Visual Speech Features Representation for Automatic Lip-reading

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：Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on. 2, pp.781-784, 2005-03. Institute of Electrical and Electronics Engineers
バージョン：
権利関係：(c)2005 IEEE

Visual Speech Features Representation for Automatic Lip-reading

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

[1] Department of Intelligent Systems, Kyushu University, Japan

[2] Department of Electronics Engineering and Computer Science, Fukuoka University, Japan

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

ABSTRACT

A fundamental task in pattern recognition field is to find a suitable representation for a feature. In this paper, we present a new visual speech feature representation approach that combines Hypercolumn Model (HCM) with HMM to perform a complete lip-reading system. In this system, we use HCM to extract visual speech features from input image. The extracted features are modeled by Gaussian distributions through using HMM. The proposed lip-reading system can work under varying lip positions and sizes. All images were captured in a natural environment without using special lighting or lip markers. Experimental results are shown to compare favourably with the results of two reported systems: SOM and DCT base systems. HCM provides better performance than both systems.

1. INTRODUCTION

Recently, visual speech recognition has undergone much interest and advancement, especially when the visual speech channel is combined to the 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].

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 and self organizing map [3].

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 [4] who modified the earlier Petajan's system by using discrete HMMs. Potamianos [7] 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.

In this paper, a novel visual speech feature representation (and lip-reading) system is proposed. Our system consists of two consecutive stages: Visual speech feature extraction and visual speech feature recognition. First stage is performed by HCM and second stage is performed by HMM. The advantage of our system is that it is capable to extract all the relevant features without reduction and without needing to lips model or lips marker. Furthermore, the proposed system may work under shifted or rotated lip positions, which is not available in some other systems such as DCT-base systems; Heckmann [6]. Further, we compare our recognition results with those achieved by both SOM [3] using same database and Discrete Cosine Transform (DCT) [6] using different database. Comparison shows that HCM has a better performance than SOM and DCT for features extraction task.

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 first stage. Simply, feature extraction process in lip-reading systems involves the derivation of salient features (phoneme) from raw data in order to reduce the amount of data used in classification. In fact, conventional lip-reading systems use different techniques to implement the feature extraction from the image. Among of these techniques are: Discrete Cosine Transform (DCT) [6] and Discrete Wavelets Transform (DWT) [7]. Regardless this fact, in our thinking, the visual feature extraction module is required to have the following two 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.

As it is explained later in section 4, using HCM satisfies these conditions.

3. HYPERCOLUMN MODEL (HCM)

Hypercolumn model is an unsupervised neural network model. Mainly it is derived from the Neocognitron neural network (NC) model [8], which is a multi-layered neural network, by replacing each NC unit cell with a Hierarchical SOM (HSOM) [9]. As it is depicted in Figure 1, HCM has two intra-layers and one output layer. The output layer consists only one SOM with 64 neurons distributed in two dimensions. 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.

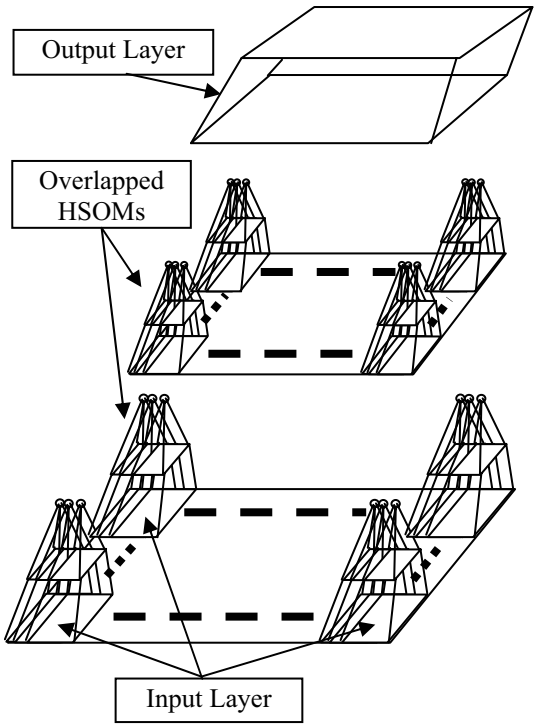

Figure 1: HCM model

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 Eq (1). 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 the distance between the neuron *c* and neuron *u*.

3.2 HCM Advantages

HCM is capable of generating ordered mappings of input data onto low-dimensional space [5]. 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 (preserving the topological order); see Figure 3. Other HCM advantages can be summarized as follows [5]:

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

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, is recognized 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 next state; see Figure 2.

Recognition process using HMM is divided into two phases: training and testing. In training phase, a training set of features and their associated transcriptions for each sentence are used to estimate the HMM parameters of that sentence. In testing phase, 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* [10].

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 Eq. (4) can be obtained 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 $P(O)$ 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,

$$P(O|\omega_i) = \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} = \prod_{t=1}^{T} b_{x(t)}(o_t) \prod_{t=1}^{T} a_{x(t)x(t+1)}. \quad (6)$$

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 Eq. (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.

