ICICS-PCM 2003
15-18 December 2003
Singapore

Modeling Continuous Visual Speech Using Boosted Viseme Models

Liang Dong[1], Say Wei Foo[2], and Yong Lian[3]

[1,3]Department of Electrical and Computer Engineering
National University of Singapore
Email: {engp0564, eleliany}@nus.edu.sg

[2]School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
Email: eswfoo@ntu.edu.sg

Abstract

In this paper, a novel connected-viseme approach for modeling continuous visual speech is presented. The approach adopts AdaBoost-HMMs as the viseme models. Continuous visual speech is modeled by connecting the viseme models using level building algorithm. The approach is applied to identify words and phrases in visual speech. The recognition results indicate that the proposed method has better performance than the conventional approach.

1. Introduction

In modern speech processing, there has been growing interest in applying visual information to an acoustic speech recognition system. The technique of retrieving speech content from visual clues such as the movement of the lip, tongue and facial muscles is commonly known as automatic lip reading. Much effort has been made on this area since 1990s. Research on automatic lip reading is largely influenced by acoustic speech processing. The techniques/tools of modeling and identifying acoustic speech are applied to visual speech processing as well, for example, Neural Network and Hidden Markov Models. However, it is not an easy task to apply them to recognize visual speech because of some distinct features of lip motion.

In this paper, our research on visual speech processing using HMM-based approach is reported. HMM-based visual speech recognition can be divided into two branches. The first branch is to model and identify the visual speech elements, such as visemes and words, with single models. The second branch is to model the continuous visual speech, such as connected-viseme, connected-word, using connected models. In this paper, the construction and performance of the boosted viseme model are briefed first. After that, the strategy of connecting the boosted viseme models is given to model continuous visual speech. We focus on the strategy of building connected AdaBoost-HMM chain by means of level building. The principle of the method is to maintain a two-dimensional probability trellis to record the accumulated likelihood scored by all the AdaBoost-HMMs at each position. The optimal viseme sequence is decoded by backtracking the probability trellis. The connected boosted viseme models are applied to identify words and phrases in our experiments. Its performance is compared with the traditional connected single-HMMs. Experimental results show that better recognition accuracy is achieved with the proposed approach.

2. Boosted Viseme Classifier

The basic visual speech element – viseme, is a special type of time sequence. It indicates a short period of lip movement that is repeated in different articulations. In natural speech, the visemes may be distorted by their context and thus demonstrate spread-out distribution. The traditional single HMM classifier is sometimes ineffective to model the viseme with such distribution.

In our approach, the visemes are modeled by Adaptive Boosted (AdaBoost) HMMs. The AdaBoost-HMM is a kind of multi-HMM classifiers. As illustrated in Fig. 1, the classifier of the viseme category d_k (k=1,2,…K) comprises L_k HMMs and L_k weights $\alpha_1, \alpha_2, \cdots \alpha_{L_k}$, which are assigned to the composite HMMs.

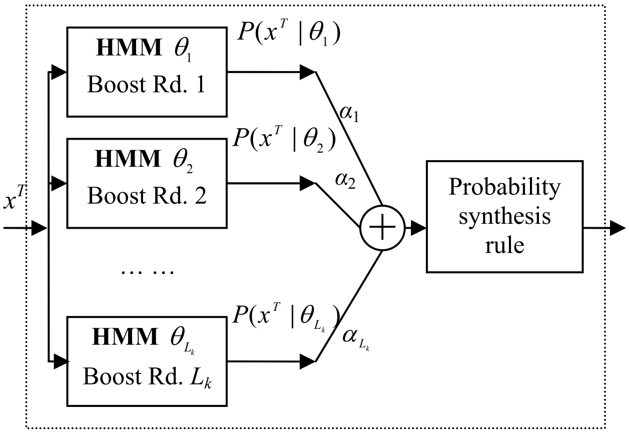

Figure 1. Block diagram of a boosted viseme model

0-7803-8185-8/03/$17.00 © 2003 IEEE

In the training phase, the composite HMMs of the AdaBoost-HMM are trained using the AdaBoosting method and the Baum-Welch estimation. The weights that are assigned to the HMMs are adjusted according to the classification error of the respective HMMs [4]. A detailed discussion about the training of the AdaBoosted-HMM is given in [2]. For the example that is illustrated in Fig. 1, HMM boosting is implemented for L_k rounds. The AdaBoost-HMM for d_k thus consists of L_k HMMs.

