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

A Combination of Hyper Column Model with Hidden Markov Model for 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

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

他

http://hdl.handle.net/2324/5856

出版情報：2004-12
バージョン：
権利関係：

KYUSHU UNIVERSITY

A Combination of Hypercolumn Model with Hidden Markov Model for Japanese Lip-Reading System

Alaa Sagheer[1], Naoyuki Tsuruta[2], Rin-Ichiro Taniguchi[1], Sakashi Maeda[2], Seiji Hashimoto[2]

[1] Dept. of Intelligent Systems, Kyushu University, Fukuoka, Japan
[2] Dept. of Electronics Engineering and Computer Science, Fukuoka University, Fukuoka, Japan
{alaa, rin}@limu.is.kyushu-u.ac.jp, {tsuruts, maeda, hasimoto}@mdmail.tl.fukuoka-u.ac.jp

Abstract. In recent years, lip-reading systems have received much attention, since they play an important role in human communication with computer especially for hearing impaired and elderly people. In this paper, we introduce a novel Japanese lip-reading system combines Hypercolumn Neural Network model (HCM) with Hidden Markov Model (HMM). In this system, we use HCM to extract the visual speech features from input image. The extracted features are modeled by Gaussian distributions, which is used in recognition phase using HMM. 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 special lighting or lip markers used. Experimental results demonstrate that the proposed system performance is supreme.

Keywords: visual speech recognition, lip-reading system, hypercolumn model, 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, see, e.g. , (Goldschen et al. 1998) and (Chen et al. 1998).

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 word or sentence. Regarding to first lip-reading stage there are two main approaches for phoneme feature extraction:

- Geometric-feature based; and
- 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, see, e.g., (Stiefelhagen et al. 1997) and (Chen 2001). 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, see, e.g., (Silsbee et al. 1996) and (Potamianos et al. 1998). Due to 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, see, e.g., (Luettin et al. 1997) and (Chan 2001).

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) (Deligne et al. 2002), or the discrete wavelet transform (DWT) (Potamianos et al. 1998). In the present work, we propose a Japanese lip-reading module categorized as appearance based approach. The proposed system combines Hypercolumn neural network Model (HCM) (Tsuruta et al. 2000), with Hidden Markov Model (HMM). HCM is used in the first stage of lip-reading system to extract the relevant phoneme features. HMM is used in the second

lip-reading stage for feature sequence recognition. The advantage of our system is that the system is capable to extract all relevant image features without reduction and without the need to lips model or lips marker. Furthermore, the proposed system may work under shifted or rotated lip positions. Experiments are performed using multiple full Japanese sentences gathered from 9 different speakers with a total number of images is 5670 image.

The outline of this paper is as follows: Basics and characteristics of HCM model are covered in Section 2. Section 3 shows our HMM-base lip reading system. Database and experiments are given in Section 4. In Section 5 experimental results are provided. Paper conclusion and future work are given in Section 6.

2 Hypercolumn Model (HCM)

Hypercolumn model is an unsupervised neural network model. Mainly it designed as a pyramidal piling up hierarchical layers derived from Neocognitron neural network (NC) model (Fukushima 1980) by replacing each NC unit cell plane with a Hierarchical Self Organizing Map (HSOM) (Lampinen et al. 1992). For the current project, we chose HCM with three layers, two of them as intra layers and the other as output layer.

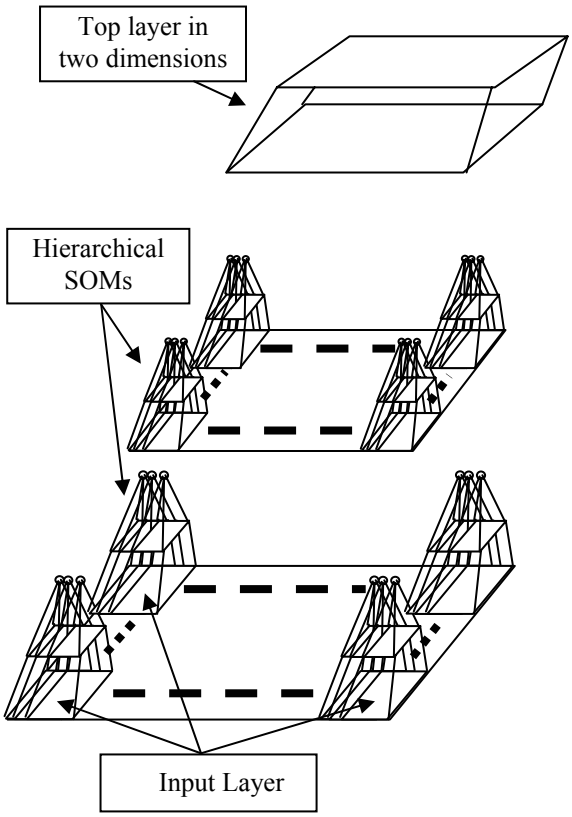

Figure 1. HCM model

2.1 HCM Intra-Layers: Structure and Features

As it is depicted 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, and 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 and inputs the winner neuron index from the first SOM layer to perform optimum feature integration and then choose the winner neuron. Moreover, this layer enables shift, rotate, scale and distort invariant recognition by a similar way as it is in the Neocognitron.

