARTICLE INFO**

**Article history:**
Received 09 July 2018
Received in revised form 15 August 2018
Accepted 28 September 2018
Available online 00 September 2019

**Keywords:**
Speech recognition
Hidden Markov model
Neural networks
Frequency analysis
Feature extraction

**ABSTRACT**

It is very important to enhance the recognition accuracy of the Arabic spoken letters. The accuracy of recognition system is affected by the feature extraction and the used classifier. An effective and robust method is proposed to evaluate speech feature to improve the performance the recognition accuracy. This work introduces applying the mel frequency cepstral coefficient (MFCC) to extract the speech features. Hidden Markov model and neural network are used as classifier tools. The objective of the proposed system is to enhance the performance by introducing three systems which are proposed to recognize the spoken Arabic letters. The first is based on neural networks. The second is based on hidden Markov model. While third system is based on combination between neural networks and hidden Markov models. The accuracy of neural network is found to be 42% with MFCC for 84 spoken letters. The hidden Markov models are statistical based approach. Its performance is found to be 98.5%. But for combination system based on neural network and hidden Markov models, the accuracy of 99.25% is obtained.

## 1. Introduction

Speech is the easiest and most straight forward method of communication between humans and also the most natural and efficient form of exchanging information among human and machine. The two main problems facing the researchers in any speech recognition systems are:

- Accuracy or Recognition Rate.
- Computational Speed.

This work introduces a trying to investigate the recognition of the Arabic spoken letters to improve the accuracy or recognition rate.

### 1.1 Definition of Automatic Speech Recognition System

Spoken language is an important means of human communication. Automatic speech recognition (ASR) is a technology that helps a computer to recognize the words that a person speaks. It has a wide application area. It can be used in command recognition (voice control interface with computer) and dictation. It can be used in helping the handicapped people to deal and interact with society. It is a technology which makes life very promising and easier.

It is important to enhance the recognition accuracy of the Arabic spoken letters. The accuracy of recognition system is affected by the feature extraction and the used classifier. An effective and robust method is proposed to extract speech feature to improve the performance the recognition accuracy. This work introduces using the mel frequency cepstral coefficient to extract the speech features to enhance the speech recognition accuracy. Hidden Markov model and neural network are used as a classifier.

* Corresponding author. Tel.: +2-01009239012
*E-mail address*: mabdalla2010@gmail.com

In recent years, the real problem of speech recognition area has introduced in many research laboratories through the world. The research ultimate goal is to produce a machine which will recognize accurately the normal speech from any speaker. This machine could be used for many wide applications. It can be used for communicating between man and computers, office automation, factory automation, security systems and aides for handicapped.

Any utterance contains information about the words being spoken and also about the identity of the speaker. In a speech recognition it is wished to select the first type of the feature and ignore the second; in a speaker recognition system we wish to do just the opposite.

The variability due to differences in dialect is another problem facing the recognition systems. It is still needed to enhance the accuracy rate of the recognition for spoken languages although much research has been done for spoken English language. The spoken Egyptian Arabic is poor in the field of recognition in comparison to English language.

1.2 Speech processing

The first step in speech analysis is to capture the speech wave. Sampling and quantization is introduced to produce a sampled data representing the speech signal. There are many types of features that can be used to represent the acoustic data. The features of speech affect the recognition accuracy. It is important to select the best feature to get high accuracy rate. These features can be summarized as [1,2 3]:

- Mel frequency cepstrum coefficients (MFCC).
- Linear prediction coefficients (PLC).
- Linear prediction cepstrum coefficients (LPCC).
- Linear frequency cepstrum coefficients (LFCC).
- Perceptual Linear predictive coefficients (PLPS).

## 2. Feature Extraction: Literature Survey

It is important to select the best feature to get high accuracy rate. Many researchers had studied and compared between different features of speech. Feature extraction is the process where speech signal is converted into sequences of feature vector coefficients which contain necessary information.