In the recognition phase, each composite HMM measures the likelihood for the input vector sequence x^T. Let $\Theta_k = \{\theta_1, \alpha_1; \theta_2, \alpha_2; \cdots \theta_{L_k}, \alpha_{L_k}\}$ denote the AdaBoost-HMM of viseme d_k, the probabilities scored by the composite HMMs are synthesized according to (1).

$$\overline{P}(x^T \mid \Theta_k) = \frac{1}{L_k} \sum_{i=1}^{L_k} \log[\alpha_i P(x^T \mid \theta_i)] \qquad (1)$$

$\overline{P}(x^T \mid \Theta_k)$ is the log likelihood normalized by the number of the composite HMMs. Since the composite HMMs of a boosted viseme classifier highlight different groups of training samples, the AdaBoost-HMM classifier obtained in this way has the ability to better cover the spread-out distribution of the samples than the traditional single-HMM classifier. The application of AdaBoost-HMM classifier is presented in [2]. An average 20% improvement on classification rate for identifying visemes is observed.

3. Continuous Speech Modeling

The success of the application of the AdaBoost-HMM on viseme modeling is encouraging. However, this is only the first step of visual speech processing. One of the most important extensions of the AdaBoost-HMM classifier is to model and identify continuous visual speech. For the problem to be solved in this paper, the visual speech elements comprising multiple visemes, such as words and phrases, are objects to be modeled and identified.

One approach of word identification is to build an AdaBoost-HMM for each word that may be appeared in the speech. This method works well in small-vocabulary cases. However, as the volume of the vocabulary increases, it is usually difficult to find enough training samples for each word. The number of the visemes, on the other hand, is vey limited. According to MPEG-4 Multimedia Standard, there are only fourteen visemes in English [5]. Each viseme corresponds to several visually similar phonemes or phoneme-like productions. As a result, it is much easier to find enough training samples for the visemes than for the words. In addition, by applying connected-viseme model, the system can make a good guess to the word that is new to it. It thus meets the requirement of text-independence of a speech recognition system.

In an AdaBoost-HMM, the composite HMMs are joint together to measure the likelihood of the input. Assume that $x^T = (x_1, x_2, \cdots x_T)$ is a T-length vector sequence that indicates a word production, θ_i^k $(\theta_i^k \in \Theta_k)$ is the i-th composite HMM of the boosted classifier Θ_k, the underlying process of computing the likelihood is given in Fig. 2, where S_0 is a null state that indicates the start of the HMM and S_e is also a null state that indicates the end of the HMM. $P_i^k(j) = P(x_1, x_2, \cdots x_j \mid \theta_i^k)$ is the accumulated likelihood scored from time 1 to j by θ_i^k. This entry can be computed using the forward variables [1]. While building a chain of AdaBoost-HMMs to match the target sequence, the key problem is to align the composite HMMs so that they start and end at the same frames.

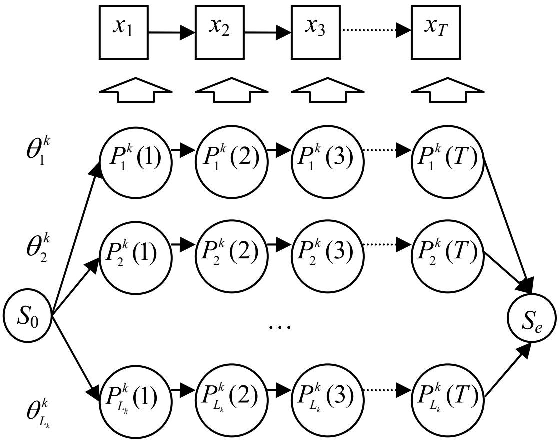

Figure 2. The underlying process of computing the likelihood by the composite HMMs of an AdaBoost-HMM

The purpose of level building is to search the sequence of reference models $\Theta^\eta = (\Theta_{l_1}, \Theta_{l_2}, \cdots \Theta_{l_\eta})$ that consists of η AdaBoost-HMMs. Such a sequence is considered optimal if it maximizes $P(x^T \mid \Theta^\eta)$. To meet this requirement, a probability trellis that records the probabilities scored by the AdaBoost-HMMs is constructed in the proposed level building approach.

