Cross-speaker viseme mapping using hidden Markov models

Dong, Liang; Foo, Say Wei; Yong, Lian

2003

Dong, L., Foo, S. W., & Yong,L. (2003). Cross-speaker viseme mapping using hidden Markov models. Proceedings of the 4th International Conference on Information, Communications and Signal Processing and the 4th IEEE Pacific-Rim Conference on Multimedia (pp. 1384-1388). Vol.3. Singapore: IEEE.

https://hdl.handle.net/10356/91219

https://doi.org/10.1109/ICICS.2003.1292692

ICICS-PCM 2003
15-18 December 2003
Singapore

Cross-Speaker Viseme Mapping Using Hidden Markov 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 method of mapping visual speech between different speakers is proposed. This approach adopts Hidden Markov Model (HMM) to model the basic visual speech element – viseme. Some mapping terms are applied to associate the state chains decoded for the visemes produced by different speakers. The HMMs configured in this way are trained using the Baum-Welch estimation, and are used to generate new visemes. Experiments are conducted to map the visemes produced by several speakers to a destination speaker. The experimental results show that the proposed approach provides good accuracy and continuity for mapping the visemes.

1. Introduction

Among the possible interactions between different media types, the interaction between audio and video has attracted the attention of the multimedia community in recent years. It has been proven with a series of experiments that the visual speech information may improve the accuracy and robustness of a purely automatic speech recognition (ASR) system [1]~[4][8]. Since 1980s, much work has been carried out in this area [1]~[9]. Some review on the development of visual speech processing can be found in [10] and [11].

In addition to improving the accuracy of speech perception, investigation on visual speech can also be applied to audio-visual mapping, cartoon animation, video games, speaker verification and multimedia telephony for the hearing-impaired. With the fast development of wide-band communication and multimedia technology, transmission of video that indicate the change of the facial area during speech becomes possible. And thus there is increasing need for processing the acquired visual speech signals. In this paper, we focus on a specific research area of visual speech processing – mapping of visual features between different speakers during speech. Research on this area serves to eliminate speaker-dependency of a visual speech recognizer. Some relevant previous work may include the construction of speech-driven models. An early facial model was proposed by Parke [12]. In 1987, a facial model called “Candide” was developed at Linkoping University [13]. In 1990, Welch *et al* studied audio-to-visual mapping using HMM for building speech-driven models [14]. The approach reported in this paper is different from the work mentioned above. Rather than mapping the acoustic speech to the visual domain, the basic visual speech elements – visemes are mapped across different speakers. The computational technique adopted for this purpose is the Hidden Markov Model (HMM). The HMM with mapping terms is first configured according to the temporal features of the visemes, and then trained to associate a source viseme with a target viseme. After adjusting the symbol emission process to guarantee the continuity of the output symbol sequences, new visemes are generated using the HMM thus obtained.

There are few researches have been conducted on cross-speaker viseme mapping in literature. As a result, we only analyzed the performance of the proposed approach but did not compare it with the previous results. The visemes produced by three speakers were mapped to a destination speaker in our experiment. The reproduced visemes can be relatively accurately identified by the viseme models of the destination speaker and they demonstrate good continuity after converted into video frames.

2. Viseme Categorization

In visual speech domain, the basic visual speech element is referred to as viseme. It is the smallest distinguishable visual speech unit. A viseme indicates a short period of lip movement that is repeated in different articulations. Like phonemes which are the basic building blocks of sound of a language, the visemes are the basic constituents for the visual representations of words.

It is commonly agreed that the relationship between phonemes and visemes is a many-to-one mapping. For example, although phonemes /b/, /m/, /p/ are acoustically distinguishable sounds, they are grouped into one viseme category as they are visually confusable, i.e. all are produced by a closed mouth shape. The visemes are grouped according to the similarities between the visual features of phonemes or phoneme-like sound productions. An early viseme grouping was suggested by Binnied *et al* [15]. Viseme grouping in [16] was proposed by analyzing the stimulus-response matrices of the acquired visual signals. The MPEG-4 multimedia standards adopt the same viseme grouping strategy for face animation, in which the visemes are clustered into 14 groups as shown in Table 1.

Table 1. The visemes defined in MPEG-4 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

By modeling and identifying the visemes, it is expected to recognize any word that can be broken up into a sequence of visemes.