There are many researches in the field of speech recognition. This field is still in need of improving the recognition accuracy especially in Arabic language. Feature vectors must have ability to differentiate among classes and should be robust to environment conditions such as noise [4, 5]. The researchers gave it much interest. Akansha discussed the advantages of MFCC and linear predictive coding [6]. He suggested combining other feature extraction with MFCC to be used for robust speech recognition systems. Suman K. Saksamudre et. al. [7] introduced different useful approaches for feature extraction of speech signals with their advantages and disadvantages. Shreye Narang and Ms. Divya Gupta [8] summarized the characteristic, advantages and disadvantages of linear predictive coding, Mel Frequency Cepstrum Relative Spectra (RASTA) and Probabilistic Linear Discriminate analysis. They concluded that MFCC is a feature extraction technique that is used widely for much speech recognition systems as it is mimic the human auditory system and it gives a better performance rate.

Namrata Dave [9] discussed the advantages and disadvantages of LPC, PLP and MFCC. He found that LPC parameter is not so acceptable because of its linear computation nature. PLP and MFCC are derived on the concept of logarithmically spaced filter bank, clubbed with the concept of human auditory system and hence they had the better response compare to LPC parameters.

M. A. Anusuya and S. K. Katti studied a comparison of different speech feature extraction techniques with and without wavelet transform to Kannada speech recognition [10]. They investigated the features based on linear predictive coefficient (LPC), Mel Frequency Cepstral Coefficients (MFCC), and Relative Spectral Transform (RAST). They showed that using the wavelet analysis with LPC, MFCC, and RAST improved the accuracy rate of the recognition.

From above discussion, one can conclude that using the wavelet analysis with LPC, MFCC and RASTA, the accuracy rate is increased. Sayf A. et. al. introduced a study about Mel Frequency Cepctrum Coefficient (MFCC) feature extraction to enhance the speech recognition rate[11].

Fang Zheng et. al. made a comparison between different implementation of MFCC [12]. They found that MFCC are more efficient. The best present algorithm in feature extraction is Mel Frequency Cepstrum Coefficient (MFCC) introduced in reference

[13]. The performance of MFCC degrades rapidly because of noise [14]. This drawback can be avoided by combining MFFC with wavelets analysis.
Yousef Ajami introduced a comparative study about using ANN and HMD for recognition of Arabic digits. The recognition system based on ANN achieved a recognition rate of 99.5% in the case of multi-speaker mode and 94.4% in the case of independent mode. On the other hand the recognition system based on HMM achieved 98% accuracy rate in the case of multi-speaker and 94.89% for speaker independent mode [15]. Many Arabic ASRS were introduced using ANN techniques [15,16,17]. Their data sets were the spoken Arabic digits.

### 3. The proposed Feature Extraction Technique

The extraction and selection of the best parametric representation of acoustic signals are important task in the design of any speech recognition system. Speech signals are Quasi–stationary signals. When speech signal is examined over a sufficient short period of time ( 5- 100msec ), its characteristics are fairly stationary. The information in speech signals is actually represented by short term amplitude spectrum of speech signal. This allows us to extract features based on the short term amplitude spectrum from speech.
To improve accuracy of the recognition system, an efficient, effective, and robust method is introduced to extract the speech features. It is found that MFCCs are more efficiently [10,12,14]. This is because MFFCs are mimic the human auditory system. MFCCs are based on known variation of the human ear's critical bandwidth with frequency. Fig.1 shows the block diagram to compute MFCC [18].
MFCC block diagram consists of seven computational stages. Each stage has its function and mathematical approaches as discussed in [18,19]. First stage is to process the passing of the signal through a filter which emphasizes higher frequencies. The energy of the signal will be increased at higher frequencies. The filter transfer function is given by [3]:

$$Y[n]=x[n]-0.95x[n-1] \tag{1}$$

The speech signal is separated into blocks with small duration. This duration is taken to be in the range 20 – 40 msec. The cepstral analysis is performed on these frames.
After partitioning the signal into frames, each frame is multiplied by a window function. This is to reduce the discontinuity introduced by framing process. Hamming window which is used for speech recognition task is given by [20]:

$$W[n] = 0.54 - (0.46)\,\mathrm{COS}[(2n\Pi)/(N-1)] \qquad 0 \le n \le N-1 \tag{2}$$

The time domain sample is converted into frequency domain using FFT. FFT is efficient algorithm of DFT. FFT is performed to calculate the magnitude frequency response for each frame [20].
The signal is applied to group of triangle band pass filters. This is to simulate the characteristics of the human's ears. The purpose of Mel filtering is to model the human auditory system that perceives sound in a linear scale [21]. The Mel frequency is computed from the linear frequency as:

$$f_{Mel} = 2525*\log_{10}(1+f/700) \tag{3}$$

where $f_{Mel}$ is the Mel frequency corresponding to linear frequency scale. The log Mel spectrum is converted into time domain using Discrete Cosine Transform (DCT). The output of Discrete Cosine Transform is called Mel Frequency Cepstrum Coefficient (MFCC).
Delta Energy and Delta Spectrum block is used to calculate the energy of the speech from the time domain signal.

