Arabic Lip-reading System: A Combination of Hypercolumn Neural Network Model with Hidden Markov Model

Sagheer, Alaa
Department of Intelligent Systems, Kyushu University

Tsuruta, Naoyuki
Department of Electronics Engineering and Computer Science, Fukuoka University

Taniguchi, Rin-ichiro
Department of Intelligent Systems, Kyushu University

出版情報：Proceedings of International Conference on Artificial Intelligence and Soft Computing. 2004, pp.311-316, 2004-09
バージョン：
権利関係：

ARABIC LIP-READING SYSTEM: A COMBINATION OF HYPERCOLUMN NEURAL NETWORK MODEL WITH HIDDEN MARKOV MODEL

Alaa El.Sagheer[1], Naoyuki Tsuruta[2], Rin-ichiro Taniguchi[1]
(1) Department of Intelligent Systems, Kyushu University,
6-1 Kasuga-Koean, Kasuga, Fukuoka 808-0196, JAPAN
(2) Department of Electronics Engineering and Computer Science, Fukuoka University
8-19-1, Nanakuma, Jonan, Fukuoka 814-0180, JAPAN
{alaa, rin}@limu.is.kyushu-u.ac.jp, tsuruta @tl.fukuoka-u.ac.jp

ABSTRACT

In recent year, lip-reading systems have received much attention, since it plays an important role in human communication with computer especially for hearing impaired or elderly people. In this paper, we introduce a new visual feature representation combines the Hypercolumn Neural Network model (HCM) with Hidden Markov Model (HMM) to achieve a complete lip-reading system. To check our system performance we introduce the Arabic language to it. According to our knowledge, this is the first time that a visual speech recognition system is applied for Arabic language. Experiments include different Arabic sentences gathered from different native speakers (Male & Female).

KEY WORDS

Lip-reading system, visual features extraction, hypercolumn neural network model, and Hidden markov model.

1. Introduction

Human-Computer Interface has undergone much advancement over the last few years as the computer's role in our lives is becoming main and vital. Lip-reading is one of the most fertile topics of interface with computer, since it can smooth the Human-Computer Interface by introducing the Human-Human interaction mechanisms into the field of Human-Computer Interface. On the other hand, by combining the visual channel to the acoustic (audio) channel for speech recognition, the resulting bimodal speech recognizer is shown to be markedly more robust, when it compared to the only acoustic counterpart [1], [2], [8].

Most of lip-reading systems consist of two stages; the first stage is a phoneme feature extraction stage and applied to each frame of the video images sequence. The second stage inputs the feature vector sequence from the first stage and recognizes the whole of a word or a sentence. Regarding to first lip-reading stage there are two main approaches for phoneme feature extraction:

1. Geometric-feature based; and
2. Appearance based.

Geometric-feature (or lip contour) based approach obtains information from the lip's height or width or color or shape or all of them, or may obtain information from other face's features such as pupils and nostril [3], [4]. In the appearance (or image transformation) based approach, the features are depending on the intensity values of the image pixels that include the lips, and they can be used directly or after some image transform [5], [6]. Due to the data's reduction involved in the first approach, considerable amount of features are lost. In contrast, the latter approach uses the entire of available features. Another advantage of the latter approach is that important features can be represented in a low-dimensional space and can often be made invariant to image transforms like translation, scaling, rotation and lighting. However, the disadvantage of the latter approach is that it needs a great amount of training data, which may not be available in some cases. Some efforts are done to combine the two approaches together or with other techniques to overcome the above disadvantages [7], [8].

In fact, conventional lip-reading systems use different techniques to implement the phoneme feature extraction on the image. Among of these techniques is the discrete cosine transform (DCT) [9], or the discrete wavelet transform (DWT) [5]. Any what the technique is, in our thinking, the phoneme feature extraction module is required to have the following three characteristics:

1. It should make a parametric feature space with low dimensionality.
2. Distributions of each phoneme in the feature space should be simple and approximated by the normal distributions.
3. Execution time of feature parameter extraction should be in the range of video camera rate.

