Combination of Hypercolumn Neural Networks Model with Hidden Markov Model Based Lip-reading for Arabic Language

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

Maeda, Sakashi
Department of Electronics Engineering and Computer Science, Fukuoka University

出版情報：Proceedings of International Workshop on Fuzzy Systems & Innovational Computing. 2004, pp.231-235, 2004-06
バージョン：
権利関係：

Combination of Hypercolumn Neural Networks Model with Hidden Markov Model Based Lip-reading for Arabic Language

Alaa El.Sagheer*, Naoyuki Tsuruta**, Rin-ichiro Taniguchi*, Sakashi Maeda**

* Kyushu University
Department of Intelligent Systems,
6-1 Kasuga-Koean, Kasuga,
Fukuoka808-0196
JAPAN
{alaa, rin}@limu.is.kyushu-u.ac.jp

**Fukuoka University
Department of Electronics Engineering
and Computer Science,
8-19-1,Nanakuma, Jonan,
Fukuoka 814-0180
JAPAN
{tsuruta, maeda} @tl.fukuoka-u.ac.jp

Abstract: In recent years *Lip-reading* systems have received much attention. Because it plays an important role in human communication with computer, namely it can easily smooth the human-computer communication and lets it closer to human-human communication. The importance of lip-reading becomes clearer when the communication environment is not so suitable for speech perception such as among persons who have permanent hearing problems, or elderly people, or under water activities, or generally in any ill conditioned environment. In this paper, we supposed that a combination of the Hypercolumn Neural Network model with the Hidden Markov Model is used to achieve a lip-reading system for Arabic language. This lip-reading system may work under varying lips positions and sizes. Experiments include different nine Arabic sentences gathered from 8 different Arabian people (Male & Female). Results show the performance of our system based Arabic Language.

1. Introduction

Human-Computer Interface (HCI) systems have undergone many advancement over the last few years as the computer's role in our lives is becoming main and vital; *Audio-Visual Dialogue* is the backbone of the interface with computer. *Lip-reading* is one of the most fertile topics of Audio-Visual dialogue research, since it can smooth the Human-Computer Interface by introducing the Human-Human interaction mechanisms into the field of Human-Computer Interface (HCI). Most of lip-reading system consists 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 sentence.

Conventional lip-reading systems have two types of phoneme feature extraction methods. First type is using the Discrete Cosine Transform (DCT) [1], the other type is specified to feature extraction such as shape of lips [2]. The phoneme feature extraction module is required to have the following three characteristics [3]:

1 It should make a parameterization feature space of 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 in the range of video camera rate.

In this paper, we will use the *Hypercolumn Neural Network Model* HCM [4] in the first stage of our lip-reading system to satisfy the above first two characteristics. In addition, a randomized technique [5] will be used to satisfy the third characteristic. *The Hidden Markov Model* (HMM) is used in the second lip-reading stage for feature sequence recognition.

Our lip-reading system may work under either varying lips positions or under different kinds of lips sizes (female or male). The main task of this system is to build a successful template-based HCM feature map; this feature map is constructed mainly from different and many learning images in order to matching the underlying variations in the lips posture. The proposed lip-reading system has performed well when applied to the Japanese language and the three above merits are satisfied [3], but we will concern within

this paper with the Arabic language.

Arabic language is counted as the sixth most widely spoken language in the world- the estimated number of Arabic native speakers is more than 250 millions. Moreover, Arabic language is the language of the "Kora'an" the holly book of more than 1000 million Muslims all over the world. For all these we can appreciate the need to involve the Arabic language in the lip-reading research applications. Despite the above facts there is no remarkable research on the trend of Arabic lip-reading system, only little bit contributions on the field of Arabic characters recognition. This research is not neither proportional with the importance of Arabic language nor enough compared with other languages of similar importance such as Spanish. Most research on the Arabic language recognition field is focused on text recognition [6, 7], speech to speech recognition [8] and speech recognition [9], but there is no bright contribution on the Audio-Visual dialogue trend. For all these reasons we will apply our lip-reading system on the Arabic language, which may consider as beginning of the applications on the lip-reading systems based Arabic language. The paper is organized as follows: an outline for both the Hypercolumn neural network (HCM) model and the Hidden Markov Model is given in section 2 and 3 respectively. In section 4, we show an overview of the skeleton of the Arabic language. Section 5 is devoted to the way of how our system deals with Arabic language. Experimental results and their analysis with a comparison with Japanese language are given in section 6. Section7 conclude the task of this paper.

