九州大学学術情報リポジトリ
Kyushu University Institutional Repository

Japanese Lip-Reading System

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

https://hdl.handle.net/2324/5861

出版情報：IEICE technical report. Neurocomputing. 104 (502), pp.31-36, 2004-12. 電子情報通信学会
バージョン：
権利関係：

社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE

日本語リップリーディングシステム

アラー サーギル† 鶴田 直之‡ 谷口 倫一郎†

†九州大学大学院 システム情報科学府 知能システム学専攻

〒816-8580 福岡県春日市春日公園 6-1

‡福岡大学 工学部 電子情報工学科

〒814-0180 福岡県福岡市城南区七隈 8-19-1

E-mail: †{alaa, rin}@limu.is.kyushu-u.ac.jp, ‡tsuruta@tl.fukuoka-u.ac.jp

あらまし リップリーディングは人間とコンピュータのインタフェースとして有用な手段の１つである．本稿では，我々が従来から研究を進めているＨＣＭ (Hypercolumn neural network Model)と隠れマルコフモデルを用いたリップリーディングシステムについて述べる．我々が提案するシステムではＨＣＭを用いて画像から特徴を抽出し，隠れマルコフモデルを用いて特徴系列の認識を行う．ＨＣＭは位置不変，サイズ不変の特徴抽出を実現できるので，本システムでは対象を撮影する位置を固定せずに，認識を行うことができる．本システムの有効性を確認するために，日本語の文章を対象としたリップリーディングの実験を行った．対象画像は顔をクローズアップしたモノクロ画像であり，比較のため，特徴抽出にＳＯＭ（自己組織化マップ）とＤＣＴ（離散コサイン変換）を用いたシステムの認識性能も評価した．実験から，ＨＣＭがＳＯＭやＤＣＴよりも画像特徴を的確に捉えており，認識性能が優れていることが示された．

キーワード リップリーディング，特徴抽出，ハイパーコラムモデル，自己組織化マップ，隠れマルコフモデル

Japanese Lip-Reading System

Alaa Sagheer† Naoyuki Tsuruta ‡ and Rin-Ichiro Taniguchi†

† Department of Intelligent Systems, Kyushu University

6-1, Kasuga-Koen, Kasuga, Fukuoka 816-8580, Japan

‡ Department of Electronics Engineering and Computer Science, Fukuoka University

8-19-1, Nanakuma, Jonan, Fukuoka 814-0180, Japan

E-mail: †{alaa, rin}@limu.is.kyushu-u.ac.jp, ‡tsuruta@tl.fukuoka-u.ac.jp

Abstract Lip-reading is one of the most fertile topics of interface with computer, since it can smooth the Human-Computer Interface by introducing Human-Human interaction mechanism. In this talk, we introduce a novel Japanese lip-reading system combines our group's model, Hypercolumn neural network Model (HCM), with Hidden Markov Model (HMM). HCM is used to extract the visual speech features while HMM is used for recognition. The proposed lip-reading system can work under varying lip positions and sizes. Our experiments were carried out using multiple sentences of Japanese language. All images were captured in a natural environment without using a special lighting or lip markers. Experimental results are shown to compare favourably with the results of two reported approaches: Self Organizing Map (SOM) using same database set and Discrete Cosine Transform (DCT) using different database set. HCM provides better performance than both approaches. This demonstrates that HCM can extract and classify features in a better manner than SOM and DCT.

Keyword Lip-reading, Visual features extraction, Hypercolumn model, Self organizing map, Hidden markov model

1. INTRODUCTION

Lip-reading has undergone much advancement over the last few years as the computer's role in our lives is becoming main and vital. Indeed, it smoothes human computer interface by introducing human-human interaction mechanism into the field of human-computer interface. The importance of lip-reading becomes clearer when the communication environment is not so suitable for speech perception such as communication among handicapped persons. In addition, it tends to be applied to areas such as speaker verification, multimedia telephony for the hearing impaired, cartoon animation, interaction with machines for handicapped and home health care systems for elderly people. Moreover, by combining the visual speech channel to acoustic 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] and [6].

