DYNAMIC AUDIO-VISUAL MAPPING USING FUSED HIDDEN MARKOV MODEL INVERSION METHOD

Le Xin, Jianhua Tao, Tieniu Tan

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100080, China
{xinle, jhtao, tnt}@nlpr.ia.ac.cn

ABSTRACT

Realistic audio-visual mapping remains a very challenging problem. Having short time delay between inputs and outputs is also of great importance. In this paper, we present a new dynamic audio-visual mapping approach based on the Fused Hidden Markov Model Inversion method. In our work, the Fused HMM is used to model the loose synchronization nature of the two tightly coupled audio speech and visual speech streams explicitly. Given novel audio inputs, the inversion algorithm is derived to synthesize visual counterparts by maximizing the joint probabilistic distribution of the Fused HMM. When it is implemented in the subsets built from the training corpus, realistic synthesized facial animation having relative short time delay is obtained. Experiments on a 3D motion capture bimodal database show that the synthetic results are comparable with the ground truth.

Index Terms— audio-visual mapping, speech driven facial animation, Baum-Welch inversion, 3D motion capture

1. INTRODUCTION

With the intensive requirements in human-computer interaction (HCI) and multimedia application, realistic audio-visual mapping has become a very popular subject in both research and flourishing industry domains. There are some basic problems such as synthesizing speech-related facial expression realistically and having a relative short time delay.

Many methods have been applied in this area during the last decade. Nowadays it is believed that Hidden Markov Models (HMMs) can achieve a high level of success among other methods. Since Yamamoto et al. [11], much attention has been paid in this direction for the purpose of accurately dynamic audio-visual mapping in sub-phonemic acoustic feature level. In some methods based on HMMs, including the remapping HMM [1], mixture-based HMM of Chen et al. [9] and Hidden Markov Model Inversion method (HMMI) proposed by Choi et al. [3], the vocal input had directly played a key role in the process of synthesizing visual counterparts. Great progress in prediction performance has been obtained from these representative HMM-based method [4]. However, some further improvement should be achieved. All these method made the assumption that it is enough to model both acoustic and visual component HMM with the same structure, no considering the individually specific structure. Furthermore, because of the intrinsic nature of loose synchronization in the two tightly coupled audio speech and visual speech streams, the training process based on the basic structure of HMM is generally complex, as the remapping HMM work showed.

Some HMM variants have received an attention for dynamic audio-visual mapping. For the consideration that the HMMI method outperformed the other two HMM-based methods [4], Xie et al. [10] presented an approach for speech animation using the coupled HMM Inversion (CHMMI). In this coupled HMM [2], two HMMs are linked together by explicitly describing the dependencies between the hidden states of the two HMMs. But it cannot handle tightly coupled series effectively for the dependence between the hidden states is only a weak representation of the statistical dependence of the observation sequences, and there are some computational drawbacks in such a globally optimization of all the parameters [8]. It is the same with the CHMMI method obviously.

In this paper, we present a new dynamic audio-visual mapping approach using a Fused HMM inversion method. As a technique of aiming at information fusion in feature level, the Fused HMM, proposed in Pan et al. [8], representing the bimodal relation explicitly, had had a successful performance in bimodal feature processing [8] [12]. It constructs a new structure linking the two component HMMs which is optimal according to the maximum entropy principle and a maximum mutual information (MMI) criterion. It has the advantage of reaching a better balance between model complexity and performance than other HMM-based fusion methods [8], resulting in the easy process of training and inversion. When it is implemented in the pre-built subsets, realistic synthesized facial animation having relative short time delay is obtained.

2. FUSED HMM AND IT'S INVERSION

Taking the advantage of modeling the data from a single sensor individually by a HMM, and according to the maximum

entropy principle and a maximum mutual information (MMI) criterion, the fusion model yields the following two structures [8], $p^{(1)}(O^a;O^v) = p(O^a)p(O^v|\hat{U}^a)$, $p^{(2)}(O^a;O^v) = p(O^v)p(O^a|\hat{U}^v)$, where $\hat{U}^a$ and $\hat{U}^v$ are the most possible hidden state sequences estimated by the Viterbi algorithm. These two structures are different essentially. $\hat{U}^a$ is required to be reliably estimated in the first one, and $\hat{U}^v$ in the second. As demonstrated by [8] [12], the first structure is preferred in bimodal speech processing, for the better reliability in estimating the best hidden states of the speech component HMM.

