PARTIAL LINEAR REGRESSION FOR AUDIO-DRIVEN TALKING HEAD APPLICATION

Chao-Kuei Hsieh and Yung-Chang Chen*

Department of Electrical Engineering
National Tsing Hua University, Hsinchu, Taiwan 300, R.O.C.
*E-mail: ycchen@ee.nthu.edu.tw

ABSTRACT

Virtual avatars in many applications are constructed manually or by a single speech-driven model which needs a lot of training data and long training time. It's an essential problem to build up a user-dependent model more efficiently. In this paper, a new adaptation method, called the partial linear regression (PLR), is proposed and adopted in an audio-driven talking head application. This method allows users to adapt the partial parameters from the available adaptive data while keeping the others unchanged. In our experiments, the PLR algorithm can retrench the hours of time spent on retraining a new user-dependent model, and adjust the user-independent model to a more personalized one. The animated results with adapted models were 36% closer to the user-dependent model than using the pre-trained user-independent model.

1. INTRODUCTION

With the rapid development of multimedia technology, the virtual avatar has been widely used in many applications, like cartoon or computer game characters and news announcers. Nevertheless, huge amount of manpower is needed in adjusting the avatar frame by frame to achieve a vivid and precise synthetic facial animation. A real-time speech-driven synthetic talking head, or so-called audio-to-visual synthesis system, is expected, which can provide an effective interface for many applications. In an audio-to-visual synthesis system, it needs a model established for describing the correspondence between the acoustic parameters and the mouth-shape parameters. In other words, the corresponding visual information is to be estimated for some given acoustic parameters, such as the phonemes, the cepstral coefficients or the line spectrum pairs.

A number of algorithms have been proposed for the task of mapping between acoustic parameters and visual parameters. The conversion problem is treated as one of finding the best approximation from given sets of training data. These approaches were briefly discussed by Chen and Rao [1], including vector quantization, Hidden Markov Models (HMM), and neural networks. However, the speech-driven systems were generally made to be user-independent for satisfactory average performance, which means a decrease in accuracy rate for a specific user. To maintain a high performance, a time-consuming retraining procedure for a new user-dependent model is unavoidable since there is no reported adaptation method for this application in the literature.

On the other hand, speaker adaptation methods have been extensively studied in the speech recognition field. There are two main categories in the adaptation methods. The first is the eigenvector-based speaker adaptation method [2]. The other is based on the acoustic model, and is simpler than the former since the normalization for the training data is not necessary. A user-independent model is statistically established with the training data of several speakers in the beginning, and the parameters are then modified with certain adaptation data of a new user. The adaptation schemes include Maximum a Posteriori (MAP) Estimation [3], Maximum Likelihood Linear Regression (MLLR) [4], VFS [5], and nonlinear neural network [6]. In these methods, they tried to adjust the model parameters to maximize the occurrence probability of the new observation data. Among them, the MLLR method is more widely adopted for its simplicity and effectiveness when the set of adaptation data is small.

In this study, we try to integrate the MLLR adaptation approach with the audio-to-visual conversion of Gaussian Mixture Model (GMM). However, to obtain the precise visual adaptation information of a new user is not feasible in a usual environment, since some markers, infrared cameras, and post-processing are needed. This makes the MLLR not fully adequate to adapt only the audio parameters while keeping the visual part the same. In other words, we require another appropriate adaptation, by means of which the new model will map the new audio parameters of a new user to the original visual movement.

A new adaptation method, called partial linear regression (PLR), is proposed in this paper. It is derived from the MLLR and put into practice in an audio-driven talking head system (Fig. 1). Rather than a time consuming retraining procedure, a simple adaptation with a small amount of additional data will be sufficient to adjust the model so as to be more applicable to the new user.

0-7803-9332-5/05/$20.00 ©2005 IEEE

The rest of the paper is organized as follows. In Section II, we describe the audio-driven talking head system which uses the Gaussian mixture model to represent the relationship between audio and video feature vectors. Section III provides a detailed description of the proposed PLR model adaptation algorithm. Some experimental results are described in Section IV, and section V concludes the paper.

