AUDIO TO VISUAL SIGNAL MAPPINGS WITH HMM

Wentao Liu, Baocai Yin, Xibin Jia, Dehui Kong

Multimedia and Intelligent Software Technology, Beijing Municipal Key Laboratory
Beijing University of Technology, Beijing 100022, China
xiaoniuge@sina.com.cn

ABSTRACT

There has been a large amount of research on speech driven face animation. Particularly, recently research efforts have been demonstrated that the Hidden Markov Model techniques could achieve a high level of success in the filed of audio/visual mapping without language information. In this paper, firstly a linear model based facial representation method was applied, which extract face feature as a global feature. Secondly a HMM based method was presented, which include a two-level frame to promote the audio-visual mapping result.

1. INTRODUCTION

The automatically audio to visual synthesis problem has intrigued researchers for years. The key issue of this research topic is to find a mapping from audio information to visual information. The goal of audio to visual mapping is to produce accurate, synchronized animations of lip movements driven by an incoming audio stream. By accurate animation we mean that the synchronization between the spoken language utterances and the animation is tight and that the synchronism with movements is as natural as possible.

Some approaches are based on great samples of lip motion. Video Rewrite [4] technique recognizes different phonemes from the input audio signal. Animation is generated by re-ordering the captured video frames which share similar phonemes as in the training video.

Chen and Rao [5] train HMMs to automatically segment the audio vectors of isolated words into state sequence. The visual feature vectors are estimated by the estimation function that is derived for each particular state. Alternatively, the state probabilities at each time stamp are evaluated. The overall visual feature of each time stamp is calculated by linear combination of the visual feature estimated for each state, which is weighted by the corresponding state probability.

Voice Puppetry [3] first trains a HMM to learn a facial dynamical model using an entropy minimization algorithm. The associated audio track is then incorporated to re-train the HMM for estimating the visual state sequence for an audio track. The most probable visual feature trajectory corresponding to the visual state sequence can be calculated by a closed form solution.

Another approach is Time Delay Neural Networks (TDNNs), which uses ordinary time delays to perform temporal processing. Curinga et al. [8] train a TDNN to map LPC cepstrum coefficients of speech signal to lip animation parameters.

Some approaches attempt to generate instantaneous lip shapes directly from the audio signals using vector quantization, Gaussian mixture model, or artificial neural networks (ANN). Vector quantization [9] is a classification based conversion approach. The audio feature is first classified into one of a number of classes. Each audio class is then mapped into a corresponding visual feature. Though it is computationally efficient, vector quantization approach often leads to discontinuous mapping results. Gaussian mixture approach [10] models the joint probability distribution of the audiovisual vectors as a Gaussian mixture. Each Gaussian mixture component generates an estimation of the visual feature given an audio feature. The estimations are then non-linearly weighted to produce the final visual estimation. Gaussian mixture approach is able to produce smoother results than vector quantization approach.

Another problem in speech driven face animation is facial feature representation. Conventional method emphasized on feature point extracting and tracking [1] [3]. A linear model based global visual feature extracting and visual feature classifying methods are employed in this paper. Based on the visual classes we build hidden Markov model to generate instantaneous lip shapes directly from the audio signals, and a two-level HMM frame is adopted in this work which can enhance the mapping accuracy.

The paper is organized as follows: the second section introduces background techniques. The third section presents the linear model based visual feature extraction. The fourth section is about the audio-visual mapping based on the HMM. The last section ends up with a conclusion and future work.

2. BACKGROUND

2.1 Hidden Markov Model

Hidden Markov Models (HMMs) are widely used in pattern recognition applications, most notably speech

recognition. The technique is about a class of stochastic processes that are usually presented as having a finite set of states, but which, in another sense, may have an infinite number of states. These processes are known variously as *Hidden Markov Models* (HMMs), functions of a Markov Chain, or stochastic finite automata. A process in the HMM class can be described as a finite-state Markov Chain with a memoryless output process which produces symbols in a finite alphabet.

2.2 Linear Face Model

A linear face model [2] is defined as a class for which novel faces or lip shapes can be represented as a linear combination of a sufficiently small number of prototypical face image. The key of this approach is the representation of an object in terms of a shape vector and a texture vector. The separation of 2D shape and texture information in face images requires correspondence to be established to a single reference face.