2.1. Learning Fused HMM

It is simple to train a Fused HMM, which includes the following three main steps in general: a) Two individual HMMs, consisting of visual component HMMs and audio component HMMs in our work, are trained independently by the EM algorithm; b) The best hidden state sequences of the audio component HMMs are found using the Viterbi algorithm; c) The coupling parameters, the conditional probability of visual observation o^v given audio states j, are determined.

In our work, it is clear that the first structure is selected, in which the coupling parameters represent the conditional probability distribution of visual observations o^v in visual component HMMs, given states j in audio component HMMs. Surly, the discrete coupling parameters in [8] can be easily extended to the continuous observations by the mixtures of the Gaussian, $b_j^a(o^v) = \sum_{k=1}^{K} c_{jk}N(o^v|\mu_{jk},\Sigma_{jk})$, where $o^v$is the visual features being modeled in visual component HMMs, and c_{jk}, μ_{jk} and Σ_{jk} are the coefficient, mean vector, and covariance matrix individually for the kth mixture Gaussian component in audio state j.

2.2. Inversion of Fused HMM

The HMM inversion algorithm was firstly proposed and applied to the robust speech recognition in [7]. Then Choi et al. [3] used HMM inversion in dynamic audio-visual mapping, whose usefulness had been demonstrated in [4]. Xie et al. [10] also derived their audio-visual conversion algorithms for the CHMMs [2].

As shown by [7] [3] [10], the optimal visual counterpart $\hat{O}^v$ can be estimated by the optimization of the following object function $L(O^v) = \log P(O^a, O^v|\lambda^{av})$, given an novel audio input O^a, where λ^{av} is the parameters of the fused HMM model. The optimization can be found by iteratively maximizing the auxiliary function $\hat{O}^v = \arg\max_{O^v} Q(\lambda^{av}, \lambda^{av}; O^a, O^v, \bar{O}^v)$, where O^v and $\bar{O}^v$ denote the old and new visual vector sequence respectively.

In our work, the fused model can be presented as

$$\begin{aligned} &P(O^a, O^v|\lambda^{av}) \\ &\quad = \kappa_1 P(O^a)P(O^v|\hat{U}^a) + \kappa_2 P(O^v)P(O^a|\hat{U}^v) \end{aligned}$$

where for constants $\kappa_1 \geq 0, \kappa_2 \geq 0$ with $\kappa_1 + \kappa_2 = 1$, $\kappa_1 > \kappa_2$. It is obvious that the two component HMMs will affect the synthesized result, but we have different reliability on them. It is an easy extension of the presentation in [8].

The object function can be expressed as

$$\begin{aligned} &\arg\max_{O^v} L(O^v)) \\ &\quad = \arg\max_{O^v} [\kappa_1 \log P(O^v|\hat{U}^a) + \kappa_2 \log P(O^v)] \\ &\quad = \arg\max_{O^v} [\kappa_1 \log \sum_{m^{av}} P(O^v, m^{av}|\hat{U}^a, \lambda^{av})) + \\ &\quad \kappa_2 \log \sum_{U^v}\sum_{m^v} P(O^v, U^v, m^v|\lambda^{av})] \end{aligned}$$

where $m^{av} = \{m^{av}_{\hat{u}^a_1 1}, m^{av}_{\hat{u}^a_2 2}, \ldots, m^{av}_{\hat{u}^a_T T}\}$, and $m^v = \{m^v_{u^v_1 1}, m^v_{u^v_2 2}, \ldots, m^v_{u^v_T T}\}$ that indicates the mixture component in visual component HMMs and the coupling parameters respectively. The influence of $\hat{U}^v$ on O^a is skipped here for lower reliability.

