ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

FACE SYNTHESIS DRIVEN BY AUDIO SPEECH INPUT BASED ON HMMS

Ling SUN

Department of Electronic Engineering and Information Science,
University of Science and Technology of China, Hefei
sunling@mail.ustc.edu.cn

Wei LAI

Department of Electronic Engineering and Information Science,
University of Science and Technology of China, Hefei
lvictor@mail.ustc.edu.cn

Ren-Hua WANG

Department of Electronic Engineering and Information Science,
University of Science and Technology of China, Hefei
rhw@ustc.edu.cn

ABSTRACT

In this paper, a HMM-based visual speech system driven by audio speech input is designed to render a face model while synchronous audio is played. Compared to many methods adopted by other researchers, there is much difference between our approach and theirs. We first train the models for every final and initial in mandarin. In this process, a large quantity of audio training data under different surroundings and spoken by different people are used. Then, the recorded synchronous audio-visual speech data are used to make the trained models more adaptive to our specific announcer. Such models are more robust in synthesis phase and satisfying performance can be achieved even when input audio speech is degraded by noises.

1. INTRODUCTION

In recent years, audio-driven facial system animation has become an interesting and evolving technique for human-computer interaction. Such kind of visual speech systems which are also called talking-heads are now practical due to the ever-increasing computing power and the performance of computer hardware. Many applications such as, for example, virtual operators or customer care/help desks on the Web require realistic "virtual agents" that look reasonably human and speak naturally. Lip-synchronization and co-articulation are essential in these applications. Based on an incoming audio stream, a face image is animated with full lip synchronization. Nowadays, many methods and tools are applied to these applications, including vector quantization (VQ), Hidden Markov Model (HMM), Artificial Neural Network etc. [1-5].

More recently, many approaches based on HMMs are proposed [6-9]. Our idea of lip parameter generation algorithm is similar to that of the work done by other researchers. However, there is much difference between our approach and theirs. In training phrase, the speech data sequences of a number of phonemes in mandarin are used to train the phonemes HMMs. Every phoneme has its corresponding lip parameters which are retrieved automatically by means of auto-lip-detecting. After the training phrase, a complete look-up table for mapping from phoneme to lip is built. In synthesis phrase, the audio stream is segmented by a speech recognition module. Using the mapping look-up table, we retrieve the lip parameters associated with the HMM state. The face image is animated by these parameters after interpolating between three basic face models. Our experiment results show that our proposed synthesis and synchronization method generates natural and synchronous talking faces from speech inputs.

This paper is organized as follows. In section 2, we describe the structure of out system. In section 3, face modeling is introduced. Then we present the structure of the HMMs we use in our system and the detailed realization method in section 4. In section 5, the system construction and experiment result are discussed. At last, we show our conclusion in section 6.

2. SYSTEM OVERVIEW

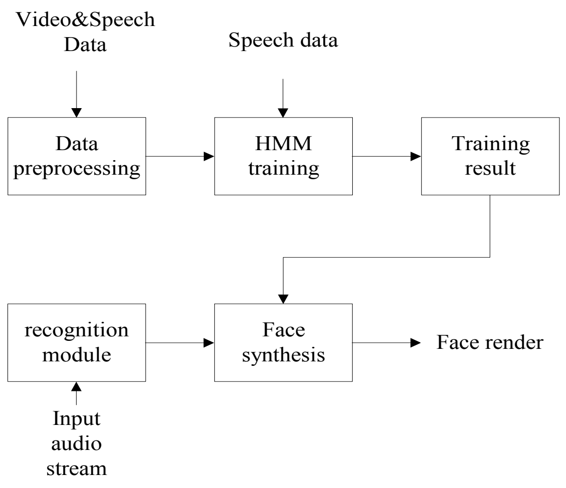


Fig.1 The overview of the visual speech system

Figure 1 shows a block diagram of our visual speech system. In training phase, we train respective model for each final and initial in mandarin. The training data used in that process are collected under different surroundings and they are uttered by different people. Speaker-independent models can be obtained through that training. Then we apply our recorded synchronous audio-visual speech data to these trained models to train them again. The aim is to make the models more adaptive to our announcer, including on the aspect of acoustic and facial characteristics as well, after which a complete look-up table for mapping from phoneme to lip is built. In synthesis phase, the input audio speech is analyzed and converted to state sequence. Lip parameters are retrieved through searching corresponding parameters associated with the state sequence. Then the face model is effectively controlled by the received lip parameters. Finally, the synthesized face is rendered while synchronized audio is played.