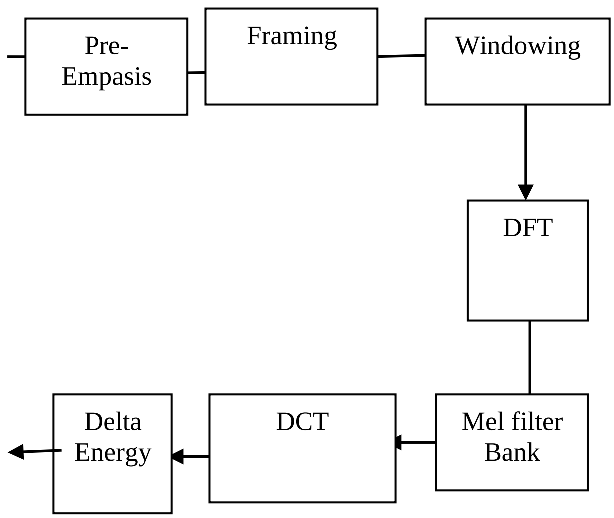


Fig. 1 MFCC block diagram

### 4. Speech Recognition Techniques

There are many methods for speech recognition techniques. They can be summarized in the following paragraphs [22, 23, 24].

**(i)-Template – based approaches**:

The basic method for computing the distance between input signal and its template is the Dynamic. Dynamic Time warping (DTW) is an efficient method in speech recognition [24, 25,26].

**(ii)- Knowledge –based approaches:**

An expert knowledge about variation in speech is hand-coded into a system. [27].

**(iii)- Learning – based approaches:**

Neural networks, support vector machine and genetic algorithm programming may be used as learning based approaches to speech recognition. Neural networks have the ability to learn. Neural networks can be used to model the speech signals.

**(iv)- Statistical based approach**

Hidden Markov models are statistical based approach. It is a flexible and successful approach to speech recognition process [28].

### 4.1 Element of HMMs

An HMM can be characterized by the following parameters [22,29, 30]:

**N**, it is the number of states in the model.

**M**, it is the number of distinct observation symbols per state.

The state transition probability distribution and it is given by:

$$a = \{a_{ij}\}$$

$$a_{ij} = P(q_{t+1}=S_j/\ q_i=s_i) \qquad 1 \le j \le N$$

The probability distribution of the observation in state j, and it is given by:

$$B = \{b_j(k)\}$$

The initial state distribution $\Pi = \{\pi_i\}$

where:

$$\Pi_i = P\{q_1=s_i\} \qquad 1 \le i \le N$$

For given values of N, M, A, B, and Π, the HMM can be used as generator to give an observation sequence [28].

### 4.2 Hidden Markov Model Algorithms

There are three algorithms associated with Hidden Markov Model [28].

(i)The forward algorithm:

it is useful for isolated word recognition. This algorithm is used to adjust or optimize the model parameters.

(ii)The viterbi algorithm:

It is useful for continuous speech recognition;

(iii) The forward – backward algorithm:

It is useful for training an HMM.

### 4 3 Parameter Estimation

To train a HMM model an initial values of the model parameters are assumed. For each training sequence, the parameters of a new model $M_{new}$ are re-estimated from those of old model $M_{old}$ until $M_{new}$ is a better model of the training sequence than $M_{old}$, that is [1]:

$$Pr\{O/M_{new}\} \geq Pr\{O/M_{old}\} \qquad (4)$$

This is repeated unite the proper results is obtained.
Baum-Welch algorithm can be used to evaluate the model parameters [30]:

### 4.4 Neural Network Classifier Approach

Artificial Neural Networks (ANNs) have been widely used during the past two decades in the speaker recognition systems. These models are composed of many simple processing elements called neurons which are referred as the McCulloch-Pitts neuron[31].