By some derivation using $\Delta = L(\bar{O}^v) - L(O^v)$, the auxiliary function can be derived as

$$\begin{aligned} &Q(\lambda^{av}, \lambda^{av}; O^a, O^v, \bar{O}^v) \\ &\quad = \kappa_1 \sum_{l=1}^{M^{av}} h_l \sum_{t=1}^{T} \log b_{\hat{u}^t_a l}(\bar{o}^v_t) + \kappa_2 \sum_{i=1}^{N}\sum_{l=1}^{M^v} H_{il} \sum_{t=1}^{T} \log b_{il}(\bar{o}^v_t) \end{aligned}$$

where

$$h_l = \frac{P(O^v, m^{av}_{\hat{u}^a_t t} = l|\hat{U}^a, \lambda^{av})}{\sum_{n=1}^{M^{av}} P(O^v, m^{av}_{\hat{u}^a_t t} = n|\hat{U}^a, \lambda^{av})} \tag{1}$$

$$H_{il} = \frac{P(O^v, u^v_t = i, m^v_{u^v_t t} = l|\lambda^{av})}{\sum_{i=1}^{N}\sum_{l=1}^{M^v} P(O^v, u^v_t = i, m^v_{u^v_t t} = l|\lambda^{av})} \tag{2}$$

$$P(O^v, m^{av}_{\hat{u}^a_t t} = l|\hat{U}^a, \lambda^{av}) = \prod_{t=1}^{T} c_{\hat{u}^a_t l} b_{\hat{u}^a_t l}(o^v_t) \tag{3}$$

$$P(O^v, u^v_t = i, m^v_{u^v_t t} = l|\lambda^{av}) = \sum_{j=1}^{N} \alpha_j(t-1) a_{ji} c_{il} b_{il}(o^a_t) \beta_i(t) \tag{4}$$

By making $\frac{\partial Q(\lambda^{av}, \lambda^{av}; O^a, O^v, \bar{O}^v)}{\partial \bar{o}^v_t} = 0$, we can find the reestimated inputs

$$\bar{o}^v_t = \frac{\kappa_1 \sum_{l=1}^{M^{av}} h_l \cdot \Sigma_l^{-1} \cdot \mu_l + \kappa_2 \sum_{i=1}^{N}\sum_{l=1}^{M^v} H_{il} \cdot \Sigma_{il}^{-1} \cdot \mu_{il}}{\kappa_1 \sum_{l=1}^{M^{av}} h_l \cdot \Sigma_l^{-1} + \kappa_2 \sum_{i=1}^{N}\sum_{l=1}^{M^v} H_{il} \cdot \Sigma_{il}^{-1}}. \tag{5}$$

3. FACIAL SYNTHESIS FROM NOVEL AUDIO

3.1. System overview

After the time synchronization in the audio-visual corpus, audio and visual features both having the same total numbers of frames are available by up-sampling visual observations.

And both audio and visual feature sequences considered in HMMs are taken at 13 (130ms time block) consecutive frames (7 backward, current, 7 forward) by tapped-delay lines [6], which is a tradeoff between the optimization of the contextual information and the requirement of undetected delay.

In our work, the idea using subsets in Hong et al. [5] is carried a step further. Those subsets are built by a two-layer clustering framework. The visual configuration is divided into 80 classes firstly. Then the audio observations in the same visual class are further classified into different sub-models. In this way, the joint audio-visual data are clustered based on their different modals in turn. This framework accounts for the many-to-many mapping between audio and visual features.

The k-means is used two times in visual configuration. The 80 visual classes found will account for not only the natural deformation in the pre-defined number of specific visual representation, but also the inter-class variation between these visual specifications.

3.2. Two times clustering in visual configuration

It has been believed that the number of the static visual configuration relative to audio speech is limited, just as the work on static viseum shows. So we focus on finding enough static visual representation in the first time clustering. The visual features from one single frame are imported to the k-means. The 40 classes found in this way represent the repertoire of facial specification.

However, it is far more enough to get a realistic facial synthesis result only in this way. So the intra-class natural facial deformation is modelled for the continuous real data. An index sequence accompanying the visual observation sequences is available. Each of index shows which cluster the current visual frame is belonged to. This index sequence will give us the basic variant in each visual class, which discloses the continuous essence of video. For the stability of video variation, only the visual shapes which have enough duration in one class are selected for the training of a Fused HMM model in that cluster.

Only a part of the whole training data is used in the above selection. Furthermore, the transformation between different visual classes should also be considered. All the frames in the existing gaps which are unused in the above selection are used in the second time clustering. These facial sub-sequences consisting of 13 consecutive frames are imputed in this time for another 40 clusters. The synthesized video output may be more continuous using the whole subsets.