Before giving the procedures of level building, the approximate number of the visemes appeared in x^T is first estimated based on the duration of x^T, e.g. from $\eta_{\min}$ to $\eta_{\max}$. The durations of the visemes are also estimated. For viseme d_k, the duration is from $d_{\min,k}$ to $d_{\max,k}$. The above assumptions are reasonable in visual speech processing. If a word-production lasts for one second, the number of the visemes appeared in it is very likely to be within 2~7. The durations of the visemes may range from 0.1sec to 0.8sec in most cases. In level building on the boosted viseme models, each level corresponds to a viseme that appears in the target sequence and the number of the levels equals to the number of the viseme models. Assume that all the composite HMMs are N-state discrete HMMs, where $\{S_1, S_2, \ldots S_N\}$ is the state set. Level building on the AdaBoost-HMMs is implemented with the following steps:

Step 1. Computing the accumulated likelihood

Starting from level 1, frame 1, the accumulated likelihood is computed for all the composite HMMs of the AdaBoost-HMM. For example, at the node at frame t on level η, $d_{\max,k}$-$d_{\min,k}$+1 accumulated probabilities are computed for θ_i^k (k=1,2,…K, i=1,2,… L_k). The probabilities are denoted as $P(t',t,\eta,\theta_i^k)$, which indicates that θ_i^k starts at frame t'.

$$P(t',t,\eta,\theta_i^k) = P(x_{t'}, x_{t'+1}, \cdots x_t \mid \theta_i^k) \quad (2)$$

The range of t' is from t- $d_{\max,k}$ to t- $d_{\min,k}$. This entry is computed using the forward variables $\alpha_i^k(t',t)$ [1].

$$\alpha_i^k(t',t) = P(x_{t'}, x_{t'+1}, \cdots x_t, s_t \mid \theta_i^k) \quad (3)$$

where s_t is the t-th state of the state chain. By summing up all the states at time t, we have

$$P(t',t,\eta,\theta_i^k) = \sum_{j=1}^{N} P(x_{t'}, x_{t'+1}, \cdots x_t, s_t = S_j \mid \theta_i^k) \quad (4)$$

Step 2. Building the probability trellis

In level building, a probability trellis as shown in Fig. 3 is built to map the frames of the observation sequence to the states of the composite HMMs at different levels. The nodes at the end of the levels are referred to as end nodes so as to separate from the other nodes. The probabilities scored by the composite HMMs are synthesized at the end nodes. At a possible end node on level η, frame t, the probabilities scored by the composite HMMs are synthesized using (5),

$$\overline{P}(x^T,t',t,\eta \mid \Theta_k) = \frac{1}{L_k}\sum_{i=1}^{L_k} \log[\alpha_i^k P(t',t,\eta,\theta_i^k)] \quad (5)$$

Since t' ranges from t- $d_{\max,k}$ to t- $d_{\min,k}$, $d_{\max,k}$-$d_{\min,k}$+1 log probabilities for Θ_k are computed and saved at the end node. The composite HMMs thus start from the same frame t' and terminate at the same frame t. For each AdaBoost-HMM, such an array of probabilities is retained.