3. FACE MODELING AND LIP PARAMETERS

To generate a realistic synthesized face, a generic face model is built. Then, the model is manually adjusted. At last, another two models are built based on the first one with mouth width and height reaching max respectively. During synthesis phrase, we operate these three models according to the lip parameters especially width and height retrieved from the neural network. Figure 2 shows the generic face model.

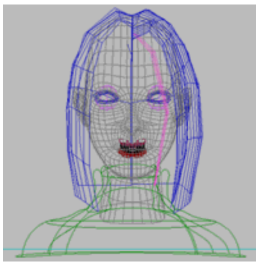


Fig. 2 Face model

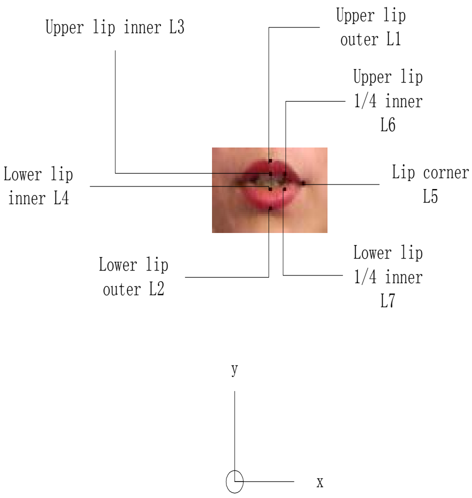


Fig 3 Lip controlling points

Figure 3 shows the controlling points of the frontal lip. Totally seven points are adopted. It is no other than the relative position of these seven points that forms the lip parameters we adopt. For example, the number of pixels in y direction between L3 and L4 constitutes the inner mouth height. Table 1 shows the frontal lip parameters.

Lip parameter	Retrieving method	Direction
P_1	$L_3 - L_4$	Y
P_2	$L_1 - L_3$	Y
P_3	$L_4 - L_2$	Y
P_4	$L_3 - L_6$	Y
P_5	$L_7 - L_4$	Y
P_6	$L_5 - L_3$	X
P_7	$L_3 - L_5$	Y

Tab. 1 Lip parameters

4. MODELS TOPOLOGY AND TRAINING

In mapping from speech to lip movement, HMMs are widely adopted. Generally, HMM-based approaches are combined with Viterbi method to obtain the mapping between them. In this section, a detailed description of our modeling and training process is presented.

4.1 Topology of Models

Considering the characteristics of finals and initials in mandarin, different topology is adopted for different finals and initials in our modeling process. The difference lies in that the number of states in the models differs with phonemes. In fact, such difference is mainly due to the duration when those phonemes are normally pronounced. In all of our models, skip between states is not permitted. Figure 4 shows the topology of final 'a'.

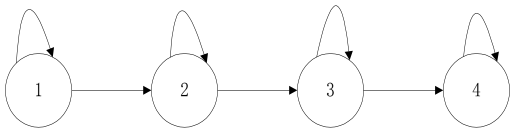


Fig.4 Model Topology of 'a'

4.2 Data Preparation

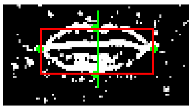
Training recognition and synthesis systems require high quality audio-visual database. Unfortunately, current audio-visual database are much smaller than needed. Moreover, no standard methodology for the collection and organization of audio-visual databases exists [3]. To get effective models, sufficient training data are necessary. As far as audio-visual speech synthesis is concerned, the training data includes visual component. To us, the preprocessing of the synchronous audio-visual data is very time consuming. To reduce the time spent on it, we must reduce the data quantity at the same time. But lack of data may results in insufficient training of models. To solve this contradiction, we first train respective phoneme model for every initial and final. Based on the pre-trained models, the recorded synchronous audio-visual speech data are used to adapt these models to the announcer's characteristic. By doing so, we can reduce the quantity of training data in the second training phase. Meanwhile, these models are more robust in synthesis phase.

The database used in our experiment includes all the initials and finals in mandarin. Each initial or final is placed in different positions in a two-word, three-word or four-word phrase. Of course, they also emerge in separate words. All the words and phrases are uttered by a female speaker. Speaker's lip is made-up in red. The PAL video (30fps) sequence is used as the training data. The sampling rate of the audio is 16-khz with 16-bit resolution.