#### 4.4.1 The Neural Network Model

The NN can be constructed using number of layers which are input layer, hidden layer(s), and an output layer [13]. The layer is called the input layer which has weights coming from the input. All subsequent layer(s) except the last one is (are) called the hidden layer(s) where their weights come from the previous layer. The last layer is called the output one. All layers have what is called bias.[31]. The weight and biases of all layers are updated through many-iterations using a training algorithm.

#### 4.4.2 Neuron Model

**E**ach neuron of the NN simulates a biological neuron [31]. The neuron unit has a multiple inputs (*x*) and only one output (*y*) which can be represented by:

$$y_{(t)} = f\left(\sum_{i=1}^{n} w_i . x_{i(t)} - \theta\right) \qquad (5)$$

$w_i$ is a weight of the connection, Θ is the bias of each neuron unit, t is time, and n is the number of inputs to the NN. The neuron has an output function f(x) called activation function. There are many activation function used in training neural networks [31].

## 5. Experiments and results

### 5.1 Data base

The data base used in this work contains the speech data files of 5 speakers (three men and 2 females). The speech files contain the spoken Arabic letters from Alf to yaha with three haraka for each letter. Each speaker repeats each letter 3x20= 60 which saved in 60 files. Each speaker recorded 60x28= 1680 files. The total data files of 5 speakers are 1680x5=8400 files. 20% of the data are selected for testing the models. 80% of the data are used to training and implementing the models. Data were recorded using a high sensitivity microphone. All data files are stored in Microsoft wave format files with 44100 Hz sampling rate, 16 bit PCM with mono channel.

### 5.2 Results and Discussion

A phonetic recognition system is implemented to recognize phonetic Arabic letter by any speaker (male, female). This is done through a robust feature of speech. After extracting the feature, three methods of analysis is used to model the speech. The acoustic signal of the Arabic letter is collected by high sensitively microphone. The signal is analysed to calculate the MFCC coefficient of each letter. The speech signal (spoken Arabic letter) in this work is modelled in three systems. The first model is based on neural networks while the second is based on hidden Markov analysis and the third model is based on combination between Hidden Markov model of each letter is implemented. Neural network for recognition system is implemented as well as the Hidden Markov model.

### 5.2.1 Neural Network model

The main challenge in designing the neural network model is choosing the number of hidden layer and the number of neurons in each layer. Each frame of speech letter is analyzed into 40 coefficients. These coefficients are the input to the input layer of the neural network. So, the neural network has to have 40 neurons. There are 84 letters. So, the output layer has to have 7 neurons. Several experiments were applied until reaching the proper number of hidden layers as well the number of neuron in each layer.

Back propagation learning algorithm is used to adjust the weights and the network parameters. The c++ source code is written to implement the back propagation algorithm. After reaching to the proper training of the neural network to adjust the weights to minimize the error, the structure of the neural network is found to one hidden layer with 240 neurons in this layer in addition to 40 input neurons and 7 neurons in the output layer.

The recognition accuracy of the neural network is found to be 42%. It is unexpected results. This is because the accuracy depends on the numbers of recognition words. The recognition word is 84 words. It is poor recognition rate with MFCC although this feature is robust. It does badly with neural network. This poorness is due to the large number of recognition words. Fig. 2 shows the training error.


Fig. 2 The normalized mean square of the trained neural network

### 5.2.2 Hidden Markov Phase (HMM)

Another method to model the spoken letters must be looked for instead of the neural network. Hidden Markov is a stochastic modeling process defined by a set of state and a number of mixture densities

#### 5.2.2.1 Building HMMs

In this work, a recognition system is developed using HMM tool box. The models are built as follow:

- Markov model has 40 states left to right transition and these models are built for each spoken word letter. The probability of each state is modeled using 40 mixtures Gaussian distribution. These models are evaluated for the output of each state.

- Training HMMs

In the training stage, the observations are used to adjust the model parameters. The parameters to be estimated are;

- The initial state probabilities $\pi_i$
- The coefficients of the state transition probabilities $\alpha_{ij}$
- The mixture coefficients $C_{im}$
- Mean vector for the mixture components $\mu_{jm}$
- Covariance matrix for each model

HMMs are trained for each spoken letter to adjust the model parameters.