In the present work, we propose a new visual feature representation (lip-reading) module combines Hypercolumn neural network Model (HCM) [11], with Hidden Markov Model (HMM). Our proposed lip-reading system considered as an example to the appearance based approach. The advantage of our system is that we use the input image directly without the need to a lot of lips model or lips marker. In addition, the system capable of using the entire image features and extracts all the relevant features. To this end, we use HCM in the first stage of lip-reading system to extract the relevant phoneme features. Using HCM for this task will satisfy

the above first two characteristics; in addition, a randomized technique (RHCM) [12] is used to satisfy the third one. HMM is used in the second lip-reading stage for feature sequence recognition. The proposed system may work under varying lips positions. In order to check our system's reliability we apply it for Arabic language without any pre-linguistic. This is the first time that a visual representation system is applied for Arabic language. For this reason we compare Arabic language recognition results with those for Japanese language.

The paper is organized as follows: Basics and characteristics of HCM model are covered in section 2. In section 3 an overview of HMM is given. Section 4 shows the structure of our lip-reading system. Database and experimental results are given in section 5. Paper conclusion and future work are given in section 6.

2. Hypercolumn Model (HCM)

The Hypercolumn model is a competitive neural network model. Mainly it designed as a pyramidal piling up hierarchical layers derived from the Neocognitron neural network (NC) model [13] by replacing each NC's unit cell plane with a Hierarchical Self Organizing Map (HSOM) [14]. HCM model able to encode gray level images, and in the same time, each layer's codebook is generated automatically. Furthermore, the output of each layer is feed directly to the following layer without any data manipulations between layers. HCM, as we see in figure 1, has two intra layers and one output layer.

2.1 HCM Intra-Layer's Structure

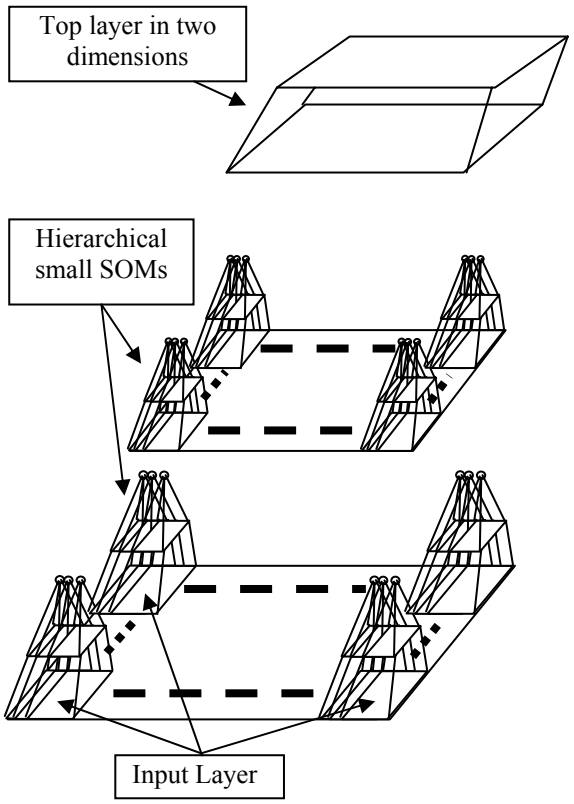
As indicated in figure (1), each HCM intra-layer has number of overlapped HSOMs. Each HSOM consists of two hierarchical SOM layers. The lower SOM layer is called feature extraction layer, since it performs features extraction by quantizing the input space and mapping it into a neuron array of low dimensional. The upper SOM layer is called feature integration layer since it inputs the winner neuron index from the first layer to perform optimum feature integration and then choose the winner neuron. Moreover, the upper SOM layer enables shift, rotate, scale and distorts invariant recognition by a similar way as in Neocognitron.

The overlapping occurred between HSOMs helps in covering all lips positions in case of lip's shift or rotates, and then enables to obtain multiple classifications for every input data. In addition, each HSOM should decide its output winner after shifting its input field in X and Y dimensions to cover different groups of neighbor pixels. Then it selects one winner for each input position such that the winner of all winners is considered as the unit's output. Finally, the array of all winner position indexes from all HSOMs in each intra layer forms the final output of that HCM layer and, in the same time, the input of the following HCM layer.