Early evidence that vision can improve speech recognition was presented by Petajan [1]. In Petajan's system, binary mouth images are analyzed to calculate the distance of geometric measures among different mouth shapes in order to identify the visual representations of word units. Mase [2] used optical flow as input for a visual speech recognizer. Subsequent researches on implementing visual speech processing also include fuzzy logic [3] and self organizing map [4].

Later, with the beginning of the 90's decade, the development of hidden Markov models (HMM) improved the speech recognition accuracy and made possible large-vocabulary recognition. HMM was first applied to visual speech recognition by Goldschen [5] who modified the earlier Petajan's system by using discrete HMMs. Potamianos [6] combined the visual features either geometrically (lip's height and width) or nongeometrically using the wavelet transform of the mouth images to form a feature vector to train the HMM-based speech recognizer. Luettin [7] used HMM based active shape models to extract active features set that includes derivative information, and compared it's performance with the performance of a static feature set.

In this paper, a novel visual speech feature representation (or lip-reading) system is proposed. Our system consists of two consecutive processes: Visual speech feature extraction and visual speech feature recognition. First process is performed by HCM and second process is performed by HMM. The advantage of our system is that it is capable to extract all the relevant features without reduction and without the need to lips model or lips marker. In addition, the proposed system may work under shifted or rotated lip positions, which is not available in some other systems such as DCT-base systems; for instance see Heckmann [9]. Further, we compare our recognition results with those achieved by both SOM [4] using same database and Discrete Cosine Transform (DCT) [9] using different database. Comparison shows that HCM has a better performance than SOM and DCT for features extraction task.

The outline of this paper is as follows: Feature extraction process is outlined in section 2. Section 3 shows basics and algorithm of HCM model. Feature recognition process by HMM is covered in section 4. In section 5, experiments details and database are provided with the advantages of using HCM. Experimental results, result's analysis and comparison with both SOM and DCT are shown in section 6. Section 7 shows paper conclusion and future work.

2. VISUAL SPEECH FEATURES EXTRACTION

It is known that, most of lip-reading systems have two core stages: First stage is visual feature extraction which applies to each frame of the video image sequence, and second stage inputs the feature vector sequence, from first stage, and recognizes whole of the target sentence. In fact, performance of lip-reading system significantly depends on the feature extraction stage. Conventional lip-reading systems use different techniques to implement the feature extraction from the image. Among of these techniques are: Discrete Cosine Transform (DCT) [9] and Discrete Wavelets Transform (DWT) [10]. Any what the technique is, in our thinking, the visual feature extraction module is required to have the following three conditions:

1. It should make a parametric feature space with low dimensionality.
2. Distributions of each phoneme 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.

3. HYPERCOLUMN MODEL (HCM)

Hypercolumn model is an unsupervised neural network model. Mainly it designed as a pyramidal piling up hierarchical layers derived from the Neocognitron neural network (NC) model [12] by replacing each NC unit cell with a Hierarchical SOM (HSOM) [13]. As it is depicted in Figure 1, HCM has two intra-layers and one output layer. The output layer is selected as only one SOM with 64 neurons distributed in two dimensions (see a snapshot for the 64 neurons in Figure 3). Each intra-layer has number of overlapped HSOM units.

Each HSOM unit consists of two hierarchical SOM layers: Lower layer performs feature extraction by quantizing the input space and mapping it into a neuron array of low dimensional. The upper layer inputs the winner neuron index from lower layer to perform feature integration and then choose the winner neuron. Moreover, this layer enables shift, rotate and distort invariant recognition by a similar way as it is in the Neocognitron.

3.1 HCM Algorithm

HCM uses an unsupervised learning scheme to construct its layer's feature map. Only one cell will activate in correspondence to one category of input

patterns, other cells respond to other categories according to the following rule:

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

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 learning phase, each time a training data item is input, the winner is selected according to 1.