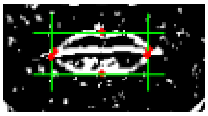
4.3 Data Preprocessing

Mel-cepstral analysis (100fps) was performed to get the acoustic training data. The acoustic observations vector was comprised of 12 Mel-cepstral features, an energy component, and the corresponding 13 delta and 13 delta-delta features.

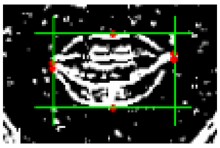
As to the lip parameters, automatic lip detection was adopted. We first retrieved all the frames from video sequence. Then edge detection and edge enhancement were carried out. The first frame was marked manually to locate the key points of the lip. From the second frame on, according to the reference supplied by the first one, automatic lip detection was performed so on and so forth. Figure 5 shows an example of our detection.



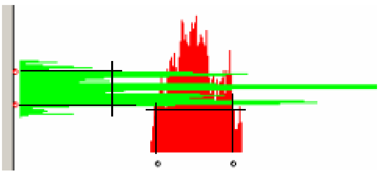
A. Manual marking result



B. Automatic detection



C. Automatic detection when mouth is wide



D. Pixel histogram

Fig. 5 Example of automatic lip detection

The detection step is described as follows.

(1) Based on the detection result of the former frame, a rectangle region around the lip is plotted. Consequently, the region to be searched in is gotten by extending certain number of pixels to left, right, upper and lower.

(2) In the search region, the luminousness of every vertical line is calculated to get a pixel histogram from left to right which is shown in the red part of Fig. 5 D. Then a threshold is defined to judge the boundary of the left and right mouth, i.e., the two green vertical lines shown in Fig. 5 B.

(3) In the search region, two points which have the max sum of luminousness plus the rounding points are selected out from the green lines respectively, whereupon lip corners are located just like the red two points shown in Fig. 5 B.

(4) Also in the search region, the luminousness of every horizontal line is calculated to get a pixel histogram from upper to lower which is shown in the green part of Fig. 5 D. Then a threshold is defined to judge the boundary of the upper and lower mouth, i.e., the two green horizontal lines shown in Fig. 5 B.

(5) According to the midline of lip corners and the boundary of upper and lower lip, the middle points of middle upper and lower lip are located.

(6) Based on the key points detected above, the lip parameters are calculated out.

(7) The same steps are performed on the subsequent frames.

4.4 Model Training

As mentioned earlier, we need to train phoneme models at first. A large quantity of speech data were used in the training. Speaker-independent models were obtained through this process. Based on these pre-trained models, the recorded synchronous audio-visual speech data were used to adapt these models to the announcer's characteristics.

The Viterbi method was adopted in our training. The Viterbi alignment assigns an input frame to the optimal audio HMM state by maximizing a likelihood of input audio parameters. The HMM-Viterbi method is composed of two processes: a decoding process that converts input audio parameters to a most likely audio HMM states by the Viterbi alignment and a table look-up process that converts an HMM state to corresponding facial parameters [9].

Because the audio parameters sequences can be converted to HMM state sequences, the synchronous lip parameter sequences can be segmented per HMM state. When lip parameters are assigned to the same HMM state, they are averaged to generate the representative lip parameter of the state. Our training and synthesis algorithms are as follows.

<Training>

1. Train the phoneme HMMs using the speech data collected under different surroundings and spoken by different people.
2. Train the models again using the synchronous audio-visual speech data on the basis of the pre-trained models.
3. Align audio parameters sequences of training data into HMM state sequences by the forced Viterbi alignment.
4. Take the average of the synchronous lip parameters for all frames associated with the same HMM state for generating a look-up table.

<Synthesis>

1. Align an input audio parameter sequence into an HMM state sequence through Viterbi alignment.
2. Retrieve the output lip parameters associated with the HMM state using the look-up table of lip parameter per HMM state.

5. SYSTEM CONSTRUCTION AND EXPERIMENT

We built a visual speech system according to the structure shown in figure 1. We experimented with many wav forms. With our synchronization strategy, the system can synthesize natural and speech-synchronized face. Figure 5 shows parameter p_1 temporal change when sentence “电脑爱好者” is uttered, where the solid line indicates the values of the original face and the dotted one indicates those of the synthesized face.