The accumulated synthesized log likelihood starting at frame 1, ending at frame t on level η, scored by the reference model Θ_k, is denoted as $P(t,\eta,\Theta_k)$. This entry is computed using the recursive means given in (6).

$$P(t,\eta,\Theta_k) = \max_{t_1<t_{\eta-1}<t_2}[P(t_{\eta-1},\eta-1) + \overline{P}(x^T,t_{\eta-1}+1,t,\eta \mid \Theta_k)] \quad (6)$$

where $t_{\eta\text{-}1}$ is the end frame of level η-1, $t_1 = t - d_{\max}(\Theta_k) - 1$ and $t_2 = t - d_{\min}(\Theta_k) - 1$. $P(t_{\eta-1},\eta-1)$ is the optimal log likelihood accumulated to level η-1, frame $t_{\eta\text{-}1}$. A general formula for computing $P(t,\eta)$, which is the optimal log likelihood accumulated to level η, frame t, is the maximization over all the reference models Θ_k (k =1,2,…K).

$$P(t,\eta) = \max_{\Theta_k} P(t,\eta,\Theta_k) \quad (7)$$

and

$$P(t,1) = \max_{\Theta_k}[\overline{P}(x^T,1,t,1 \mid \Theta_k)] \quad (8)$$

Let $F(t,\eta,\Theta_k)$ denote the starting frame of $P(t,\eta,\Theta_k)$ at level η, we have,

$$F(t,\eta,\Theta_k) = \arg\max_{t_1<t_{\eta-1}<t_2}[P(t_{\eta-1},\eta-1) + \overline{P}(x^T,t_{\eta-1}+1,t,\eta \mid \Theta_k)] + 1 \quad (9)$$

$F(t,\eta,\Theta_k)$ is retained as the backpointer for path finding in Step 3.

The index of the best-matched viseme model is also retained at the end node.

$$\Theta_{best}(t,\eta) = \arg\max_{\Theta_k \in \{\Theta_1,\Theta_2 \cdots \Theta_K\}} P(t,\eta,\Theta_k) \quad (10)$$

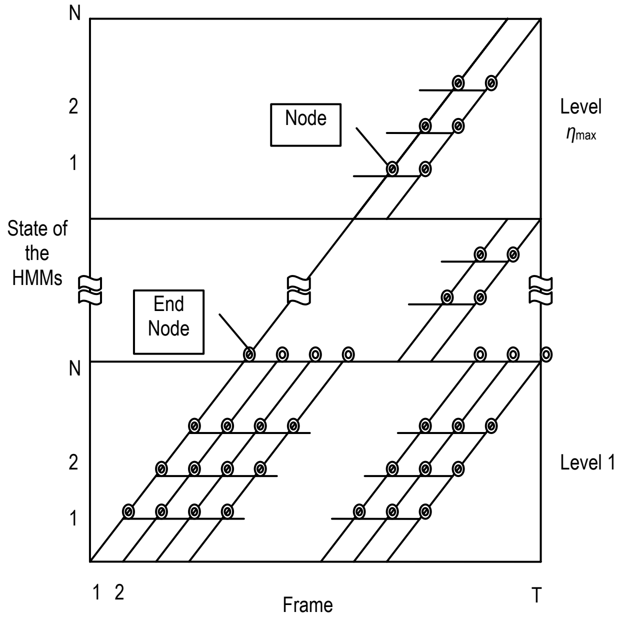

Figure 3. The probability trellis obtained from level building

Step 3. Path finding

With the above probability trellis, backpointers and model indexes, the optimal sequence of the AdaBoost-HMMs for x^T is searched by backtracking the probability trellis. At a possible level number η ($\eta_{\min} \le \eta \le \eta_{\max}$) and frame T being the end frame of level η, the best viseme model at level η — $\Theta_{best}(T,\eta)$, is obtained from (10). Using (9), the optimal starting frame of level η — $F[T,\eta,\Theta_{best}(T,\eta)]$, is computed and also the end frame of level η-1, $F[T,\eta,\Theta_{best}(T,\eta)]-1$, is obtained. The above iteration continues until level 1 at frame 1. The optimal log likelihood scored for x^T by the sequence of the AdaBoost-HMM is $P(T,\eta)$.