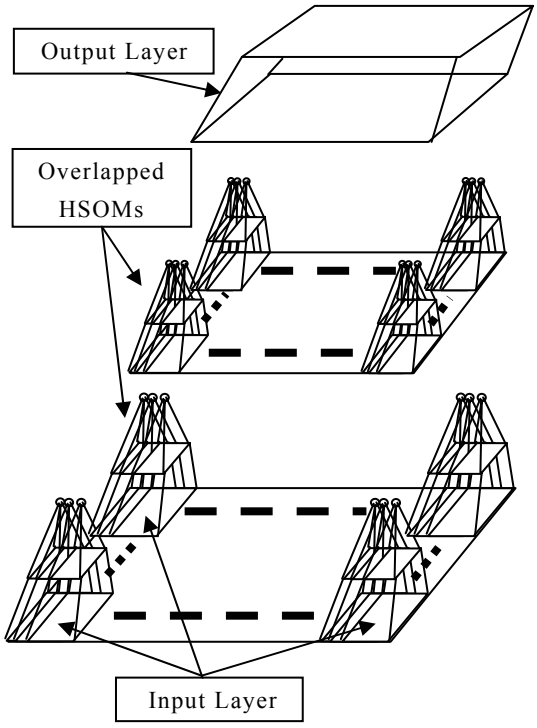

Figure 1: HCM model

The weight vectors are updated according to:

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

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

where $\alpha(t)$ is the learning rate and $\sigma^2(t)$ is a factor used to control the neighborhood range. The term $\|r_c - r_u\|$ is referring to distance between the neuron *c* and neuron *u*.

4. HMM-BASE VISUAL SPEECH FEATURES RECOGNITION

Needless to say that HMM holds the greatest promise among the various techniques used for visual speech recognition studied so far due to its capabilities in handling either the variability or the sequence of speech features. Visual speech features, extracted by HCM, will recognize using HMM. One HMM is constructed for each phoneme and continuous speech is recognized by joining the phonemes together to make any required word or sentence using pronunciation dictionary. Each HMM has five states from left to right and allows self-loops and sequential transitions between current state and the next state; see Figure 2.

Recognition process using HMM is divided into two stages: training stage and testing stage. In training stage, a training set of features and their associated transcriptions for each sentence are used to estimate the HMM parameters of that sentence. In testing stage, unknown features are transcribed and then the probability of each model generating that sentence is calculated. Finally, most likely model identifies the target sentence.

4.1 Visual Speech Features Modeling

For modeling visual speech features, consider a visual observation *O* of uttered sentence is represented by the following sequence of features vectors: $O = o_1, o_2, o_3, \ldots, o_T$, where o_t is the feature vector extracted at time t. Each HMM is initialized using a uniform segmentation, followed by iterative segmentation using Viterbi alignment approach. Each model parameters are further re-estimated using Baum–Welch procedure.

It is demonstrated in HMM base approach that each HMM representing a particular utterance is defined by the parameter set: $\omega_i = (\mathbf{A}, \mathbf{B}, \pi)$ where $\mathbf{A}=\{a_{ij}\}$ is a matrix of state transition probabilities from state *i* to state *j*, **B** is vector of observation (or output) probabilities $b_j(o)$ for state *j*, and π is the vector with probabilities π_i of entering the model at state *i* [14].

HMM-base recognition is performed using Viterbi algorithm which calculates the most likely state sequence for each HMM of having generated the observed sequence. In other words, sentence recognition is performed by estimating the following maximum a posteriori probability:

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

The probability included in 4 can obtain using Bayes rule:

$$P(\omega_i|O) = \frac{P(O|\omega_i)P(\omega_i)}{P(O)} \quad (5)$$

where $P(\omega_i)$ represents the prior probability of a category *i* and it is assumed to be equal for all categories,

and $P(O)$ is assumed to be constant for all categories. So, for simplicity, both $P(\omega_i)$ and $P(O)$ are ignored. For a state sequence has the form of x(1), x(2),…,x(T) of any model, the most probable spoken sentence depends only on the likelihood $P(O \mid \omega_i)$ and can be estimated as the product of state transition probabilities a_x and output probabilities $b_x(o)$ of the most likely state sequence. In other words,

$$P(O|\omega_i) = \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)}$$

$$= \prod_{t=1}^{T} b_{x(t)}(o_t) \prod_{t=1}^{T} a_{x(t)x(t+1)} \qquad (6)$$

4.2 Requirements for Phoneme Features Space

According to the above representation of each sentence, it is clear that our system is concerned with the output probability

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

of 6. To approximate this probability, the feature space should be low dimensionality and every distribution function should be simple. In the current project we approximate such probability using Gaussian distribution. In addition, we demonstrated, experimentally, that HCM can generate a low dimensional feature space. From this point of view, conditions 1 and 2, which mentioned in section 2, are satisfied.