2.2 HCM Output Layer's Structure

It is clear from figure (1) that the size of the hierarchical layers in HCM model decreases as the layer index increases. The HCM output layer (top layer) is selected as only one SOM, and it supposed that this layer integrate all information from the input patterns. The output layer's feature map neuron is in two dimensions not in one dimension as the other SOMs in the lower layers. Clearly, the idea of two dimensions means higher number of neurons, which will be more convenient not only if we have a high dimensional learning data but also if we need to achieve a finer-grained classification of the learning data. Finally, due to the competitive learning nature of HCM, only one cell in the output layer will be activated in correspondence to the category of the input pattern, other cells respond to other categories.

Figure (1): HCM Model



2.3 HCM Algorithm

HCM uses an unsupervised learning algorithm to construct its layer's feature maps. As it mentioned before that only one neuron *c* from each HSOM network will activated according to the following rule:

$$\|I - W_c\| = \min_u(\|I - W_u\|) \qquad (1)$$

where W_u is the neuron weight vector. Strictly speaking, the winner neuron *c* is the neuron that has the nearest weight vector to the input data *I*. So in the learning phase, each time a training data item is input, the winner is selected according to equation 1. The weight vectors are updated according to:

$$W_u(t+1) = W_u(t) + h_{cu}[I(t) - W_u(t)] \qquad (2)$$

$$h_{cu} = \alpha(t).\exp\left(\frac{\|r_c - r_u\|}{2\sigma^2(t)}\right) \quad (3)$$

where $\alpha(t)$ is the learning rate and is decreased gradually toward zero and $\sigma^2(t)$ is a factor used to control the neighborhood range. The term $\|r_c - r_u\|$ is referring to the distance between the neuron *c* and neuron *u*. We conclude that not only the winner neuron but also the neighborhood neurons of it are trained simultaneously in every iteration phase.

The learning process line of HCM is applied beginning from bottom to top, layer by layer, where the standard learning algorithm of HSOM [14] is used to train each HSOM's unit neurons. The learning process of the first and second layers of each HSOM network is carried out as it is explained in table 1 and table 2, respectively.

Table 1: 1st HSOM Layer Learning Algorithm

- Initial i=0;
- While (i=Num of Iterations) do {
  1- For the first layer, initialize the neurons weights randomly. In addition, initialize both of neighborhood range and learning rate parameters
  2- For all input positions (considering the shift positions) apply the competitions according to equation (1) and select the winner neuron.
  3- Update the codebook of the winner and its neighbors as in equations (2) and (3).
  4- i++ }

Table 2: 2nd HSOM Layer Learning Algorithm

- Initial j=0;
- While (j=Num of Iterations) do {
  1- For the second layer, consider the input of this layer as the winner position index of the first layer, apply the competition as in equation (1) and select the winner.
  2- Update the codebook of the winner and its neighbors as in equations (2) and (3).
  3- j++ }

Finally, to reduce the recognition time to the range of video camera rate, a randomization technique (RHCM) [12] is applied in the competition process.

2.3 HCM Model's Advantages

HCM capable of generating ordered mappings of the input data onto low-dimensional topological structure, which is very useful in analyzing high-dimensional data. Another capability is the partitioning of input data into clusters in such a way that gathering the similar data items into one cluster. This is possible as HCM preserving the topological orders of the input data items, and keeping close in the output space those data items which are closer in the input space by using any distance measure function. Other HCM's advantages can be summarized as follows:

1. HCM can recognize invariant objects with variations in position and size.
2. Accept random initialization for the network weights.
3. No preprocessing for the input images is needed.
4. HCM enables rotate and distort invariant objects recognition.

3. Hidden Markov Model (HMM)

HMM may define as a variable-size collection of random variables with an appropriate set of conditional independence properties. It is known that, HMM is widely used in many practical applications such as DNA, pattern and speech recognition. Among the various techniques used for visual speech recognition studied so far, the HMM holds the greatest promise due to its capabilities in handling either the variability or the sequence of speech signals

3.1 HMM specifics

Our system's second stage uses HMM to recognize each sentence. Generally, in the recognition process using HMM, there are two main phases: training phase and testing (recognition) phase. In the training phase a training set of features and their associated transcriptions for each sentence are used to derive a set of features reference models. In the testing phase, unknown sentences are transcribed and then the probability of each model generating that sentence is computed.