The advantage of overlapping occurred between HSOMs is that it helps in covering all lips positions in case of lip's shift or rotate, and then enables us to obtain multiple classifications for every input data. Each

HSOM unit works by selecting one winner for each input position, the winner of all winners is considered as the unit's output. Then, the array of all winner position indexes from all HSOMs in each intra-layer forms the final output of that HCM intra-layer and, in the same time, the input of the following HCM layer. The output of each layer is feed directly to the following layer without any data manipulations between layers.

In the current project, first HCM intra-layer includes 25 x 18 HSOMs. Through the competition phase, each HSOM's neurons are competing 3 times in 3 different positions in X and Y dimension with shift step 2 times in each dimension. Second HCM intra-layer includes 10 x 6 HSOMs and the neurons are competing 2 times in 2 different positions with shift step 1 in each dimension.

2.2 HCM Output Layer: Structure and Features

As it is shown in Figure 1, the size of the hierarchical layers of HCM model decreases as the layer index increases. The HCM output layer (top-layer) is selected as only one SOM with 64 neurons distributed in two dimensions. In this project, we chose SOM in two-dimensions not in one dimension as other SOMs in the intra-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 wish to achieve a finer-grained classification for learning data. Especially, if we know that HCM top layer integrates all information from the intra-layers. Finally, due to the competitive learning nature of HCM, only one cell from this layer will be activated in correspondence to the category of the input pattern, other cells respond to other categories.

2.3 HCM Algorithm

HCM uses an unsupervised learning scheme to construct its layer's feature maps. As it is mentioned above, only one neuron *c* from each HSOM unit will activate according to the following rule:

$$\|I - W_c\| = \min_u(\|I - W_u\|), \tag{1}$$

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 1. The weight vectors are updated according to:

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

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

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*. The learning process line of HCM is carried out beginning from bottom to top, layer by layer, where the standard learning algorithm of HSOM (Lampinen 1992) is used to train each HSOM's unit neurons.

2.4 HCM Advantages

Indeed, the fact of combining the 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 the input data onto some low-dimensional topological structure. This capability is very useful in analyzing high-dimensional data. 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 the ability of HCM to 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 function of distance measure. Other HCM advantages can be summarized as follows (Tsuruta 2000):

- HCM can recognize distort and invariant objects with variations in position and size.
- Accept random initialization for network weights.
- No preprocessing for the input images is needed.

3 HMM-Base Lip-Reading System

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. The visual features, which are extracted by HCM, will recognize using HMM. Recognition process by HMM is divided into two stages: learning (or training) stage and testing stage. In learning stage, a learning set of features and their associated transcriptions for each sentence are used to estimate the HMM parameters of that sentence. The parameter's estimation process is performed using Baum-Welsh procedure. 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.

3.1 Visual Speech Features Recognition

To recognize the visual speech features, consider a visual observation O of each uttered sentence is represented by the following sequence of features vectors:

$$O = o_1, o_2, o_3, \ldots, o_T, \tag{4}$$

where o_t is the feature vector extracted at time t. For each sentence, we build an HMM with five states. Each HMM is initialized using a uniform segmentation, followed by iterative segmentation using Viterbi alignment approach. The model's 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 the matrix of state transition probabilities from state i to state j, $\mathbf{B}$ is the 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, see, e.g., (Rabiner 1989).

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

$$\arg\max_i \{P(\omega_i | O)\}, \tag{5}$$

The probability included in 5 can obtain using Bayes rule:

$$P(\omega_i | O) = \frac{P(O|\omega_i)P(\omega_i)}{P(O)}, \tag{6}$$

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. For simplicity, both $P(\omega_i)$ and $P(O)$ are ignored.

Therefore, the most probable spoken sentence depends only on likelihood $P(O \mid \omega_i)$. To calculate this likelihood, it is known that Markov model is a finite state machine changes state once every time unit and at each time t when a state j is entered, a speech vector o_t is generated from the probability density $b_j(o_t)$. Transition from state i to state j is governed by the transition probability a_{ij}. Thus, the required likelihood $P(O|\omega_i)$ for a state sequence x(1), x(2),……,x(T) of any model can be estimated by

$$\begin{aligned} P(O|\omega_i) &= \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \\ &= \prod_{t=1}^{T} b_{x(t)}(o_t) \prod_{t=1}^{T} a_{x(t)x(t+1)} \end{aligned} \tag{7}$$

i.e. the product of state transition probabilities a and output probabilities $b_x(o)$ of the most likely state sequence

4 Experiments

We explored the performance of the proposed lip-reading system using Japanese language data set. One of the most challenges in lip-reading applications 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 type using different speakers for training and testing, to investigate how well the system generalizes to new speakers. Namely, experiments are achieved by 9 different Japanese adults, such that each adult uttered 9 Japanese sentences. Table 1 shows the uttered sentences and their translation to English.