3.3. Audio-visual mapping

Audio observations in the same visual class are further classified (6 sub-classes). In each subsets built by the two-layer clustering framework, one 3 states right-left visual component HMM and one 4 states right-left audio component HMM are learned, and the coupling GMM is fitted after the best hidden state sequences of the audio component HMM are estimated.

In the synthesis stage, the best cluster series corresponding to the continuous speech input is firstly estimated using Viterbi algorithm by maximizing the probability of the audio component HMM in all subsets given audio features. The best hidden state sequences $\hat{U}^a$ of the audio component HMM are also computed. Then the following steps will be done with $\kappa_1 = 0.8, \kappa_2 = 0.2$, until the change of visual subsequence is lower than predefined values in this cluster: 1) (Initialization) Initial the visual output based on the visual component HMM; 2) (E-step) Compute middle variants by Function (1), (2); 3) (M-step) Compute updates of visual output given middle variants by Function (5). The current time visual feature is chosen simply as the middle frame in the subsequence.

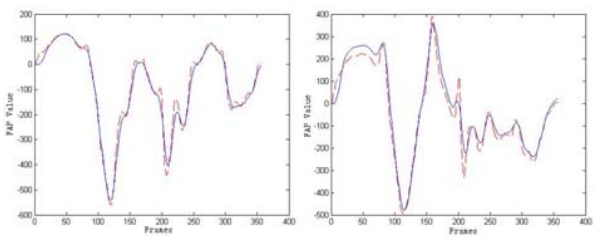
4. EXPERIMENTAL RESULTS AND CONCLUSION

A audio-visual database acquired by ourselves using a commercially available motion capture system, MotionAnalysis, with 8 cameras and a 75 Hz sampling rate, is used for the evaluation of our approach. The 3D facial motion trajectories of a subject articulating the pre-designed corpus with 3d markers on the face and its accompanying time-aligned audio are recorded simultaneously.

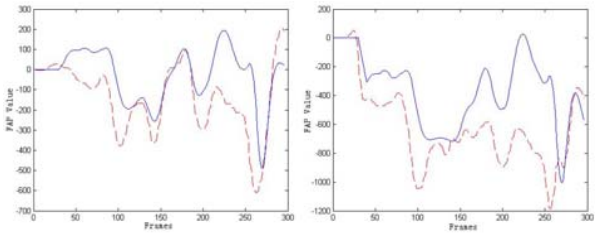
In our system, a phoneme-balanced corpus covering all single phoneme and most of the frequently used bi-phoneme combination existing in Chinese is designed. The same 129 sentences are repeated five times, in which four times are used as training data, one time as validation data. Another 9 different sentences are used as test data. The duration of each sequence is about 2 5s. So the total duration of the sequences in our database is about 3000s. The audio wave is sampled on the rate of 44100 Hz with 16 bit resolution. 12D MFCC coefficient and 1D energy parameters, their delta, and delta-delta parameters are all computed every 10ms to capture more dynamics in the vocal feature. 50 markers compatible with FDPs in MPGE-4 standard are selected to obtain powerful facial movement encoding. Since we focus on speech driven facial animation, only 8 markers on the outer lip boundary are used. The 10 FAPs in group 8 are computed from the 3D trajectories of the chosen markers to represent the lip movement. Here the FAP sequences are up-sampled to 100 fps in order to have the same total frame number as the vocal feature.

Certainly, the motion capture data must be pre-processed before FAPs are computed: 1) Gaps and aberrations in trajectory data should be filled or fixed frame by frame; 2) The 5 markers put on the head are used to compensate the global head movement; 3) The 3D trajectories of face markers in the first frame have to be normalized into an upright position in positive XYZ space, compatible with the definition of neutral face in MPEG-4 standard.