2. AUDIO-DRIVEN TALKING HEAD SYSTEM

2.1. System Architecture

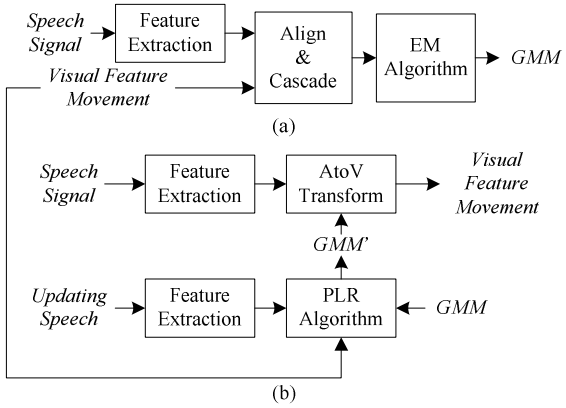

Figure 1. (a) Flowchart of training phase; (b) Flowchart of testing and updating phases

The flowcharts of the training and the testing phases of our audio-driven talking head system are described in figure 1. In the audio signal processing, we extract 10th-order line spectrum pair (LSP) coefficients from every audio frame of 240 samples. In the training phase, the frame rates of the audio and video signal generally differ from each other. After labeling the beginning and the ending points of every training word manually, we use linear interpolation to align the audio and visual feature vectors and cascade them into a single vector. The Gaussian mixture model, derived by the EM algorithm, is then adopted to represent the distribution of the audio-visual vector.

In the testing phase (Fig 1(b)), to obtain the optimal estimator of the visual vector from an audio vector is actually to calculate the conditional expectation using the trained GMM. However, voice features of the current user may be distinct from others'. An adaptation would be necessary to modify the pre-trained GMM more suitable to the new user.

2.2. Gaussian Mixture Model

The density function of Gaussian mixture model is defined as

$$p(\boldsymbol{y}) = \sum_{i=1}^{m} \pi_i f_i(\boldsymbol{y}) \text{ where } \boldsymbol{y}^T = \begin{bmatrix} \boldsymbol{v}^T & \boldsymbol{a}^T \end{bmatrix} \text{ and}$$

$$f_i(\boldsymbol{y}) = \frac{1}{(2\pi)^{d/2} \Delta_i^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu}_i)^T \overline{\boldsymbol{S}}_i^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_i) \right\},$$

where $\boldsymbol{v}$ is the visual feature vector, $\boldsymbol{a}$ is the audio feature vector, d is the dimension of vector $\boldsymbol{y}$, π_i is the weighting of the i-th Gaussian kernel, $\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_{v,i}^T & \boldsymbol{\mu}_{a,i}^T \end{bmatrix}^T$ is the mean vector, and $\overline{\boldsymbol{S}}_i = \begin{bmatrix} \overline{\boldsymbol{S}}_{vv} & \overline{\boldsymbol{S}}_{va} \\ \overline{\boldsymbol{S}}_{av} & \overline{\boldsymbol{S}}_{aa} \end{bmatrix}$ is the covariance matrix.

2.3. Audio-to-Visual Conversion

For two vectors $\boldsymbol{v}$, $\boldsymbol{a}$ modeled as jointly Gaussian, the optimal estimator of $\boldsymbol{v}$ given the value of $\boldsymbol{a}$ in mean-squared error sense is actually the conditional expectation of $\boldsymbol{v}$ given $\boldsymbol{a}$. For a Gaussian mixture model, similarly, the marginal probability function of $\boldsymbol{a}$ can be obtained from

$$f_a(\boldsymbol{a}) = \int \sum_{i=1}^{m} \pi_i N(\boldsymbol{a}, \boldsymbol{v}; \boldsymbol{\mu}_i, \overline{\boldsymbol{S}}_i) d\boldsymbol{v} = \sum_{i=1}^{m} \pi_i N(\boldsymbol{a}; \boldsymbol{\mu}_{a,i}, \overline{\boldsymbol{S}}_{a,i})$$