Let us consider that each sentence is represented by a feature vector sequence as; $O = o_1, o_2, o_3, \ldots, o_T$, where o_t is the feature vector at time t. So, the lip-reading process can be regarded as the computation process of:

$$\arg\max_i \{p(\omega_i | O)\} \quad (4)$$

where ω_i is the ith target sentence. We can express about the probability included in (4) using Bayes's rule:

$$P(\omega_i | O) = P(O | \omega_i) P(\omega_i) / P(O) \quad (5)$$

So, for a given set of $P(\omega_i)$, the most probable spoken sentence depends only on the likelihood $P(O | \omega_i)$. To calculate this likelihood, it is assumed that the feature vector sequence is generated by a Markov model corresponding to each sentence, where Markov model is a finite state machine changes state once every time unit. The transition from state i to state j is governed by the discrete probability a_{ij}. At each time *t* when a state j is entered, a feature vector o_t is generated from the probability density $b_j(o_t)$. Then, the likelihood $P(O, X|M)$ for a state sequence $X = \{x(1), x(2), \ldots\ldots, x(T)\}$ of a model *M* is estimated by

$$P(O,X|M) = \sum_{X} a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \quad (6)$$

Supposing that X is unknown, the required likelihood $P(O \mid \omega_i)$ is estimated by summing over all possible state sequences, that is

$$P(O|M) = \sum_{X} a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \quad (7)$$

Given a set M_i corresponding to sentences or words w_i, equation 4 is solved by using equation 5 and by supposing that:

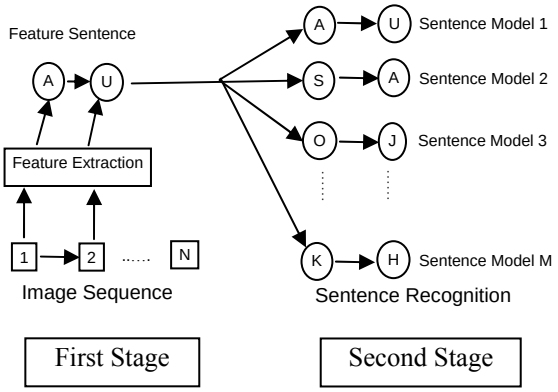
$$P(O|\omega_i) = P(O|M_i) \quad (8)$$

A good overview of HMM and its using in speech recognition is in [15].

4. A Lip-Reading System

Automatic recognition of speech using the video sequence of the speaker's lips has attracted significant interest, especially for those who need to combine the visual channel with the audio channel to produce a robust speech recognition system. It is well known that human speech perception is enhanced by seeing the speaker's lips providing useful information to the receiver. In other words, the possibility of communicating with computers through lip-reading presents an opportunity to change profoundly the way humans, especially handicapped persons, interact with machine.

Figure (2): Overview of Lip-reading Systems



Most of lip-reading systems consist of two stages, the first stage is a phoneme feature extraction and it is applied for each frame of the video image sequence. The second stage inputs the feature vector sequence from the first stage and recognizes the whole of a word (or sentence). Figure (2) shows the Lip-Reading system's stages. In our case, the first stage is achieved by using HCM neural network model. Neural network stands as one of the leading techniques that used in the modeling process of audio and visual speech processing especially when combined with HMM [16-17].

4.2 HCM Structure

The HCM network, in this paper, structured from three hierarchical layers. According to previous architecture of HCM in section 2, first HCM layer includes 25 x 18 HSOMs, while second layer includes 10 x 6 HSOMs. Through competition phase, neurons are competing 3 times in a 3 different positions in X and Y dimension with shift step 2 times in each dimension for the first HCM layer. For the second HCM layer, the neurons are competing 2 times in 2 different positions with shift step 1 in each dimension. Number of neurons of the top layer is 64 neurons in two dimensions.

4.3 Features Extraction (First Stage)

The first stage of our lip-reading system, and according to the previous representation of each sentence in equations 4-8, is concerned with the output probabilities:

$$\prod_{t=1}^{T} b_{x(t)}(o_t)$$

of equation 7. In the HMM training process, which is well known as Baum-Welch algorithm, the output probabilities are represented by Gaussian mixture densities. Therefore, the performance of the lip-reading system depends on both the dimensionality of the feature space and whether the output distributions are simple or not in the feature space to be approximated by Gaussian densities. From this point of view, the requirements 1 and 2 in section 1 are satisfied.

4.4 Features Recognition (Second Stage)

In the second stage (recognition stage), we use continuous density HMMs as a mean of statistical pattern matching for every sentence. For each sentence we build an HMM with 5 states, each state having an observation probability distribution modeled by a Gaussian probability. We train each HMM using the HTK toolkit [18].