Both quantitative evaluation and qualitative observation results are shown here. Figure 1 shows comparisons between

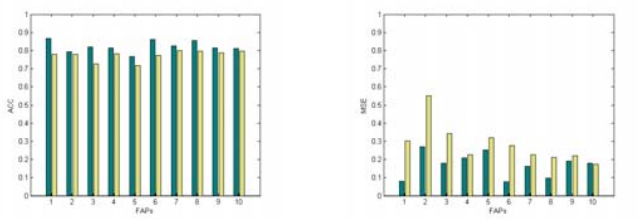


(a) Validation data



(b) Test data

Fig. 1. Comparisons between synthesized FAP stream results (blue line) with the corresponding ground-truth data (red dash line) frame by frame. (a) Validation sentence; (b) Test sentence. In each figure, left is about FAP 51 (lower_t_midlip_o), and right is about FAP 52 (raise_b_midlip_o).



(a) ACC (b) MSE

Fig. 2. The (a) ACC and (b) MSE for each FAPs on the whole database with validation set (in green) and test set (in yellow).

synthesized FAP stream results with the corresponding ground-truth data frame by frame. Prediction performance about our method is measured quantitatively using the normalized mean square error (MSE) $\varepsilon = \frac{1}{T \cdot \sigma_v^2} \sum_{t=1}^{T} (\hat{v}_t - v_t)^2$ and average correlation coefficient (ACC) $\rho = \frac{1}{T} \sum_{t=1}^{T} \frac{(\hat{v}_t - \mu_{\hat{v}})(v_t - \mu_v)}{\sigma_{\hat{v}} \sigma_v}$ for each FAPs on the whole database shown by Figure 2, where v_t and $\hat{v}_t$ denote the recorded and the predicted FAP stream individually, T is the total number of frames in the database, and μ_v, $\mu_{\hat{v}}$, σ_v and $\sigma_{\hat{v}}$ are corresponding mean and standard deviation.

Finally, A 2D image-based MPEG-4 facial animation engine is used to access our synthesized FAP streams qualitatively. Figure 3 shows some frames of synthesized talking head.

We can see from the synthesized results in the experiments that our method shows a good performance which is comparable with the ground truth. Our Fused HMM Inversion method and the two layer clustering framework can result in an accurately prediction having a relative short time delay.

Fig. 3. Five frames in one synthesized sequence.

Acknowledgement

The work was supported by the National Natural Science Foundation of China (No. 60575032) and the 863 Program (No. 2006AA01Z138).

5. REFERENCES

[1] M. Brand. Voice puppetry. In *Siggraph*, 1999.

[2] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *CVPR 1997*, June 1997.

[3] K. Choi and J.-N. Hwang. Baum-welch hmm inversion for reliable audio-to-visual conversion. In *Proc. IEEE Int. Workshop Multimedia Signal Processing*, 1999.

[4] S. Fu, R. Gutierrez-Osuna, A. Esposito, P. K. Kakumanu, and O. N. Garcia. Audio/visual mapping with cross-modal hidden markov models. *IEEE Transactions on Multimedia*, 7(2):243–252, April 2005.

[5] P. Y. Hong, Z. Wen, and T. S. Huang. Real-time speech-driven face animation with expressions using neural networks. *IEEE Transactions on Neural Networks*, 13(4):916–927, July 2002.

[6] P. K. Kakumanu. *Audio-visual processing for speech driven facial animation*. Master Thesis, Wright Stage University, 2002.

[7] S. Moon and J.-N. Hwang. Robust speech recognition based on joint model and feature space optimization of hidden markov model. *IEEE Transaction on Neural Networks*, 8(2):194–204, March 1997.

[8] H. Pan, S. E. Levinson, T. S. Huang, and Z.-P. Liang. A fused hidden markov model with application to bimodal speech processing. *IEEE Transactions on Signal Processing*, 52(3):573–581, March 2004.

[9] R. R. Rao, T. Chen, and R. M. Mersereau. Audio-to visual conversion for multimedia communication. *IEEE Trans. On Industrial Electronics*, 45(1):15–22, Feb 1998.

[10] L. Xie and Z.-Q. Liu. Speech animation using coupled hidden markov models. In *ICPR 2006*, August 2006.

[11] E. Yamamoto, S. Nakamura, and K. Shikano. Lip movement synthesis from speech based on hidden markov models. *Speech Communication*, 26(1):105–115, May 1998.

[12] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T. S. Huang, and S. Levinson. Audio-visual affect recognition through multi-stream fused hmm for hci. In *CVPR 2005*, June 2005.