3. Viseme Classifier

3.1 Feature extraction from the video

In our experiments, the raw data indicating a viseme is video clip that is sampled at 25 frames per second. The image frame of the video reveals the lip area of the speaker during viseme production, which is shown in Fig. 1. Eleven geometric measures are extracted from the raw image to build a feature vector. These geometric measures give the thickness, position and curvature of the lip. They are chosen as they uniquely determine the lip shape and best characterize the dynamics of lip movement.

The collected feature vectors are put through normalization, principal component analysis (PCA) and quantization. They are finally clustered into groups using *K*-means algorithm. For the experiments conducted in this paper, 128 clusters (code words) are used in the vector database (code book) for each speaker.

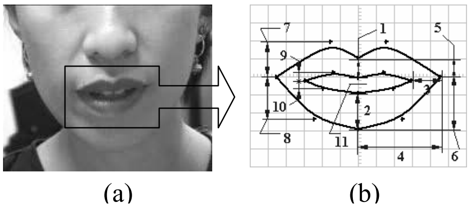

Fig. 1 (a) original image (b) extracted image
1: thickness of the upper bow
2: thickness of the lower bow
3: thickness of the lip corner
4: position of the lip corner
5: position of the upper lip
6: position of the lower bow
7: curvature of the upper-exterior boundary
8: curvature of the lower-exterior boundary
9: curvature of the upper-interior boundary
10: curvature of the lower-interior boundary
11: length of the tongue (if visible)

3.2 Review of HMM principles

The visemes are modeled by the popular probabilistic framework of Hidden Markov Models (HMMs). HMM is basically a quantization of a time process into discrete states. For an *T*-length observation sequence, say $x^T = (x_1, x_2, \cdots x_T)$, it is assumed to be emitted from a sequence of hidden states $s^T = (s_1, s_2, \cdots s_T)$ that is generated by an HMM, where $s_i \in S^N$ and $S^N = \{S_1, S_2, \cdots S_N\}$ is the state set. If the output takes discrete and finite values, say $O^M = \{O_1, O_2, \cdots O_M\}$, an *N*-state-*M*-symbol HMM $\theta(\pi, A, B)$ is determined by the following three components:

1.) The probabilities of the initial states:
$\pi = [\pi_i]_{1\times N} = [P(s_1 = S_i)]_{1\times N} \quad (1 \le i \le N)$, where s_1 is the first state in the state chain.
2.) The state transition matrix :
$A = [a_{ij}]_{N\times N} = [P(s_{t+1} = S_j \mid s_t = S_i)]_{N\times N} \quad (1 \le i, j \le N)$, where s_{t+1} and s_t are the t+1-th and the t-th states.
3.) The symbol emission matrix :
$B = [b_{ij}]_{N\times M} = [P(O_j \mid S_i)]_{N\times M} \quad (1 \le i \le N, 1 \le j \le M)$.

If the above parameters are properly trained, the HMM can well model the temporal features of the observation sequence. One of the popular approaches of training the HMM is the Baum-Welch estimation. For $\theta(\pi, A, B)$, we define the forward variables $\alpha_t(i) = P(x_1, x_2, \cdots x_t, s_t = S_i \mid \theta)$ and backward variables $\beta_t(i) = P(x_{t+1}, x_{t+2}, \cdots x_T \mid s_t = S_i, \theta)$ for x^T, the parameters of the HMM are then estimated through the following Expectation-Maximization (EM) recursion [17].

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j(x_{t+1}) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i) \beta_{t+1}(j)} \quad (1)$$

$$\bar{b}_j(O_m) = \frac{\sum_{\substack{t=1 \\ s.t. x_t = O_m}}^{T} \alpha_t(j) \beta_t(j)}{\sum_{t=1}^{T} \alpha_t(j) \beta_t(j)} \quad (2)$$

where O_m is the m-th symbol in the symbol set. After a sufficient number of training epochs, a local maximum point of the likelihood $P = P(x^T \mid \theta)$ is attained.

3.3 HMM modeling of visemes

Our study on human speaking habit reveals that while a speaker is articulating single phoneme (or producing a viseme), the lip can be assumed to experience three phases. The first is the initial phase, which is the course from the mouth is closed and relaxed to get ready to make the sound. During this phase, there is usually no sound articulated and the lip is characterized with sharp changes. The next is the articulation phase, which is the course that the lip poses to make the sound until the sound is made. The change of the lip shape during this phase is not so violent as the previous one and there is usually short stable moment in the phase. The third is the end phase. The mouth will restore from the articulation state to relaxed state. Fig. 2 illustrates the three phases and the lip shapes within each of them while the speaker is articulating the phoneme /u/.