The linear model is defined as the set of images I^{model}, parameterized by $\mathbf{B} = \{b_0, b_1, ..., b_N\}$, $\mathbf{C} = \{c_0, c_1, ..., c_N\}$,

$$S = \sum_i c_i S_i, \quad T = \sum_i b_i T_i \qquad (1)$$

Summation $\sum_i c_i S_i$ constrains the shape of every modeled image to be a linear combination of the prototype shapes. Summation $\sum_i b_i T_i$ constrains the texture of every modeled image to be a combination of prototype textures. The coefficients vector **c** is the feature vector used for classification of lip shape in this work.

3. LINEAR MODEL BASED VISUAL FEATURE REPRESENTATION

3.1 Linear model based lip feature

In conventional approaches, to obtain facial articulation data, computer vision techniques have to be applied for the information extracting from a video of an unmarked face. Researchers track many individual feature points on the face [1], such as corners and creases of the lips. We apply linear face model [2] to extract the global feature of lips instead of a set of feature points.

$$I^{model} \circ \left(\sum_{i=0}^{N} c_i S_i \right) = \sum_{j=0}^{N} b_j T_j \qquad (2)$$

I^{model} is the linear model from prototypical lip image, $\{c_i\}$ is the shape coefficient of one lip image, and $\{b_i\}$ is the texture coefficient of it. Any training image can be expressed by the two sets of coefficients. In another word, we can use linear coefficients as the visual feature, especially shape coefficient.

When training, we need to obtain training data set directly from given video. Hundreds of lip image are randomly selected from the source video using and a linear face model is learned by Bootstrapping algorithm [11]. The training frames of source video are then projected to linear model space, and two coefficient vectors (**B** and **C**) are obtained.

3.2 Lip Shape Clustering

The linear face model provides a accurate feature representation of shape vector-$\mathbf{C} = \{c_0, c_1, ..., c_N\}$ and texture vector-$\mathbf{B} = \{b_0, b_1, ..., b_N\}$, which locates a lip image in lip space accurately. Because there is little difference in texture for special man's lip, we only concern the shape vector for clustering the lip images.

For our method, we build audio-visual mapping from vocal tract to a facial set. The facial set is no explicit relationship to phoneme. First all training images are expressed by the linear model. We then employ SOFM [12] for clustering. We get 40 centroid vectors from coefficients of training images. Combining the centroid vector (shape vector) with principal texture vector by linear model, 40 prototype lip images for visual class is synthesized (Figure 1). The synthesized images from linear model compose of the facial set for audio-visual mapping.

Figure 1 Synthesized center image from linear model

The synthesized center images (parameterized by **B** and **C**) are clustered from large number of source lip images, and we use them to denote the 40 visual classes.

4.TWO-LEVEL MODEL FOR MAPPING STRATEGY

4.1 Fundamental Audio-visual mapping

The key issue of speech driven face animation always is focused on audio/visual mapping. Previous approaches of audio-visual mapping employed a remapping model [3]. The states are first mapped into regions of facial configuration space, and each state explains one frame of training sequence. Then the facial HMM is reused to estimate another set of output probabilities. This associates audio features to each facial state, resulting in a vocal

HMM.

In this paper, we express the content of a frame as a mixture of several states. That's mean one model corresponds to a visual class which is clustered with linear model features. We all know that the mapping from vocal to lip shape is many-to-many. Many sounds are compatible with one lip shape and a same utterance may lead to different lip shapes. The best way to deal with the mapping problem is to make full use of context information forward and backward in time. Here we use context information of audio to affirm a frame of face image. Five consecutive audio frames that include current frame are considered, 2 backward frames and 2 forward frames. The time of the frame we want is the center of a series of contextual audio data and the size of each audio unit (frame or window) is near the duration of a video frame. Figure 2 shows the relation between the face frame and the contextual audio data.

We divided lip space into 40 classes and the same number of hidden Markov model is built. Each HMM is attached to a lip shape class. In training stage, using Baum-Welch algorithm and two-level frame, parameters of HMMs for Audio-visual mapping are figured out.

First of all, every training image (for HMM) is parameterized by previously built linear face model. Then current image is assigned to one of 40 visual classes, and the time-aligned contextual audio of that image will be an observation of that visual class. Finally 40 HMMs (first layer face model) are obtained by training their observations.