2. Hypercolumn Neural Network Model (HCM)

The HCM is an unsupervised neural network model [4]. HCM is composed mainly by a pyramidal piling up hierarchical layers derived from the Neocognitron Neural Network (NC) model by replacing the NC's unit cell planes with two Self Organizing Map (SOM) layers, i.e. Hierarchical SOM (HSOM), as indicated in figure 1. These two SOM layers (HSOM) allow features extraction through the first one of them and feature integration through second one. Second layer inputs the index of winner neuron from the first layer. Combining the structure of NC with SOM lets the HCM model combines the advantages of NC and SOM neural networks too. The HCM model merits can be summarized as follows: 1. HCM can recognize invariant objects with variations in positions and sizes. 2. Accept random initialization for the network weights. 3. No preprocessing for the input images is needed. 4. HCM enables rotate, scale and distortion invariant objects recognition.

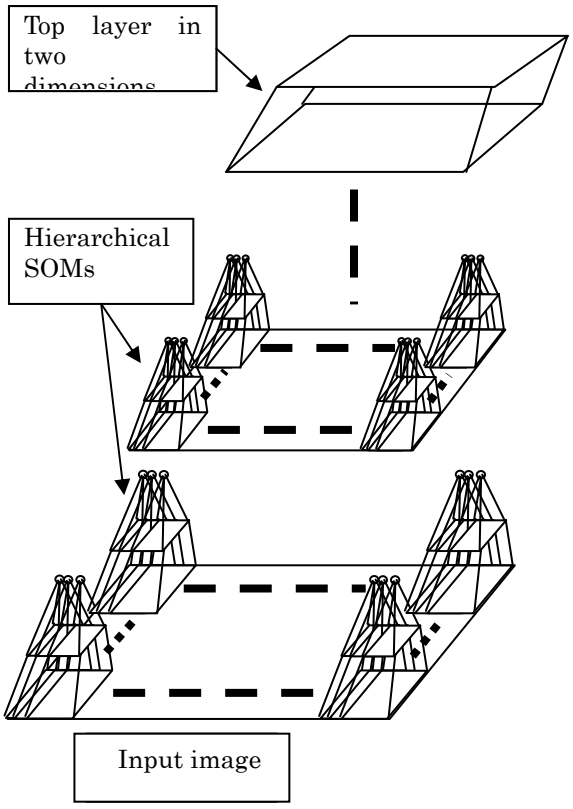

Figure 1 HCM model

As indicated in figure 1, the size of the hierarchical layers in HCM model decreases as the layer index increases. The HCM top-layer is selected as only one SOM, its feature map neurons is in two dimensions not in one dimension as the other SOMs in the lower layers. Since two dimensions means higher number of neurons which will be more convenient not only if we have a high dimensional learning data (images) but also if we need to achieve a finer-grained classification of the learning data.

HCM uses unsupervised learning algorithm to construct its layer's feature maps. The learning process line of HCM is applied beginning from bottom to top layer by layer, where the standard learning algorithm of HSOM is used to train each HSOM's unit neurons. To introduce the shift invariant recognition only one feature map for each HSOM is trained in spite of the shifts

happened in input positions. The drawback of HCM is its execution time may still long. So a randomization technique, and then it will denoted by (RHCM), may apply to the HCM model directly to reduce the recognition time to the range of video camera rate [5]. To do that two main parameters PUR (Pixel Usage Ratio) and NR (Neighborhood Range) should be small, but this sometimes is not satisfied.

3. Hidden Markov Model (HMM)

3.1 HMM definition

HMM is widely used in many practical applications such as pattern recognition, DNA and especially in speech recognition because their ability to handle either the variability or the sequent of the speech signals. HMM may define as a variable-size collection of random variables with an appropriate set of conditional independence properties.

3.2 HMM specifics

Our system uses a continuous density HMMs as a means of statistical pattern matching for every word or sentence. In the recognition process using HMM there are two main phases: the training phase and the recognition phase. In the first phase a training set of features is used to derive a set of reference models. In the other phase the probability of generating the test observation is computed for each reference model of the first phase. Namely, if we consider each spoken sentence (or word) be represented by a feature vector sequence as

$$O = \{ o_1, o_2, o_3, \ldots, o_T \},$$

So the lip-reading can be regarded as the computation process of

$$\arg\max_i \{ P(w_i \mid O) \},$$

where w_i is the i th target sentence or word. According to Bayes's rule:

$$P(w_i \mid O) = P(O \mid w_i)\, P(w_i)/P(O), \qquad (1)$$

the most probable spoken sentence depends only on the likelihood $P(O \mid w_i)$.