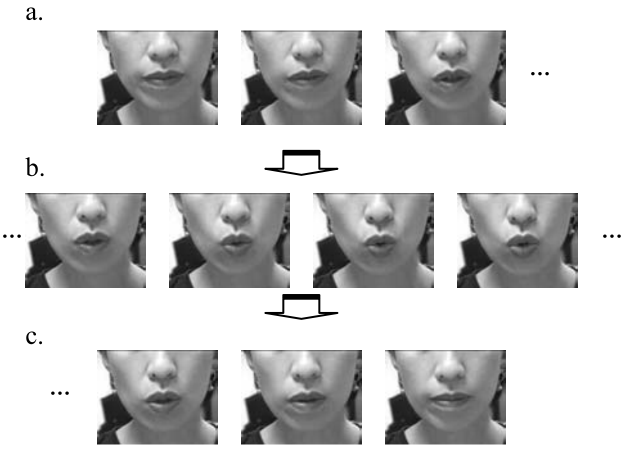

Fig. 2. The three phases during viseme production (a) Initial phase (b) Articulation phase (c) End phase

The HMM used for modeling viseme is the three-state left-right HMM as shown in Fig. 3. The states of the HMM are denoted as the initial state, articulation state and end state.

The initial values of the three states are configured according to the statistical features of the three phases of viseme production, i.e. setting the symbol emission probabilities of the initial state, articulation state and end state approximate to that of the initial phase, articulation phase and end phase [18][19].

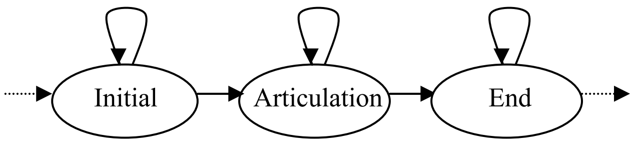

Fig. 3. The three-state left-right HMM framework used for modeling viseme

4. Viseme Mapping

The facial features are different from person to person. The visual speech such as viseme thus shows strong speaker-dependency. Our task is to map the viseme produced by one speaker (source speaker) to the same viseme produced by another speaker (destination speaker). For ease of subsequent explanation, we refer the viseme models (HMMs) of the source speaker as the source models and the viseme models of the destination speaker as the destination models.

4.1 Mapping terms of the HMMs

Assume that $\{O_1^s, O_2^s, \cdots O_M^s\}$ and $\{S_1^s, S_2^s, \cdots S_N^s\}$ are the symbol set and state set for the source models, and $\{O_1^d, O_2^d, \cdots O_{M'}^d\}$ and $\{S_1^d, S_2^d, \cdots S_{N'}^d\}$ are the symbol set and state set for the destination models, where N is the state number and M is the symbol number of the source model, and N' and M' are those of the destination model. For the k-th viseme as illustrated in Table 1, a source model θ_k^s and a destination model θ_k^d are configured with the approach as mentioned in 3.3. θ_k^s is then trained using the Baum-Welch estimation.

Given a training sample of viseme k (k=1,2,…14) produced by the source speaker $-(o_1^s, o_2^s, \cdots o_T^s)$ (it is referred to as the source sequence), where o_i^s denotes the i-th observed symbol in the sequence, the optimal state chain $(s_1^s, s_2^s \cdots s_T^s)$ is decoded using the Viterbi search [17], where s_i^s stands for the i-th state in the decoded state chain. An observation sequence $(o_1^d, o_2^d, \cdots o_T^d)$ (destination sequence) of T-length is selected from the training samples of viseme k of the destination speaker. The optimal state chain $(s_1^d, s_2^d \cdots s_T^d)$ for the destination speaker is also decoded using the Viterbi search, where o_i^d and s_i^d have the same meaning as in the source model. The state chains of the source model and destination model are associated with each other by the mapping terms $c(1)$, $c(2)$, $c(3)$ … as shown in Fig. 4.