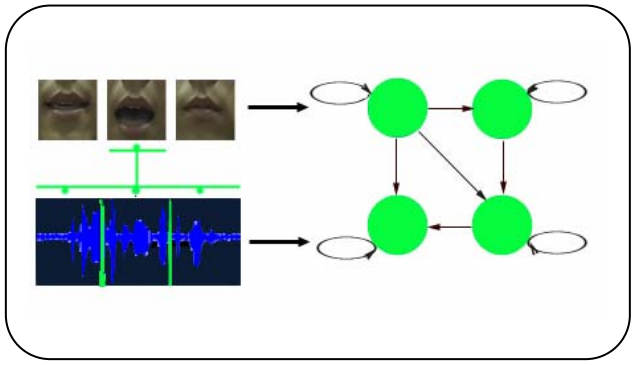

Figure 2. Audio-visual mapping HMM

4.2 Two-level HMM

We can find it in experiments that the monolayer HMM could hardly achieve a satisfying result. The reason is audio-to-visual mapping is many to many. Especially, for one lip appearance class, there will be very different audio observations. One monolayer model can be regarded as a combination of result parameters $\lambda = \{\lambda_1, \lambda_2, ..., \lambda_n\}$ of different training sequences. The parameters come from large number of training sequences will affect much in the result combination model. That's mean when an unfrequent vocal signal come up with one lip shape class, the result model of that lip shape class will express little information about the unfrequent vocal signal. In order to solve the problem, we propose a two- level model.

The first level hidden Markov model is same with usual combination model. Its parameters come from training sequences according to one lip shape class. The second layer composed of several sub-models called audio sub-models. The sub-models' training sequences are same kind of audio observations. Figure 3 shows the two-level model.

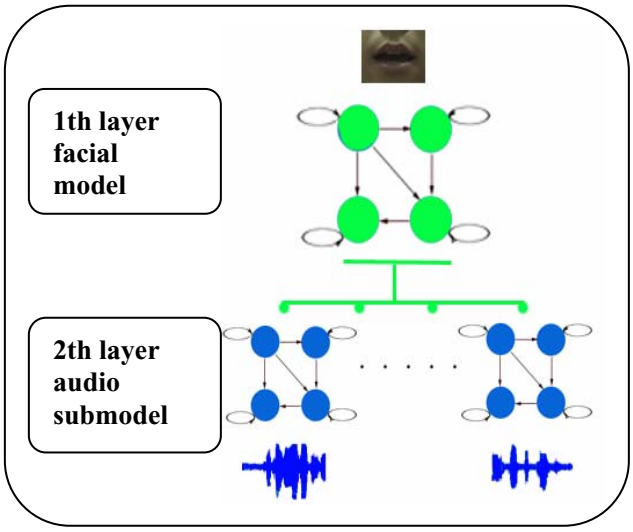

Figure 3. Two-level HMM

We use SOFM (self-organizing map) to cluster the audio observations in same visual class. N sub-models were built with N classes of audio observations. For each visual model we build one audio sub-model set, finally we get 40 sub-model sets attached to 40 visual models.

Though we mainly concern the difference of audio data for sub-model, the same visual class still has difference in visual space. In order to classify the sub training data accurately, we have to concern the visual feature. In sub-model training, for the sub-model clustering we use following distance,

$$d = d_{audio} + \alpha \cdot d_c \qquad (3)$$

Adjusting the value of α (0-1), the result could be promoted. d_{audio} is the Euclidean distance of audio feature, d_c is the distance of shape vector **C**. Generally d_c is small, so the result d has more information of audio difference.

16 LPCC is employed as the audio feature vector, since we use 5 frames audio feature as the observation the vector is 80 LPCC. In training stage, when the first level visual model is training, the corresponding audio

observations are recorded. A sub-model is trained with its own audio data subclass,

$$\lambda_* = \arg\max_{\lambda}[P(O_{sub} / \lambda, V)] \qquad (4)$$

and λ_* is the parameter of sub-model and O_{sub} is the audio observations of one subclass, V is some visual class.