5. EXPERIMENTS

It is known that, one of the most challenges in visual speech recognition domain is to cope with the large variability across speakers, due to individual appearance and individual lip movements and sizes. We therefore performed our experiments according to speaker-independent-base rule. Namely, we achieved experiments using Japanese database set includes 9 full sentences; see Table 1, uttered by 9 different Japanese adults, such that each adult uttered all sentences using different speakers for both training and testing phases.

The image database set includes 5670 gray images; image size is 160x120 pixels. We divided the image database set into two subgroups: First group, *training group*, has 3780 gray images gathered from 6 speakers. The second group, *test group*, has 1890 gray images gathered from 3 speakers different completely than those of the first group. Each uttered sentence represented by 70 visual frames.

Japanese Sentence	English Meaning
1- ATAMA ITAI	- A headache in head
2- SENAKA ITAI	- A pain in back
3- ONAKA SUITA	- Feel hungry
4- MUNE ITAI	- A pain in chest
5- TEACHI ITAI	- A pain in limbs
6- ATAMA OMOI	- Heavy head
7- ONAKA ITAI	- A pain in stomach
8- MUNE KURUSHI	- Difficult breath
9-TEACHI SHIBIRERU	-Spasm in hand and leg

Table 1: Japanese Sentences Database Set

It has been remarked that a considerable part of the existing work in visual speech recognition domain has been explored only by small data sets such as digits from 1 to 10 or isolated words or words repeated more than one time by same speaker as it is in [7], [9], [16] and [17]. Of course handling full grammatical sentences with a complete meaning has a higher challenge than isolated words.

5.1 HMM-Base Phoneme Recognition: Example

As an example for phoneme recognition process using our system, consider the word IT*A*I of the first sentence in Table 1; ATAMA ITAI. Specifically, consider the phoneme *A* of this word. In the structure of HMM in Figure 2, first and fifth node represents start and end node, respectively.

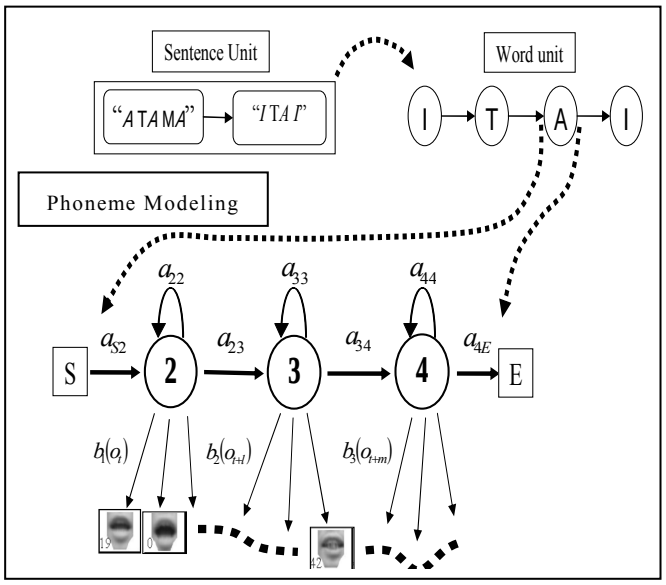

Figure 2: HMM-Base Lip-Reading for *ATAMA ITAI*

The other three middle nodes, from node 2 to node 4, are representing phoneme A. Definitely; node 3 represents the stable state of the phoneme, while node 2 and node 4 are

representing changing status from the previous phoneme T to the following phoneme I of the word *ITAI*. Since images have large dimensionality, then feature extraction process using HCM model is applied to draw a low-dimensional continuous space. In recognition process, each node from 2 to 4 outputs one frame image at the transition space according to the output probability $b_x(o)$. These output probabilities are approximated by single Gaussian distribution. Finally, the most likely state sequence of each HMM is calculated using Viterbi algorithm to match the best model of phoneme A.