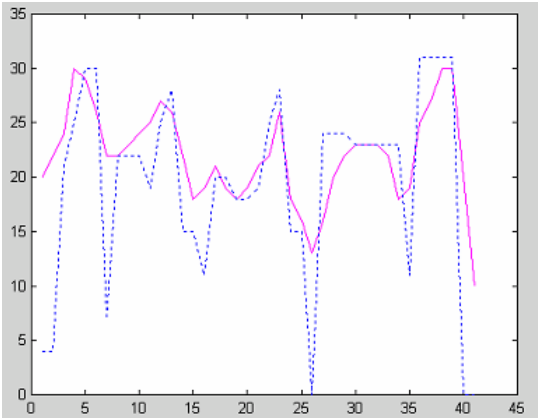


Fig. 5 The comparison between parameter p_1 of the original face and synthesized face

6. CONCLUSION

A visual speech synthesizer driven by audio input based on HMM is designed in this paper. We first train respective acoustic model for every initial and final. Based on the pre-trained models, the recorded synchronous audio-visual speech data are used to adapt these models to the announcer’s characteristics. After generating the look-up table using HMM-Viterbi method, we retrieve the lip parameters for rendering during synthesis phase. Through our control on synch our face rendering and audio playing can achieve good performance on the aspect of synch.

HMM has achieved great success in speech recognition. In recent audio-visual speech research, it also has been widely used. More recently, there is a trend that MPEG-4 is used as the interface between speech and video by more and more researchers in this field. Moreover, face animation parameters defined in MPEG-4 are applied to the control of face models in many multimodal techniques, which makes the control more flexible and convenient.

Currently, we have been combining the face synthesis defined in MPEG-4 with our work and already achieved some fruits. In our future work, the emphasis will be placed on the models refining and trying different synthesis strategies.

7. REFERENCES

[1] D.G.Stock and M.E Hennecks, *Speechreading by Humans and Machine,* Springer-Verlag, Berlin, 1996

[2] F.I Parker and K.Waters, *Computer Facial Animation,* ch.8 A L Peters, Wellesley, MA, 1996

[3] D.R.Hill, A. Pearce, B.Wyvill, “Animation speech: an automated approach using speech synthesized by rule,” *The Visual Computer,* 3, pp.277-289, 1998

[4] M.M.Cohen and D.W.Massaro, *“Modelilng coarticulation in synthetic visual speech,”* In N.M,.Thalmann and D.Thalmann, eds., Models and Techniques in Computer animation, pp.139-156, Springer-Verlag, Tokoyo, 1993

[5] K.Water and T.M.Levergood, “DECface: an automatic lip-synchronization algorithm for synthesis faces,” *Technica Report CRL 93/4,* DEC Cambridge Research Laboratory, Cambridge, MA, Sep, 1993

[6] Masuko, T.; Kobayashi, T.; Tamura, M.; Masubuchi, J.; Tokuda, K. , “Text-to-visual speech synthesis based on parameter generation from HMM,” Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, Volume: 6, Page(s): 3745 -3748 vol.6, 1998

[7] Eli Yamamoto, Satoshi Nakamura, Kiyohiro Shikano, “Speech-to-lip movement synthesis based on the EM algorithm using audio-visual HMMs,” Multimedia Signal Processing, 1998 IEEE Second Workshop on, 1998 Page(s): 53 -58

[8] Williams, J.J.; Katsaggelos, A.K., Randolph, M.A., “*A hidden Markov model based visual speech synthesize,*” Acoustics, Speech, and Signal Processing, 2000, ICASSP '00, Proceedings. 2000 IEEE International Conference on, Volume: 4, 2000 Page(s): 2393 -2396 vol.4

[9] *Nakamura, S.*, “HMM-based transmodal mapping from audio speech to talking faces ”, *Neural Networks for Signal Processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop , Volume: 1 , 2000 Page(s): 33 -42 vol.1*

[10] S. Morishima, “*Real-time talking head driven by voice and its application to communication and entertainment,*” AVSP98

[11] Yuhas, B.P.; Goldstein, M.H., Jr.; Sejnowski, T.J. “*Integration of acoustic and visual speech signals using neural networks*” IEEE Communications Magazine, Volume: 27 Issue: 11, Nov. 1989 Page(s): 65 -71