And the conditional expectation of v given a can be derived as

$$E[\boldsymbol{v} \mid \boldsymbol{a}] = \int \boldsymbol{v} \frac{f_{av}(\boldsymbol{a}, \boldsymbol{v})}{f_a(\boldsymbol{a})} d\boldsymbol{v} = \sum_{i=1}^{m} \frac{\pi_i}{f_a(\boldsymbol{a})} \int \boldsymbol{v} N(\boldsymbol{a}, \boldsymbol{v}; \boldsymbol{\mu}_i, \overline{\boldsymbol{S}}_i) d\boldsymbol{v}$$

$$= \sum_{i=1}^{m} \left(\frac{\pi_i N(\boldsymbol{a}; \boldsymbol{\mu}_{a,i}, \overline{\boldsymbol{S}}_{aa,i})}{f_a(\boldsymbol{a})} \right) \left(\boldsymbol{\mu}_{v,i} + \overline{\boldsymbol{S}}_{va,i} \overline{\boldsymbol{S}}_{aa,i}^{-1} (\boldsymbol{a} - \boldsymbol{\mu}_{a,i}) \right)$$

which is the optimal estimator for GMM in mean-squared sense

3. PARTIAL LINEAR REGRESSION

In MLLR of mean adaptation [4], the purpose is to maximize the likelihood of the new observation data by linear-regressively adjusting the mean vectors of every Gaussian kernel, i.e. $\boldsymbol{\mu}_i' = \overline{\boldsymbol{W}} \begin{bmatrix} 1 \\ \boldsymbol{\mu}_i \end{bmatrix}$.

The MLLR method performs well even when the amount of the adaptation data is insufficient. It modifies every single value in all the mean vectors of the Gaussian kernels. In other words, if we can gather both audio and visual adaptation data at the same time, the MLLR will be qualified for the task of model adaptation in the audio-to-visual conversion. Unfortunately, to obtain the precise visual adaptation information, the 3-dimensional movements of specific control points, of a new user is not feasible in an ordinary environment, since some markers, infrared cameras, and post-processing are needed. Only audio adaptation data is available. This makes the MLLR not conformable to our demand. In our application, we may merely want to adapt audio mean vectors, $\boldsymbol{\mu}_{a,i}$, and keep the correspondence between audio and visual vectors

unchanged. Another appropriate adaptation is indispensable, by means of which the new model will map the audio parameters of a new user to the original visual movement. MLLR is then modified and integrated with the concept of conditional expectation used in audio-to-visual conversion part, mentioned in Section 2.3.

In Section 2.3, with the conditional expectation, the corresponding visual information, $E[\boldsymbol{v}|\boldsymbol{a}]$, can be estimated for some given acoustic parameters in the audio-to-visual conversion. Oppositely, we can evaluate the audio information from its corresponding visual parameters, $E[\boldsymbol{a}|\boldsymbol{v}]$, by the same token. The concept of our proposed PLR is, by adjusting the audio mean vectors $\boldsymbol{\mu}_{a,i}$ linear regressively, to minimize the distance between the adaptation data a and the optimal estimator of given the value of $\boldsymbol{v}$, corresponding to $\boldsymbol{a}$. To do so, the new user has to pronounce the words we have pre-defined.

Suppose we have J adaptation data aj, j=1,2,…J. Our final goal is to *minimize* $\sum_{j=1}^{J}\left\|\boldsymbol{a}_j - E\left[\boldsymbol{a} \mid \boldsymbol{v}_j\right]\right\|$, i.e., to solve the equation