In HMM based approach, 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}. Furthermore, 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 $P(O, X \mid M)$ for a state sequence $X = \{x(1), x(2), \ldots\ldots, x(T)\}$ of a model M is estimated by

$$P(O, X \mid M) = \prod_{t=1}^{T} b_{x(t)}(o_t)\, a_{x(t)x(t+1)} = \prod_{t=1}^{T} b_{x(t)}(o_t) \prod_{t=1}^{T} a_{x(t)x(t+1)}, \qquad (2)$$

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

$$P(O \mid w_i) = P(O \mid M) = \sum_X P(O, X \mid M), \qquad (3)$$

A good introduction of HMM models and its using in speech recognition is in [10].

4. Arabic Language

Arabic language considered as the sixth most widely spoken language in the world. Even though there are no any clearer pre-efforts to achieve an Arabic lip-reading system parallel to the need of the Arabic people to such system. So, in this paper, we apply our lip-reading system on the Arabic language for the first time. By the way our lip-reading system has applied before on the Japanese language and the given results proved that our system's performance is well on the Japanese language [3].

4.1 Knowledge about Arabic Language

Arabic language is a cursive language written from right to left. It has 28 letters (25 consonants and 3 vowels) each of them has 2-4 written representations depending on the position. There are two main types of spoken (or written) Arabic:

1. Modern Standard Arabic MSA- the universal language of the Arabic speaking world that is understood by all Arabic speakers. It is the language of the vast majority of written materials and of formal TV shows, lectures, etc
2. Classical Arabic - the language of the "Kora'an" and classical literature. It differs from MSA mainly in style.

The Arabic language have three long vowels /a:/, /i:/ and /u:/ and three short vowels which are not usually marked, except in classical

Arabic, and never used in the MSA type. Our experiments implemented on the MSA type.

4. The HMM Based Lip-reading System

Most of lip-reading system consists of two stages, figure 2; the first stage is a phoneme feature extraction and it is applied to 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. Our system achieves the first stage by using HCM concerning with the output probability $\prod_{t=1}^{T} b_{x(t)}(o_t)$ of equation (2) above. The second stage achieved by using the HMM.

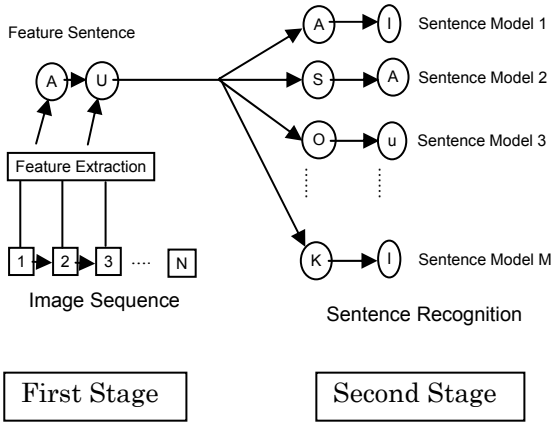

Figure 2 Lip-reading System

4.1 Lip-reading Database

In our project, we implemented the system on 9 different Arabic sentences where most sentences consist of three words and, at a first time, we interest ourselves with the condition that each sentence must have a number of the Arabic vowels beside, of course, the Arabic consonants. Definitely, our system's word list includes the 3 vowels and 14 consonants from the 25 consonants. Depending on this way, it is possible to recognize the spoken Arabic sentences through the sequence of vowels in each sentence. The learning images are captured directly by using EVI-G20 Sony camera where the size of each image is 160x120 pixels. The images database is divided into two groups. First group, *training group*, has 4320 gray images gathered from 6 different Arabian persons (male & female), while second group, *test group*, has 2160 gray images gathered from 3 Arabian persons (only male), one of them belongs to the training group. Each spoken sentence represented by 80 visual frames, this big number of frames due to the natural long of Arabic letters pronunciation. Table 1 shows some of the Arabic used sentences and their English meanings.