An HMM, $\lambda$, is a 5 tuple consisting of

- N is the number of states
- The starting state probabilities $(\pi_1, \pi_2, \ldots, \pi_N)$
- M is the number of possible observations
- The state transition probability $P(q_{t+1}=S_j/q_t=S_i) = a_{ij}$
- The observation probability $P(X_t=k/q_t=S_i) = b_i(k)$

The forward recursion algorithm is used to calculate and adjust the model parameters of 84 HMMs for each Arabic spoken letter. It works by guessing the initial parameter values, then adjusting the model parameters through training the HMM models.
The accuracy of HMM model is found to be 98.25% for training data while it found to be 60% for test data. It is clear that the HMM performance is poor. This is due the large number of spoken letters (84 spoken letters). It is obvious that the performance of HMM model is better than the neural network performance for the same features.

### 4.2.3 Neural –Hidden Markov Model Phase

In this phase, a system based on neural networks and Hidden Markov models is used to model the spoken Arabic letters. Firstly the neural network recognizes the letter without haraka. After recognizing the letter, neural network activates the hidden system to recognize the haraka. After training the neural network and building the hidden Markov model for each letter, the accuracy is found to be 99.25%.

## 6. The Conclusions

Speaker recognition is one of the most important fields of research in our daily life. It has been a challenging problem all over the past decades and still continuing just to obtain a better recognition rate.
This work introduces three approaches for spoken Arabic letters recognition. The first one is based on neural network. The second is based on hidden Markov models. The third approach depends on a combination from neural networks and hidden Markov models. An effective and robust feature is proposed for three systems. To improve the accuracy of recognition system MFCCs are used where they are mimic the human auditory system. Markov model is used as a classifier. The objective of the proposed system is to enhance the performance by introducing different systems. Three systems are proposed to recognize the spoken Arabic letters. The accuracy of neural network is found to be 42% with MFCC for 84 spoken letters. The hidden Markov model is based on probabilities. Its performance is found to be 98.5%. But for combination system based on neural network and hidden Markov models, the accuracy of 99.25% is obtained.

## 7. Future Work

The recognition of phonetic Arabic letters can be investigated using the feature based on the energy per frame. A comparison between using different feature can be made to select the most significant feature.

## References

[1] Sumank Saksamudre, P. P. Shrishrimal, and R. R. Deshmukh, "A Review on Different Approaches for Speech Recognition Systems" International Journal of Computer Applications" Volume 115, No. 22, April 2015

[2] Bishnu Prasad Dasl, and Ranjan Parkh, " Recognition of Isolated Words Using Features based on LPC, MFCC, ZCR and STE, with Neural Network Classifiers", International Joural of Modern Engineering Research, Vol. 2, Issue 3, May- June 2012.

[3] Mayur R. Gamit, Prof. Kinnal Dhameliya, and Ninad Bhatt, " Speech Recognition: A Review" International Journal of Emerging Technology and Advanced Engineering, Volume 5, Issue 2, Febrary 2015.

[4] Sonia Sunny, David Peter S. and K. PouloseJacob" Recognition of Speech Signals: An Experimental Comparison of Linear Predictive Coding and Discrete Wavelet Transforms", International Journal Of Engineering Science and Technology (IJEST), April 2012.

[5] Anjivani S. Bhabad and , Gajanon K. Kharate, "An overview of Technical Progress in Speech Recognition "International Journal of Advanced Research in Computer Science and software Engineering, March 2013, volume 3, issue3.

[6] Akansha Madan and Divya Gupta"Speech Feature Extraction and Classification: A Compartive Review" International

Journal of Computer Applications, volum 90, No, 9, March 2014.

[7] Suman K. Saksamudre, P. P. Shrishrimal and R. R. Deshmukh, "A Review on Different Approaches for Speech Recognition System" Internationalof Computer Applications, volume 115, N. 22, April 2015.

[8] Shreye Narang and Ms. Divya Gupta, "Speech Feature Extraction Techniques: A Review" International Journal of Computer Science and Mobile Computing, Vol. 4, Issue 3, March 2015.