$$\arg\min_{\bar{\boldsymbol{A}},\boldsymbol{b}} \sum_{j=1}^{J}\left\|\boldsymbol{a}_j - \sum_{i=1}^{m}\left[\begin{array}{l}\dfrac{\pi_i N\left(\boldsymbol{v}_j;\boldsymbol{\mu}_{v,i},\bar{\boldsymbol{S}}_{vv,i}\right)}{f_v\left(\boldsymbol{v}_j\right)}\times \\ \left(\bar{\boldsymbol{A}}\boldsymbol{\mu}_{a,i}+\boldsymbol{b}+\bar{\boldsymbol{S}}_{av,i}\bar{\boldsymbol{S}}_{vv,i}^{-1}\left(\boldsymbol{v}_j-\boldsymbol{\mu}_{v,i}\right)\right)\end{array}\right]\right\| \quad (1)$$

where

$$f_v\left(\boldsymbol{v}_j\right)=\int\sum_{i=1}^{M}\pi_i N\left(\boldsymbol{a},\boldsymbol{v}_j;\boldsymbol{\mu}_{v,i},\bar{\boldsymbol{S}}_{vv,i}\right)d\boldsymbol{a}=\sum_{i=1}^{M}\pi_i N\left(\boldsymbol{v}_j;\boldsymbol{\mu}_{v,i},\bar{\boldsymbol{S}}_{vv,i}\right)$$

After defining $\dfrac{\pi_i N\left(\boldsymbol{v}_j;\boldsymbol{\mu}_{v,i},\bar{\boldsymbol{S}}_{vv,i}\right)}{f_v\left(\boldsymbol{v}_j\right)}=w_{i,j}$, (1) becomes

$$\arg\min_{\bar{\boldsymbol{A}},\boldsymbol{b}} \sum_{j=1}^{J}\left\|\begin{array}{l}\left[\boldsymbol{a}_j-\sum_{i=1}^{m}w_{i,j}\left(\bar{\boldsymbol{S}}_{av,i}\bar{\boldsymbol{S}}_{vv,i}^{-1}\left(\boldsymbol{v}_j-\boldsymbol{\mu}_{v,i}\right)\right)\right] \\ -\sum_{i=1}^{m}w_{i,j}\left(\bar{\boldsymbol{A}}\boldsymbol{\mu}_{a,i}+\boldsymbol{b}\right)\end{array}\right\| \quad (2)$$

Let $\boldsymbol{a}_j-\sum_{i=1}^{m}w_{i,j}\left(\bar{\boldsymbol{S}}_{av,i}\bar{\boldsymbol{S}}_{vv,i}^{-1}\left(\boldsymbol{v}_j-\boldsymbol{\mu}_{v,i}\right)\right)\triangleq\boldsymbol{d}_j$.

We now have

$$\arg\min_{\bar{\boldsymbol{A}},\boldsymbol{b}} \sum_{j=1}^{J}\left\|\left(\boldsymbol{d}_j-\boldsymbol{b}\right)-\bar{\boldsymbol{A}}\sum_{i=1}^{m}w_{i,j}\boldsymbol{\mu}_{a,i}\right\|\triangleq\arg\min_{\bar{\boldsymbol{A}},\boldsymbol{b}} \sum_{j=1}^{J}\left\|\left(\boldsymbol{d}_j-\boldsymbol{b}\right)-\bar{\boldsymbol{A}}\boldsymbol{c}_j\right\|$$

For simplicity, we can solve this question as a least square problem. That means

$$\left[\boldsymbol{c}^T\quad\boldsymbol{1}\right]\cdot\left[\begin{array}{c}\bar{\boldsymbol{A}}(k,:)^T\\ \boldsymbol{b}(k)\end{array}\right]=\boldsymbol{d}^T(k),\ \boldsymbol{c}=\left[\boldsymbol{c}_1\ldots\boldsymbol{c}_J\right],\boldsymbol{d}=\left[\boldsymbol{d}_1\ldots\boldsymbol{d}_J\right]$$

$$\left[\begin{array}{c}\bar{\boldsymbol{A}}(k,:)^T\\ \boldsymbol{b}(k)\end{array}\right]=\left(\left[\begin{array}{c}\boldsymbol{c}\\ \boldsymbol{1}\end{array}\right]\cdot\left[\boldsymbol{c}^T\quad\boldsymbol{1}\right]\right)^{-1}\left[\begin{array}{c}\boldsymbol{c}\\ \boldsymbol{1}\end{array}\right]\boldsymbol{d}^T(k) \quad (3)$$

4. EXPERIMENTAL RESULTS

4.1. Experimental Data

The audio ground truth data was captured with a microphone with a 8KHz and 16-bits mono channel, and the facial movement was captured by 6 infrared cameras, 120 fps, with 27 particular markers (1 is the root) stuck on certain feature points of the user's face (Fig 2).