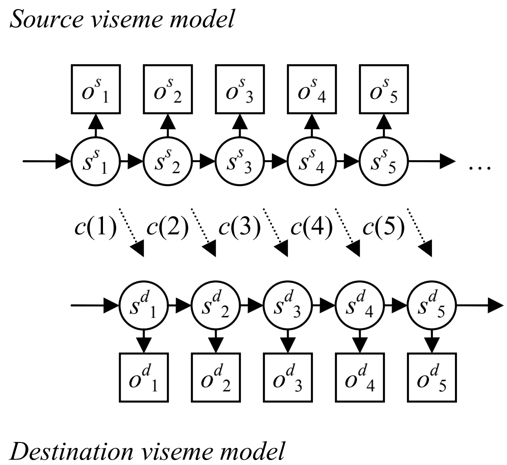
Source viseme model
o^s_1 o^s_2 o^s_3 o^s_4 o^s_5
s^s_1 s^s_2 s^s_3 s^s_4 s^s_5 …
$c(1)$ $c(2)$ $c(3)$ $c(4)$ $c(5)$
s^d_1 s^d_2 s^d_3 s^d_4 s^d_5 …
o^d_1 o^d_2 o^d_3 o^d_4 o^d_5
Destination viseme model

Fig. 4 Mapping of the source model to the destination model

These mapping terms come from the mapping matrix $C = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,N'} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,N'} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N,1} & c_{N,2} & \cdots & c_{N,N'} \end{bmatrix}$, where $c_{i,j} = P(S^d_j \mid S^s_i)$. The state chain $(s^d_1, s^d_2 \cdots s^d_T)$ can be looked as the symbols emitted by $(s^s_1, s^s_2 \cdots s^s_T)$. By combining o^s_t and $s^d{}_t$ as the t-th observation symbol of the source sequence, training of the source model thus becomes the process of adjusting the state transition matrix and the symbol emission matrix to maximize the likelihood $P(s^d_1 + o^s_1, s^d_2 + o^s_2, \cdots s^d_T + o^s_T \mid \theta^s_k)$. A uniform distribution is assigned to the initial values of the coefficients in C.

$$c_{ij} = 1/N \quad (i = 1,2,\cdots N, j = 1,2,\cdots N') \tag{3}$$

The Baum-Welch estimation is carried out again for this purpose. After a sufficient number of EM iterations, the maximum-likelihood source model for $d_k - \theta^d_k$, is obtained. The mapping terms are the "state emission" probabilities $P(S^d_j \mid S^s_i)$ (i=1,2,…N, j=1,2,… N').

5.2 Viseme generation

With the source model, a viseme produced by the source speaker is mapped to a destination sequence with the following steps.

1.) Assume that $y = (y^s_1, y^s_2, \cdots y^s_T)$ is a source sequence indicating the production of viseme k. The optimal state chain $(s^s_1, s^s_2 \cdots s^s_T)$ of the source model is decoded using the Viterbi search.

2.) A state chain $(s^d_1, s^d_2 \cdots s^d_T)$ of the destination model, together with its likelihood, is generated using the mapping terms.

3.) An observation sequence $y' = (y^s_1, y^s_2, \cdots y^s_T)$ is then generated by the state chain $(s^d_1, s^d_2 \cdots s^d_T)$.

The mapping of the source sequence to the destination sequence is thus realized. However, the above approach does not consider the continuity of the generated destination sequence. The lip shape may change abruptly in the sequence. To solve this problem, some restricts are added to the destination model.

For the destination model θ^d_k, at time t, if the decoded state s_t= S_i (i=1,2,…N') and the symbol obtained at time t-1 is $o_{t\text{-}1}$, the symbol emission coefficient is changed as in (4).

$$b'(O_j \mid S_i) = \frac{b(O_j \mid S_i) e^{-E(o_{t-1}, O_j)}}{\mu} \tag{4}$$

where $E(o_{t-1}, O_j)$ is the Euclidean distance between $o_{t\text{-}1}$ and O_j and μ is a normalization factor to make $b'(O_j \mid S_i)$ $(j = 1,2,\cdots M)$ a distribution. With such modification, the symbol obtained at time t – o_t, is more likely to be close to $o_{t\text{-}1}$ and thus the continuity of the destination viseme is improved.

5. Experiments

Experiments are conducted to test the performance of the proposed strategy. The visemes produced by three speakers are mapped to a destination speaker. The accuracy of such mapping is first studied. We define that, if the mapped destination viseme can be correctly identified by a viseme model, a correct classification is made; otherwise an error occurs. The average recognition rates are listed in Table 2. For example, θ_1 is the average recognition rate of the viseme samples of the destination speaker recognized by the viseme models of the destination speaker, and θ_2 denotes the mapped visemes (Speaker 1 to the destination speaker) recognized by the viseme models of the destination speaker. The results show that θ_1 is close to θ_2. It indicates the generated visemes are similar to the actual visemes produced by the destination speaker.