[9] Namrata Dave, " Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition" International Journal of Advance Research in Engineering and Technology, volume 1, Issue VI, July 2013

[10] M. A. Anusuya and S. K. Katti," Comparison of different Speech Extraction Techniques with and without Wavelet Transform to Kannada Speech Recognition " International Journal of Computer Applications, Volume 26, No. 4, July, 2011

[11] Sayf A. Majeed, Hafizah Husain, Salina abdul Samad , and Tariq F. Idbeay, " Mel Frequency Cepstral Coefficients (MFCC) Feature Extraction Enhancement in the Application of Speech Recognition : A Comparison Study" Journal of Theoretical and Applied Information Technology, 10th September, 2015, vol.79, No. 1, ISSN 1882- 8645.

[12] Fang Zheng, Guoliang Zhang and Zhanjiang Song" Comparison of Different Implementation of MFCC" , J. of Computer science & Technology, 16 (6), sep. 2001.

[13] S. Davis and P. Mermelstein, " Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences ", Acoustic, Speech and Signal Processing, IEEE Transaction, Vol.23, 1980.

[14] N. S. Nehe and R. S. Holambe, "Isolated word Recognition Using Normalized Teager Energy Cepstrum Feature", Advanced in computing Control & Telecommunication Technology, Act o9 International Conference on 2009.

[15] Yousef Ajam Alotaibi, " Comparative study of ANN and HMM to Arabic Digits Recognition Systems" JKAU, Eng. Sci. vol.19, No. 1,2008.

[16] Yousef Ajam Alotaibi, "High performance Arabic Digits Recognition Using Neural Networks ", International Joint Conference on Neural Network 2003, IJCNN, Portand, Oregan.

[17] Yousef Ajam Alotaibi, "Spoken Arabic Digits Recognizer Using Recurrent Neural Networks", International Symposium on Signal Processing and Information Technology, ISSPIT, Rome, Italy 2004

[18] Lindasalwa Muda, Mumtaj begam, and I. Elmvazuthi, " Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques" Journal of Computing, Volume 2, Issue 3, March 2010.

[19] Santosh V. Chapaneri, " Spoken Digits Recognition Using Weighted MFCC and Improved Feature for Dynamic Time Warping" International Journal of Computing, Volume 2, Issue 3, March 2010.

[20] Sayp A. Majeed, Haflzah Husain, Salina Samad Technology, Vol. 79, No. 1, and Tariq F. Idbena, " Mel Frequency Cepstral Coefficient (MFCC) Feature Extraction Enhancement in the Application of Speech Recognition: A Comparison Study" Journal of Theoretical and Applied Information Sept. 2015.

[21] L. Rabiner and B. H. Juang, "Fundamentals of speech Recognition " Prentic Hall, 1993.

[22] Bhupinder Singh, Neha Kapur, and Puneet Kaur, "Speech Recognition with Hidden Markov Model: A Review " International Journal of Advanced Research in Computer Science and Software Engineering, Vol.2 Issue, March 2012, ISSN: 2277128.

[23] M. Matton, " Distance measure for template based and pattern recognition" Ph.D thesis, numerical analysis and applied mathematic, 2009.

[24] J. Tebelskis, " Speech Recognition using Neural Networks" PH.D Thesis , school of Computer Science; Mellon University, Pittsburgh PA, 1995.

[25] Bich Ngoc Do, "Neural Networks for Automatic Speaker, Language and Sex Identification" PH. D. thesis, Faculty of Arts, University of Groningen, 2015.

[26] F. Itakura, "Minimum Prediction Principle Applied to Speech Recognition" IEEE, Trans. On Acoustic, Speech and Signal Processing, 23(1), February 1975.

[27] P. Jackso, "Introduction to Expert System" Addison- Weskey, 1999.

[28] Gernot A. Fink, " Markov Models for Pattern Recognition From Theory to Application" Springer – Verlag, Berlin Heidlberg, 2008.

[29] Lawrence R. Rabiner, " A Tutorial on Hidden Markov Models and selected Applications in Speech Recognition" 1989.

[30] Oliver Cappe, Eric Moulines and Tobias Ryden, "Inference in Hidden Markov Models" Springer Science+ Business Inc, 2005Proceeding of IEEE, vol.77, No. 2, February

[31] Jacek M. Zurada "Introduction to Artificial Neural Systems" West Publishing Company 1992.