In synthesizing stage, the input vocal data was first apply to all first level models then most probable n visual classes were selected whose sub-models were used to output a probability with Viterbi algorithm,

$$P(\lambda) = Max[P(O / \lambda_{sub-\mathrm{mod}el})] \qquad (5)$$

and $P(\lambda)$ is the maximum probability of all sub-models from one visual class.

Num of Sub model	NO	2	3	4	5
Error	0.142	0.061	0.025	0.008	0.007

Table 1. Comparison between the synthesized video and ground truth

The evaluation of the algorithm has been made by re-synthesizing the reserved video and the video frames both original and synthesized are classed to certain visual classes. In our experiment 36000 frames of video (25fps) have been used. A part of source corpus is reserved to test the synthesis result. The vocal tract of that part of corpus is the input, and a result visual class sequence is the output. The source video frames from that part of corpus (ground truth) are also classified to visual classes. Comparing the ground truth with synthesized results, an error table (table 1) show the difference between the monolayer HMM and two-level HMM, and the difference for variant number of sub-model.

5. CONCLUSION

In this paper, linear model coefficients are proposed as the visual representation of facial movements. New frame for hidden Markov model is adopted that make a final mapping decision based on sub-model probability. Experimental results show that this approach achieves good results. In future work we will focus on real time synthesis for learnable method and the facial expression learning during an utterance will be concerned.

6. ACKNOWLEDGE

This work is supported by the National 863 Program of China under Grant No. 2001AA114160, the National Natural Science Foundation of China under Grant No. 60375007 and the Natural Science Foundation of Beijing of China under Grant No. D070601-01.

7. REFERENCE

[1] S. Basu, N. Oliver and A. Pentland, "3D modeling and tracking of human lip motion," *ICCV'98*, 337-343, Bombay, India, 1998.

[2] Thomas Vetter and Tomaso Poggio. Linear object classes and image synthesis from a single example image. A.I. Memo 1531, MIT, 1995

[3] M. Brand. "Voice Puppetry". In *Proc. SIGGRAPH99*, pages 21-28, 1999.

[4] C. Bregler, M. Covell, and M. Slancy, "Video rewrite: driving visual speech with audio", *SIGGRAPH' 97*, 1997.

[5] T. Chen, and R.R. Rao, "Audio-Visual Integration in Multimodal Communications," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 837--852, May 1998.

[6] T. F. Cootes, C. J. Taylor, et al., "Active shape models - their training and application," *Computer Vision and Image Understanding*, 61(1):38-59, Jan. 1995.

[7] M. Covell and C. Bregler, "Eigen-points," *Proc. IEEE ICIP*, vol. 3, pp. 471-474, 1996.

[8] S. Curinga, F. Lavagetto, and F. Vignoli, "Lip movements synthesis using Time- Delay", *Proc. EUSIPCO-96*, Trieste, 1996.

[9] S. Morishima, K. Aizawa and H. Harashima, "An intelligent facial image coding driven by speech and phoneme," *Proc. IEEE ICASSP*, p.1795. Glasgow, UK, 1989.

[10] R. Rao and T. Chen, "Exploiting audio-visual correlation in coding of talking head sequences" in *Picutre Coding Symposium' 96*, Melbourne, Australia, March, 1996.

[11] Vetter, T., Jones, M. and Poggio, T., "A bootstrapping algorithm for learning linearized models of object classes," IEEE Conference on Computer Vision and Pattern Recognition, 1997

[12] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, pp.1464- 1480, 1990.

[13] S. Kshirsagar and N. Magnenat-Thalmann, "Lip Synchronization Using Linear Predictive Analysis", *Proceedings of IEEE International Conference on Multimedia and Expo*, New York, August 2000.

[14] Ezzat, T., Geiger, G., Poggio, T., 2002. "Trainable Videorealistic Speech Animation", SIGGRAPH 2002 pg 389-398.

[15] J. P. Lewis, "Automated lip-sync: Background and techniques", *J. Visualization and Computer Animation*, 2:118-122, 1991.

[16] D.W. Massaro, J. Beskow, *et al.*, "PictureMy Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks", in *Proc. AVSP'99*, Santa Cruz, USA.

[17] I. Pandzic, J. Ostermann, and D. Millen, "User evaluation: Synthetic talking faces for interactive services," The Visual Computer, vol. 15, Issue 7/8, 4 November 1999, pp. 330-340.