For each possible number of levels η ($\eta_{\min} \le \eta \le \eta_{\max}$), a sequence of constitute elements $\Theta^\eta = (\Theta_{t_1}, \Theta_{t_2}, \cdots \Theta_{t_\eta})$ with maximum likelihood $P(x^T \mid \Theta^\eta)$ is decoded. Thus $\eta_{\max} - \eta_{\min} + 1$ sequences with various numbers of viseme models are obtained. Some systems take all these sequences as the possible pattern candidates and further processing is carried out to identify them while some systems only require one sequence. The best sequence is obtained by maximizing over all the possible level numbers as in (11),

$$\Theta_{best} = \underset{\Theta^\eta:\eta_{\min} \le \eta \le \eta_{\max}}{\arg\max} P(x^T \mid \Theta^\eta) \qquad (11)$$

Θ_{best} is the overall optimal sequence of the viseme models. In the experiments conducted in this paper, both the best viseme sequences obtained at different level numbers and the overall best viseme sequence are computed.

4. Experiments

The boosted viseme models are first built in our experiments. According to MPEG-4 Multimedia standards, the basic visual speech elements are clustered into 14 visemes as shown in Table 1.

Table 1. The visemes defined in MPEG-4 Multimedia Standards

Viseme Number	Corresponding Phonemes	Examples
0	none	(silence and relax)
1	p, b, m	push, bike, milk
2	f, v	find, voice
3	T, D	think, that
4	t, d	teach, dog
5	k, g	call, guess
6	tS, dZ, S	check, join, shrine
7	s, z	set, zeal
8	n, l	note, lose
9	r	read
10	A:	jar
11	e	bed
12	I	tip
13	Q	shock
14	U	good

The raw data indicating viseme productions are video frames that reveal the shape of the mouth during articulation, which is shown in Fig. 4. The sampling rate of the video is 50 frames per second. The respective image undergoes edge detection/tracking, normalization, quantization and principal component analysis, are finally converted into sequence of feature vectors. The detailed processing is given in [3][6]. In our experiments, 3-state-128-symbol left-right HMMs are adopted to model the visemes. To make the states of the HMM physically associated with the process of viseme production, the vector sequence indicating viseme production is manually partitioned into three phases:

1.) Initial phase, which starts from a closed mouth or the end of the last viseme production to the beginning of sound production of the viseme investigated.
2.) Articulation phase, which is the time when the sound is produced.
3.) End phase, which is the course when the mouth restore to the relaxed state or transit to the production of the next viseme.

An example of partitioning the viseme production into three phases is given in Fig. 4. The composite HMMs of the AdaBoost-HMM are initialized according to the statistical features of the three phases.

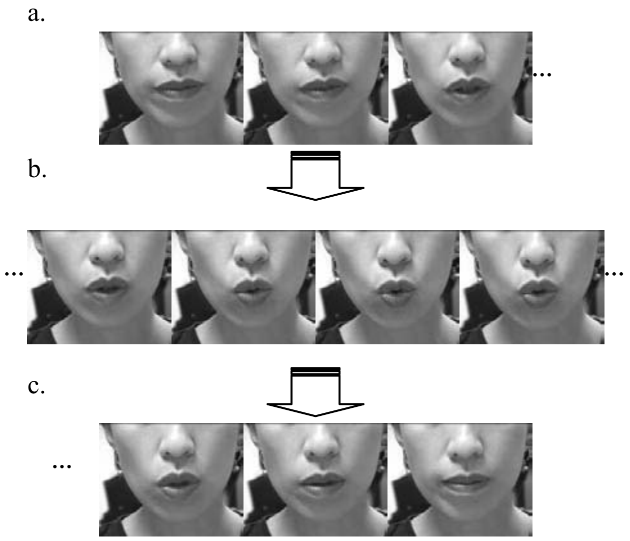

Figure 4. The three phases of viseme production
a) Initial phase
b) Articulation phase
c) End phase

As mentioned in Section 1, the temporal features of the visemes are easily distorted by their context. The training samples of a viseme should thus be chosen to cover the spread-out distribution. In our experiments, 200 samples are drawn for each viseme with 100 for training and the other 100 for testing. For a certain viseme, its samples are video clips that are manually extracted from various word-productions, for example, samples of viseme /A/ include the video clips segmented out of *right* and *hide*. The training process and the HMM boosting process are detailed in [2] and [3]. After boosting, each boosted viseme model comprises 15~20 composite HMMs.