Table 2. The average recognition rates scored for the actual viseme samples and the mapped visemes

θ_1/θ_2	Speaker 1	Speaker 2	Speaker 3
Destination Speaker	0.80/0.75	0.80/0.75	0.80/0.65

The continuity of the obtained sequence is also investigated. The standard we applied is relatively subjective. We map the vector sequence indicating viseme production back to the video frames. The playback of the video frames shows that the movement of the lip is most of the time stable and reasonable.

6. Conclusion

The strategy proposed in this paper is a simple method of mapping visemes between two speakers. By training some mapping terms for the HMM, a viseme produced by the source speaker can be mapped to the destination speaker. The state chain of the source speaker is loosely associated with that of the destination speaker. The mapped visemes can thus be generated with great flexibility. Experiments show that the obtained visemes can be accurately identified by the models of the destination speaker. And by adding some restrictions to the symbol emission coefficients of the viseme model, the continuity of the mapped viseme is also guaranteed.

References

[1] E. D. Petajan, "Automatic lipreading to enhance speech recognition," Ph.D thesis, University of Illinois at Urbana-Champaign, 1984

[2] A. Adjoudani and C. Benoit, "On the Integration of Auditory and Visual Parameters in an HMM-based ASR," Speechreading by Humans and Machines, Edited by D. G. Stork and M. E. Hennecke, NATO ASI Series, pp. 461-472, 1996

[3] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," IEEE Trans. on Speech and Audio Processing, Vol. 4 Issue 5, pp. 337 -351, Sep 1996

[4] M. Tomlinson, M. Russell and N. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 821-824, 1996

[5] A. J. Goldschen, "Continuous automatic speech recognition by lipreading," Ph.D dissertation, George Washington University, Washington, Sep. 1993

[6] B. P. Yuhas, M. H. Goldstein and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," IEEE Communication Magzine, pp. 65-71, 1989

[7] C. Bregler and S. Omohundro, "Nonlinear manifold learning for visual speech recognition," IEEE International Conference on Computer Vision, pp. 494-499, 1995

[8] P. Silsbee and A. Bovik, "Computer Lipreading for Improved Accuracy in Automatic Speech Recognition," IEEE Trans. Speech and Audio Processing, Vol. 4, No. 5, pp. 337-351, 1996

[9] D. G. Stork and H. L. Lu, "Speechreading by Boltzmann zippers," Machines that learn, Snowbird, UT, 1996

[10] Tsuhan Chen, "Audiovisual Speech Processing," IEEE Signal Processing Magazine, Jan. 2001

[11] D. G. Stork and M. E. Hennecke, "Speechreading: An overview of image processing, feature extraction, sensory integration and pattern recognition techniques," The Second Int. Conf. on Automatic Face Gesture Recognition, pp. xvi-xxvi, Oct. 1996

[12] F. I. Parke, "Parameterized models for facial animation," IEEE Computer Graphics and Applications, pp. 61-68, Nov. 1982.

[13] M. Rydfalk, "CANDIDE: A parameterized face," Linkoping University, Sweden, Report LiTH-ISY-I-0866, Oct. 1987

[14] W. J. Welsh, A. D. Simon, R. A. Hutchinson and S. Searby, "A speech-driven 'talking-head' in real time," Proceedings of Picture Coding Symposium, pp. 7.6-1 - 7.6-2, Cambridge USA, 1990

[15] C. Binnie, A. Montgomery and P. Jackson, "Auditory and visual contributions to the perception of consonants," Journal of Speech Hearing and Research, Vol. 17, pp. 619-630, 1974

[16] E. Owens and B. Blazek, "Visemes Observed by Hearing Impaired and Normal Hearing Adult Viewers," Journal of Speech Hearing and Research, Vol 28, pp. 381-393, 1985

[17] 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

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

[19] Say Wei Foo, Yong Lian, Liang Dong, "A two-channel training algorithm for hidden markov model to identify visual speech elements", Intel. Symposium on Circuits and Systems, (ISCAS '03), Vol. 2, pp. 572 -575, 2003

[20] Say Wei Foo, Liang Dong, "A boosted multi-HMM classifier for recognition of visual speech elements", IEEE Intel. Conf. on Acoustics, Speech, and Signal Processing, (ICASSP '03). 2003, Vol. 2, pp. 285 -288, 2003