5.2 Advantages of using HCM

Indeed, the fact of combining structure of both NC and SOM models lets HCM inherits the advantages of both models too. First of all, HCM is capable of generating ordered mappings of input data onto low-dimensional topological structure. This capability is very useful in analyzing high-dimensional data. Figure 3 shows a two-dimensional feature map (space) learned by HCM and depicts the codebook images of the 64 neurons of output layer.

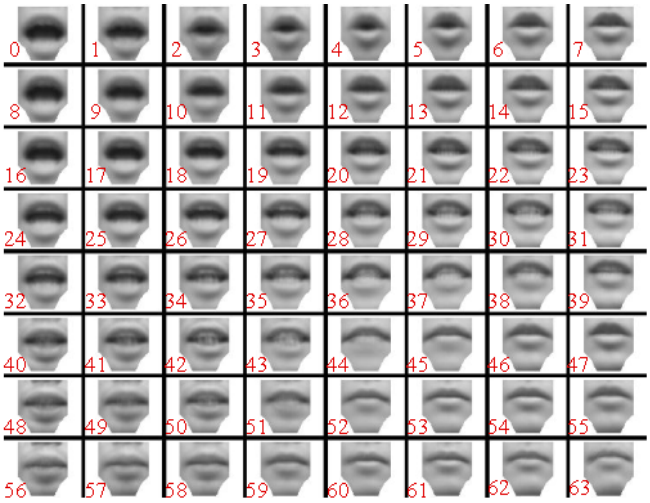

Figure 3: *Feature Map* of the output layer of HCM

Another challenge is the ability to partitioning the input data into clusters in such a way that gathering the similar data items into one cluster. This is due to HCM's ability to preserving the topological orders of input data items. In the same time, it is able to keeping close in the output space those data items which are closer to each other in the input space by using any function of distance measure; see Figure 4.

6. RESULTS & COMPARISON

Needless to say that, the target of improving the recognition accuracy stills an essential task in the field of pattern recognition. Table 2 shows our recognition results using current system (HCM+HMM) and those results achieved by *Tsuruta et al.* [4] using (SOM+HMM) system in case of using a full rank covariance matrix.

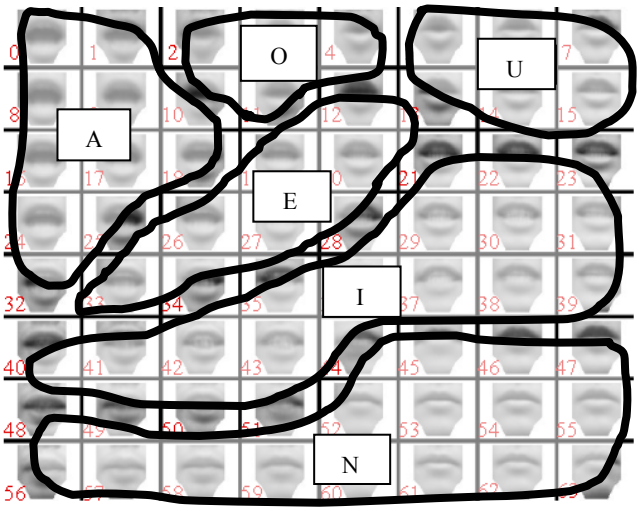

Figure 4: Vowels and consonant Japanese letters distributed on the *Feature Map Neuron* of Figure 3. Due to pronunciation similarities between both (A & E) and (O & U), HCM preserves their topological orderings by keeping each couple close in final output feature space.

	Training Data		Testing Data	
Target	SOM	***HCM***	SOM	***HCM***
Word	92.6	**95.4**	83.3	**90.7**
Sentence	88.9	**94.4**	70.7	**81.5**

Table 2: Recognition Accuracy Results

Both experiments are performed using same database, same dictionary and same phoneme models. Comparison between both systems' results demonstrates that HCM/HMM system's results are so supreme than those using SOM/HMM system in both word and sentence units for training and testing data set.

This means that HCM can extract and separate the features clearly in the feature space better than SOM. However, some dwindling occurred in the testing phase especially for sentence unit. This dwindling may due to pronunciation similarity between some Japanese characters like (A and E) and (O and U); see Figure 4. Furthermore, this dwindling may be caused by one or both of the following factors:

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

In future work, we will pay attention to these two factors.