As we explained above, there are two phases through this stage, training phase and recognition (or testing) phase. In the first phase, each training image is quantized into discrete sequence of phonemes. A training set of features and their associated transcriptions, for each sentence (or word), are used to estimate the parameters of HMMs in order to derive a set of reference models. In the other phase, unknown features (or new user) are transcribed using specific HTK tools and then the probability of each model generating that sentence is computed. Finally the most likely model identifies the sentence.

5. Experiments

5.1 Sentences Database

Arabic language is considered as the sixth most widely spoken language in the world, the estimated number of Arabic native speakers around 300 millions over the world. Despite this fact there is no remarkable research on

the trend of visual features representation based Arabic language. In the current experiments, we use 9 different sentences including 26 words; most sentences have 3 words, uttered by 9 native Arabian speakers. The only contribution in the trend of Arabic Speech recognition has been tested only by small data set (Arabic digits) [19]. Of course handling full sentence has a higher challenge. The uttered sentences and their translation to English are in table 3.

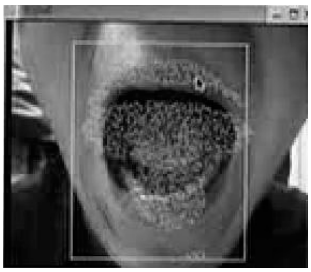
Table 3: Arabic Sentences

English Meaning	Arabic sentence
1- A pain in my teeth	- الم فى ضروسى
2- A headache in head	- صداع فى راسى
3- A swelling in my back	- ورم فى ظهرى
4- A pain in my gum	- الم فى لثتى
5- The Arabic Salutation	- سلام عليكم
6- A swelling in my leg	- ورم فى قدمى
7- A pain in my back	- الم فى ظهرى
8- A swelling in my tooth	- ورم فى ضرسى
9- A pain in my head	- الم فى راسى

5.2 Image Database

The image database set includes 5760 gray images captured directly by using EVI-G20 Sony camera where the size of each input image is 160x120 pixels. We divided the image data set into two subgroups. First group, *training group*, has 4320 gray image gathered from 6 different Arabian adults (male & female), while the second group, *test group*, has 2160 gray images gathered from 3 persons, one of them belongs to the training group. Some persons in both groups have moustaches or beards or both. Each uttered sentence represented by 80 visual frames. Figure (3) shows a snapshot of the online system.

Figure (3): A Snapshot of the Online System



5.3 Experimental Results & Analysis

Introducing the Arabic language to our system gives the recognition accuracy results summarized in table 4 for both word unit and full sentence unit. In the current experiment, we used a single Gaussian density to approximate the output distributions. Our results in table 4 are in case of using a full rank covariance matrix for each HMM state. In table 4, we remark that there is a difference between the sentence accuracy result and the word accuracy result. This is due to the nature of the Arabic letters when they are used in speech.

Table 4: Arabic Recognition Accuracy

Target type	Learning data	Test data
Word	82.1	79.5
Sentence	63.0	62.9

As an example, when someone vocalizes specific separate Arabic letters the mouth will take the same shape (or appearance) for each letter. In our case, this will let the system in a confusion status especially in case of sentence recognition. Furthermore, we observe that there is a dwindling occurred for the test data accuracies. This may due to one or both of the following:

1. The estimated covariance was too small.
2. The number of learning data was not enough.

To overcome the above-mentioned drawbacks we suggest to using large learning data set. On the other hand, since there are no pre-research in the field of visual speech recognition for Arabic language, we compare the above results with those for the Japanese language. Table 5 shows the recognition accuracy results for the Japanese language under same experimental conditions [10].

Table 5: Japanese Recognition Accuracy

Target type	Learning data	Test data
Word	92.6	83.3
Sentence	88.9	70.7

It is clear that Japanese language performance is better than Arabic language performance. This may justify by the nature of both languages. For example, Japanese depends by nature on vowels sequence of sentence, which considered as an aid to recognize the Japanese characters than Arabic that has not the same merit. However, regarding to the novelty of research of visual speech processing for Arabic sentences, our results consider a good start to continue in this trend.