The boosted viseme models are connected by means of level building to model a connected-viseme sequence. Before decoding the sequence, the approximate duration of the visemes is estimated as shown in Table 2.

Table 2. The approximate durations of the visemes

Viseme Number	**Corresponding Phonemes**	**Durations: $d_{\min}$~$d_{\max}$ (sec.)**
1	p, b, m	0.1 ~0.6
2	f, v	0.1~0.7
3	T, D	0.2~0.8
4	t, d	0.1~0.7
5	k, g	0.1~0.7
6	tS, dZ, S	0.3~0.8
7	s, z	0.2~0.8
8	n, l	0.1~0.6
9	r	0.1 ~0.6
10	A:	0.2~0.8
11	e	0.1~0.6
12	I	0.1~0.8
13	Q	0.2~0.8
14	U	0.2~0.8

In our experiments, seven words and simple phrases are chosen to be identified with the proposed approach. For each words/phrases, 100 visually clearly produced samples are drawn. Since different viseme sequences can be decoded if different level number is chosen, the classification rate is computed in two means.

1.) $\boldsymbol{\mu_1}$: At the correct level number, if the correct viseme sequence, e.g. /z/+/U/ for *zoo*, is decoded, a correct classification is made; otherwise, an error occurs. The classification rate computed in this way for the 100 samples is denoted as μ_1 in Table 3.
2.) $\boldsymbol{\mu_2}$: Over all the level numbers, if the correct viseme sequence is decoded, a correct classification is made; otherwise, an error occurs. It means that the decoded viseme sequence satisfies (11). The classification rate computed in this way is denoted as μ_2 in the table.

Table 3. Classification rates of the single-HMMs and AdaBoost-HMMs

Words/ Phrases	**Classification Rate**			
	Single HMM		**Boosted HMM**	
	$\boldsymbol{\mu_1}$	$\boldsymbol{\mu_2}$	$\boldsymbol{\mu_1}$	$\boldsymbol{\mu_2}$
zoo	63%	35%	67%	57%
right	39%	28%	43%	40%
deck	27%	17%	31%	17%
transit	25%	15%	33%	26%
banana	34%	24%	42%	25%
we are	59%	40%	58%	51%
use up	32%	26%	41%	30%

Clearly μ_2 has stricter requirement than μ_1. The classification rates of the connected single-HMMs (traditional approach) are also listed in the table. It is observed that both μ_1 and μ_2 increase with the proposed approach. As a result, level building on the boosted viseme models is an effective means of continuous visual speech processing.

5. Conclusion

AdaBoost-HMM is a kind of multi-HMM classifier that is specially designed to model the stochastic processes with spread-out distribution. Application of the AdaBoost-HMM to model continuous process is an important extension of the method. In this paper, the strategy of level building on the AdaBoost-HMMs is proposed. The strategy is applied to visual words/phrases recognition with some success. This approach can also be applied to other situations especially when the observed data have spread-out distribution, for example, speech recognition, handwriting recognition and speaker identification.

References

[1] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition", Proc. IEEE, Vol. 77, No. 2, pp 257-286, Feb. 1989

[2] Say Wei Foo, Liang Dong, "A boosted multi-HMM classifier for recognition of visual speech elements," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 285 - 288, 2003

[3] Say Wei Foo, Liang Dong, "Recognition of Visual Speech Elements Using Hidden Markov Models," Proceedings of 3rd IEEE Pacific Rim Conf. on Multimedia, pp. 607-614, 2002

[4] Robert E. Schapire, "A brief introduction to boosting," Proc. of the Sixteenth International Joint Conference on Artificial Intelligence, pp. 1401-1405, 1999

[5] M. Tekalp and J. Ostermann, "Face and 2-D mesh animation in MPEG-4," Image Communication J. Aug. 1999

[6] X. Z. Zhang, C. Broun, R. M. Mersereau and M. A. Clements, "Automatic speechreading with applications to human-computer interfaces," Eurasip Journal on Applied Signal Processing, Special Issue on Joint Audio-Visual Speech Processing, pp. 1228-1247, 2002