On the other hand, comparison with other systems

performance reported elsewhere in the literature with different databases and conditions is not a fair way for judgment. Nevertheless, our results are measure well against the 70-80% different word accuracies by *Heckmann et al*. 2002 [9] who used DCT-base system. Another drawback in DCT-base that it is not a shift invariant recognition approach. In other words, a precise positioning of the region around the mouth from which features are derived is required to perform DCT experiments. In contrast, HCM enables shift and rotate invariant object recognition [8]. Finally, in DCT-base system a lot of coefficients, compared with HCM, have to be selected from each image frame in order to perform recognition experiments.

7. CONCLUSION & FUTUR WORK

It is known that, performance of visual systems significantly depends on the extracted features from images. Simply, feature extraction process in lip-reading systems involves the derivation of salient features from raw data in order to reduce the amount of data used in classification. In this paper, we proposed a novel visual speech features representation system. The proposed system is a combination of HCM neural network model with HMM. The system performance is examined using multiple sentences of Japanese language. Comparison turned out that our system accuracy results are so higher than others. Even though, the drawback of our system is the recognition time is still higher than the recognition time of SOM/HMM. One of our urgent future tasks is to modeling the recognition parameters of HCM.

REFERENCE

[1] E. D. Petajan, "Automatic lip-reading to enhance speech recognition,"IEEE Global Telecommunications Conf: 265-272, 1984.

[2] K. Mase and A. Pentland, "Automatic lip-reading by optical flow analysis," Systems and Computers in Japan, 22 (6): 67-75, 1991.

[3] P. L. Silsbee and A. C. Bovik, "Computer lip-reading for improved accuracy in automatic speech recognition," IEEE Trans. Speech Audio Processing, 4 (5): 337-351, 1996.

[4] N. Tsuruta, H. Iuchi, A. Sagheer, T. Tobely, "Self–Organizing Feature Maps for HMM Based Lip-reading," The 7th Int. conf. on Knowledge-Based Intelligent Information & Engineering Sys, KES, 2: 162-168, 2003.

[5] A.J. Goldschen, O.N. Garcia, and E. Petajan," Continuous optical automatic speech recognition by lip-reading," the 28th Annual Asilomar conf. on Signal Systems, and Computer, 1:572-577, 1994.

[6] G. Potamianos, E. Cosatto, H.P. Graf, and D.B. Roe, "Speaker independent audiovisual database for bimodal ASR," European Tutorial Workshop Audiovisual Speech Processing, Greece, 1997.

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

[8] N. Tsuruta, R. Taniguchi, M. Amamiya," Hypercolumn Model: A Combination Model Hierarchical Self-Organizing Maps and Neocognitron for Image Recognition," Systems and Computers in Japan, 31(2): 49-61, 2000.

[9] M. Heckmann, K. Kroschel, C. Savariaux, F. Berthommier, "DCT-Based Video Features For Audio-Visual Speech Recognition," Proc. Of Inter. Conf. on Spoken Language Processing, ICSLP: 1925-1928, 2002.

[10] G. Potamianos, H. P. Graf, and E. Cosatto, "An Image Transform Approach for HMM Based Automatic Lip Reading," Int. Conf. on Image Proc., 3:173-177, 1998.

[11] T. Tobely, N. Tsuruta, M. Amamiya, "A randomized Model for the Hypercolumn Neural Network for Gesture Recognition," Journal of Comp., Sys, and Signals, 3 (1): 14-18, 2002.

[12] K. Fukushima, "Neocognitron," Biol Cyberetics, 36 (4) : 193-202, 1980.

[13] J. Lampinen, E. Oja, "Clustering Properties of Hierarchical Self Organizing Maps," J. of Math Imaging and Vision, 2, 1992.

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

[15] HMM Tool Kit, (online) { http://htk.eng.cam.ac.uk}.

[16] I. Matthews, T. Cootes, A. Bangham, and S. Cox, "Extraction of Visual Features for Lip-Reading," IEEE Trans. on Pattern Ana. and Machine Intel. 24(2), 2002.

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