Table 1. Japanese Sentences

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

It has been remarked that a considerable part of the existing work in the 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, see, e.g., (Luettin et al. 1997), (Chen 2001) and (Heckmann et al. 2002). Of course handling full grammatical sentences with a complete meaning has a higher challenge than isolated words.

4.1 Image Database

Image database set includes 5670 gray images captured directly by using EVI-G20 Sony camera, where the size of each input image is 160x120 pixels. We divided the image database set into two subgroups. First group, *training group*, has 3780 gray images gathered from 6 different Japanese adults. Second group, *test group*, has 1890 gray images gathered from 3 Japanese adults different completely than those of first group. Each uttered sentence represented by 70 visual frames. Figure 2 shows a snapshot of our online lip-reading system. It is clear in Figure 2 that user is required to put his/her mouth into the shown white rectangle. After easy training for user, recognition accuracy was almost the same average with off-line case.

Figure 2. A Snapshot of the Online System

5 Experimental Results

Table 2 shows recognition accuracy results for both word unit and sentence unit in case of using a full rank covariance matrix.

Table 2. Recognition Accuracy using HCM

Target	Learning data	Testing data
Word	95.4	90.7
Sentence	94.4	81.5

Although, we remark that our system (HCM+HMM) performs high accuracy especially for learning data, some dwindling is occurred for the testing data results especially for the sentence unit, this dwindling may be caused by one or both of the following reasons:

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

6 Conclusion & Future Work

In this paper, we proposed a novel Japanese lip-reading system. The proposed system is a combination of HCM neural network model with HMM. HCM model is used as a feature extractor, while HMM is used as a feature recognizer. The system performance is examined using multiple sentences of Japanese language. Image database set includes 5670 gray image for training and testing phases. All images were captured in a natural environment without special lighting or lip markers used. Although, our system shows considerable results, some dwindling happened in the testing phase. To overcome this dwindling we intend to increase the training data set in our future work.

References

Chan, M. T. (2001), "HMM-Based Audio-Visual Speech Recognition integrating Geometric-and Appearance-Based Visual Features," *IEEE Workshop on Multimedia Signal Processing*, pp. 9-14.

Chen, T. (Jan 2001)," Audiovisual Speech Processing, Lip Reading and Lip Synchronization," *IEEE Signal Processing Magazine*, pp. 9-21.

Chen, T. and Rao, R. R. (May 1998), "Audio-visual integration in multimodal communication," *Proceeding IEEE*, vol. 86, pp. 837–852.

Deligne, G., Potamianos, C., and Neti, C. (2002), "Audio-visual Speech Enhancement with AVCDCN," *Int. Conf. on Spoken Language Processing, ICSLP02*, pp.1449-1452.

Fukushima, K. (1980), "Neocognitron," *Biol Cyberetics*, vol. 36, no. 4, pp. 193-202.

Goldschen, A.J., Garcia, O.N., and Petajan, E. (1994), "Continuous optical automatic speech recognition by lipreading," *Proceeding the 28th Annual Asilomar conference on Signal Systems, and Computer*, vol. 1, pp. 572-577.

Heckmann, M., Kroschel, K., Savariaux, C. and Berthommier, F. (2002) "DCT-Based Video Features For Audio-Visual Speech Recognition," *Proceeding of Int. Conf. on Spoken Language Processing, ICSLP02*, pp. 1925-1928.

HMM Tool Kit (available on line): http://htk.eng.cam.ac.uk.

Lampinen, J. and Oja, E. (1992), "Clustering Properties of Hierarchical Self Organizing Maps," *Math. Imaging and Vision*, vol.2.

Luettin, J. and Thacker, N. (1997), "Speechreading using Probabilistic Models," *Computer Vision and Image Understanding,* vol. 65, no. 2 , pp 163-178.

Potamianos, G., Graf, H. P., and Cosatto, E. (1998),"An Image Transform Approach for HMM Based Automatic Lip Reading," *Proceeding of the Int. Conf. on Image Processing*, vol. III, pp. 173-177.

Rabiner, L.R. (Feb. 1989), "A tutorial on Hs and Selected Applications in Speech Recognition," *Proceeding of the IEEE*, vol. 77, no. 2, pp. 257-286.

Silsbee, P. L. and Bovik, A. C. (1996), "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 5, pp. 337–351.

Stiefelhagen, R., Meier, U., and J. Yang (1997), "Real-Time Lip-Tracking for Lipreading," *Proceeding. of Eurospeech 97*.

Tsuruta, N., Taniguchi, R., and Amamiya, M. (2000), "Hypercolumn Model "A Combination Model Hierarchical Self- Organizing Maps and Neocognitron for Image Recognition," *System and Computer in Japan*, vol. 31, no. 2.

Yuhas, B. P., Goldstein, M. H., and Sejnowski, T. J. (1989), "Integration of acoustic and visual speech signals using neural networks," *IEEE Communication Mag.*, vol. 27, pp. 65-71.