Table 1 Samples of used sentences

English Meaning	Arabic sentence
1- A pain in my back	1- الم فى ظهرى
2- A headache in head	2- صداع فى راسى
3- A swelling in my leg	3- ورم فى قدمى
4- A pain in my gum	4- الم فى لثتى
5- Hello or good bye (the Arabic Salutation)	5- سلام عليكم

4.2 HCM Structure

The HCM network structured from three hierarchical layers. Each layer consists from a number of HSOM neural networks, except the top layer which consists from only one- two dimensional- SOM neural network instead. Namely, the first layer includes 25 x 18 HSOMs, and the second layer includes 10 x 6 HSOMs. In the competition phase, for the first layer, each HSOM neurons is compete 3 times in a 3 different positions in X and Y dimension with shift steps 2 in each dimension, while in the second layer the competition is done 2 times in 2 different positions with shift step 1 in each dimension. Number of neurons in the top layer is 8x8 neurons.

4.3 Lip-reading System

Five states HMMs were considered for each phoneme and the all were learned by using the HTK [11]. In the HMM's training phase, a training set of features is used where each training image will quantized into discrete sequence of vowels in order to derive a set of reference models. Then each group of winners of the same category will treated as a vowel state. In the HMM's recognition phase the sequence of these states will compared with the reference model of each sentence. The probability of generating the test observation is computed for each reference model. Finally the features are organized as the sentence (or word) whose model gives the highest accuracy rate.

5. Results and Analysis

In this section, experimental results of applying our lip-reading system on the Arabic language are shown. The accuracy in the case of using a full rank covariance matrix for each state is shown in table 2.

Table 2 Recognition accuracy for Arabic

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

Since there is no pre-research in the topic of lip-reading based Arabic language we will compare the above results with those we already got for the Japanese language. Table 3 shows the results of our system using the Japanese language [3]:

Table 3 Recognition accuracy for Japanese

Target type	*Learning data*	*Test data*
Word	95.4	83.3
Sentence	94.4	74.1

This dwindling occurred in Arabic's results than Japanese's results may justify by the nature of both languages. For example in Japanese, all consonants, except for a special consonant, are combined with one of 5 vowels. Compared with this, since most of Arabic consonants are independent by themselves, the above phenomenon is not naturally happened in Arabic. If the case like that, it is easier for our system to get higher recognition accuracy for Japanese than Arabic languages, especially, regarding to the scarceness of vowels in Arabic than Japanese. In addition, there is a similarity in vocalization (or phonation) of some Arabic letters lets the system in a confusion status especially in case of sentence's recognition. Furthermore, the Arabic language differentiate than other language by it's including some letters are not available in any other language, such as the letters ض (DAAD) and غ (GHAIN) which led to more confusions.

6. Conclusion

In this paper, a lip-reading system for the Arabic language, for the first time, was introduced. The proposed system is a combination of HCM neural network with the HMM. Even though Hidden Markov Model HMM 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] S. Deligne, G. Potamianos, C. Neti,: "Audio-visual Speech Enhancement with AVCDCN", Int. Conf. on Spoken Language Processing, pp.1449-1452,2002.

[2] U. Meier, R. Stiefelhagen, J. Yang, A. Waibel,: "Towards Unrestricted Lip-reading", In Second International Conference on Multimodal Interface (ICMI99), 1999.

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

[4] N. Tsuruta, R. Taniguchi, M. Amamiya,: "Hypercolumn Model: A Modified Model of Neocognitron Using Hierarchical Self Organizing Maps", in the Intr. Work Conf. on ANN, IWANN99, 1999.

[5] T. Tobely, N. Tsuruta, M. Amamiya,:A randomized Model for the Hypercolumn Neural Network for Gesture Recognition, Int. J. of Comp., Sys, and Signals, vol. 3, No.1 ,pp. 14-18, 2002.

[6] M. Altuwajiri, M. Bayoumi,: "A Thining Algorithm for Arabic Characters Using ART2 Neural Network", IEEE Transactions On Circuits And Systems-II: Analog and Digital Signal Processing, Vol. 45, No. 2, Feb., 1998.

[7]Y. A. El-Imam, "Phonetization of Arabic: rules and algorithms" J. of Computer Speech and Language, available on line at www.scincedirect.com

[8] K. Kirchhoff et. All, "Novel Approach to Arabic Speech Recognition" Report from the 2002 JOHN-HOPKINS Summer Workshop.

[9] H. Bahi, M. Sellami,: "Combination of Vector Quantization and Hidden Marcov Models for Arabic Speech Recognition" ACS/IEEE International Conference on Computer Systems and Applications (AICCSA'01), 2001.

[10] L.R. Rabiner, "A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceeding of the IEEE, Vol. 77, No. 2, pp. 257-286, Feb. 1989.

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