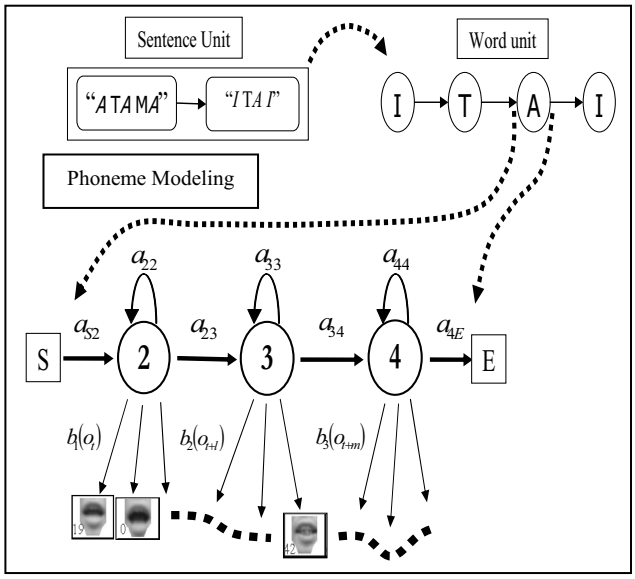

Figure 2: HMM-Base Lip-Reading for *ATAMA ITAI*

4.3 HMM-Base Phoneme Recognition: Example

Here, we show an example of speech recognition process using our system. For instance, consider the word IT*A*I of the first Japanese sentence in Table 1; ATAMA ITAI (A Headache in head). 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. The other three nodes, from node 2 to node 4, are representing phoneme A. 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 current and from the current to the following phoneme I, respectively.

Since images have large dimensionality, then feature extraction process using HCM model is applied to draw a low-dimensional continuous space. In the recognition process, each node from 2 to 4 outputs one frame image at the transition space according to the output probability $b_x(o)$, as it is shown in Figure 2. These output probabilities are approximated by single Gaussian distribution. Finally, the most likely state sequence of each HMM is calculated using Viterbi algorithm to represent the phoneme A.

5. EXPERIMENTS

It is known that, one of the biggest 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 using different speakers for both training and testing phases. Namely, we performed our experiments using Japanese database set includes 9 full sentences uttered by 9 different Japanese adults, such that each adult uttered all sentences one time. Table 1 shows the uttered sentences and their English meaning.

Table 1: Japanese Sentences Database Set

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

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.

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 [6] and [12] for examples. Of course handling full grammatical sentences with a complete meaning has a higher challenge than isolated words.

6. EXPERIMENTAL RESULTS & COMPARISON

Needless to say that, the target of improving the recognition accuracy stills an essential target in the field of pattern recognition. Table 2 shows our recognition results using current system (HCM+HMM) and those results performed by *Tsuruta et al.* [3] using (SOM+HMM) system in case of using a full rank

covariance matrix. Both experiments are performed using same database, same dictionary and same phoneme models.

Table 2: Recognition Accuracy Results

	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

Fairly comparison between both systems demonstrates that HCM/HMM system's results are better than those using SOM/HMM system in both units of word and sentence for training and testing phases. 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 3. 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 measured well against the 70-80% different word accuracies by *Heckmann et al.* 2002 [6] who used DCT-base system. Another drawback, DCT-base system is not a shift invariant recognition approach. In other words, a precise positioning of the region around the mouth is required to perform DCT experiments.

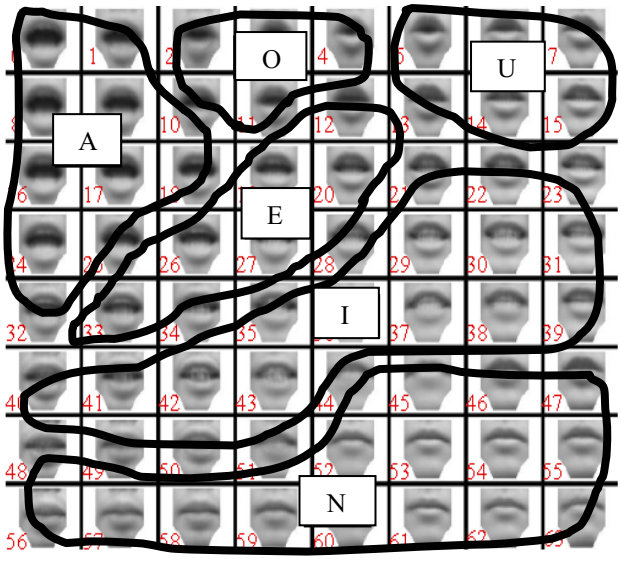

Figure 3: *Feature Map Neuron* of HCM includes Japanese vowels and consonant. Due to pronunciation similarities between both (A & E) and (O & U), HCM preserves their topological orderings by keeping each couple close in final output feature space.

In contrast, HCM enables shift and rotate invariant object recognition [5]. Finally, in DCT-base system experiments, a lot of coefficients have to be selected from each image frame.

7. CONCLUSION & FUTUR WORK

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 higher than others. Even though, the drawback of our system is the recognition time is still longer than the recognition time of SOM/HMM. One of our urgent future tasks is to modeling the recognition parameters of HCM.

8. 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] 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, KES03, 2: 162-168, 2003.

[4] 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.

[5] 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.

[6] 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.

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

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

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

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

[11] HMM Tool Kit, [Online] Available: http://htk.eng.cam.ac.uk

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