6. Conclusion & Future Work

In this paper we introduced a new lip-reading system. The proposed system is a combination of HCM neural network with HMM and applied to Arabic language. According to our knowledge, this is the first time a lip-reading system is applied to Arabic language. Even though, the system shows a good accuracy in case of Arabic word recognition especially for learning data, some dwindling happened in the Arabic sentence recognition level. To overcome a part of the sentence accuracy dwindling, we intend to increase the learning data set. In addition, it is supposed that the Hidden Markov Model is designed mainly to help in the process of speech recognition; more efforts have to do to use it in Arabic lip-reading systems, at least on the level of phoneme recognition part. For example, using an optimum number of HMM states per Arabic phoneme may be a guarantee for result's improvement.

References

[1] A.J. Goldschen, O.N. Garcia, & E. Petajan, Continuous optical automatic speech recognition by lipreading. *Proc. 28th Annual Asilomar conf. on Signal System, and Computer*, vol. 1, 1994, 572-577.

[2] T. Chen, & R. R. Rao, "Audio-visual integration in multimodal communication," *Proc. IEEE*, 86, 1998, 837–852.

[3] R. Stiefelhagen, U. Meier, & J. Yang, Real-Time Lip-Tracking for Lipreading. *Proc. of Eurospeech* 97, 1997.

[4] T. Chen, Audiovisual Speech Processing, Lip Reading and Lip Synchronization. IEEE Signal Processing Magazine, Jan. 2001, 9-21.

[5] G. Potamianos, H. P. Graf, & E. Cosatto, An Image Transform Approach for HMM Based Automatic Lip Reading. *Proc. of the Int. Conf. on Image Processing, 3*, 1998, 173-177.

[6] P. L. Silsbee,& A. C. Bovik, Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Trans. Speech Audio Processing, 4* (5), 1996, 337–351.

[7] J. Luettin, & N. Thacker, Speechreading using Probabilistic Models. *Computer Vision and Image Understanding, 65* (2), 1997, 163-178.

[8] M. T. Chan, HMM-Based Audio-Visual Speech Recognition integrating Geometric-and Appearance-Based Visual Features. *Proc. of IEEE Workshop on Multimedia Signal Processing,* 2001, 9-14.

[9] S. Deligne, G. Potamianos, & C. Neti, Audio-visual Speech Enhancement with AVCDCN. *Int. Conf. on Spoken Language Processing,* 2002, 1449-1452.

[10] N. Tsuruta, H. Iuchi, A. Sagheer, & T. Tobely, Self–Organizing Feature Maps for HMM Based Lip-reading. *Proc. of 7th Int. conf. on Knowledge-Based Intelligent Information & Engineering Systems, KES2003, II*, 2003, 162-168.

[11] N. Tsuruta, R. Taniguchi, & M. Amamiya, Hypercolumn Model: A Combination Model Hierarchical Self- Organizing Maps and Neocognitronfor Image Recognition. *System and Computer in Japan, 31*(2), 2000.

[12] T. Tobely, N. Tsuruta, & M. Amamiya, A randomized Model for the Hypercolumn Neural Network for Gesture Recognition. *Intentional J. of Comp., Sys, and Signals,3* (1), 2002, 14-18.

[13] K. Fukushima, Neocognitron. *Biol Cyberetics, vol. 36* (4), 1980, 193-202.

[14] J. Lampinen, & E. Oja, Clustering Properties of Hierarchical Self Organizing Maps. *Journal of Math. Imaging and Vision, 2*, 1992.

[15] L.R. Rabiner, A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *IEEE Proc., 77* (2), 1989, 257-286.

[16] B. P. Yuhas, M. H. Goldstein, & T. J. Sejnowski,, Integration of acoustic and visual speech signals using neural networks. *IEEE Communication Mag., 27*, 1989, 65-71.

[17] G. I. Chiou, & J.-N. Hwang, Image sequence classification using a neural network based active contour model and a hidden Markov model. *Proc. IEEE International Conf. Image Processing, ICIP-94., 3*, 1994, 13-16.

[18] HMM Tool Kit (http://htk.eng.cam.ac.uk).

[19] H. Bahi, M. Sellami, Combination of Vector Quantization and Hidden Marcov Models for Arabic Speech Recognition, *Proc. of ACS/IEEE Int. Conf. on Computer Systems and Applications (AICCSA'01)*, 2001.