Figure 2. A snapshot of video grabbed from the digital video camera during motion capturing

For each of the 3 male subjects in our experiment, we recorded 413 Chinese words. The start and the end point of each word were labeled manually, and 10 LSP coefficients were calculated from each voice segment of 240 samples, and then cascaded with their corresponding visual feature vectors. We choose the horizontal and the vertical moving distance of the 8 points around the mouth as the visual feature, because only these points are more related to the audio variation. The dimension of the visual vectors is 16. In order to preserve the mouth movement information, only the visual data of one person was adopted and used for the other subjects as a duplicate. In this way, there could be a standard for comparing the estimated result from the audio-to-visual conversion no matter a user-dependent or user-independent model was used. However, the frame number of audio vectors is absolutely not the same as that of visual vectors because of the different sample rate and the varying length between different people. Therefore, we normalized the audio vectors to the visual vectors in each word by linear interpolation and resulted in 19315 audio-visual vectors totally for each user. GMMs with 10 kernels were used to approximate the relationship between the audio feature and the visual feature factor.

4.2. Experiment I

The odd audio-visual vectors and the even vectors are used as the training data and testing data, respectively. Three kinds of model were established:

(1) For each of the 3 users, a user-dependent voice to mouth movement model ($Model_i$, i=1,2,3) was established using their own audio training data and same visual data.

(2) Training the user-independent voice to face movement model ($Model_{all}$) using the mixed data (the $3k+i$ vectors from the training data of the i-th user, where $k\in$Z and i=1,2,3.) of all the users.

(3) With the proposed PLR method, we adjusted the model Modelall with the other audio-visual vectors, not used in the training of $Model_{all}$, to obtain $Model_{all\text{-}i}$, i=1,2,3 , which will be more consistent to user i.

After the GMMs are established and adapted, we can directly derive the corresponding facial movement vector with a given audio vector. In the testing phase, the visual vector conducted from a certain model was compared with the ground truth data. The mean values of the difference with the mouth width normalized to 100 are recorded in table I. As the result shows, the relationship between the performances of the GMMs applied to the i-th user is:

$$Model_i > Model_{all-i} > Model_{all}, i = 1,2,3$$

Table 1. Mean of the difference between the original data and the value obtained from GMM

GMM	*Mean*		
	User 1	*User 2*	*User 3*
$Model_{all}$	4.45	4.57	4.60
$Model_{all\text{-}1}$	*3.72*	4.51	4.55
$Model_{all\text{-}2}$	4.73	*4.22*	4.56
$Model_{all\text{-}3}$	4.41	4.44	*4.21*
$Model_1$	*2.95*	4.28	4.34
$Model_2$	4.39	*3.42*	4.56
$Model_3$	4.67	4.35	*3.30*

4.3 Experiment II

Instead of using the all odd audio-visual vectors as the training data, we randomly choose 5, 10, 15, 20, 25, and 30 words from the recorded 413 Chinese words, and use the audio-visual vectors of these selected words as the adaptation data. The user-dependent models trained in experiment I, $Model_i$, were consequenctly used as the reference. Each random selection was given 3 trials. The whole 413 words were used as the testing data and the difference between the estimated visual vector and the ground truth was shown in figure 3. The value '0' in the adapting words axis means that no adaptation is implemented, and the corresponding value is the result of the original user-dependent model, $Model_i$, applied on the new user. The word 'odd' stands for using the odd vectors as the adaptation data as in experiment I (about 206 words, quantitatively).

As the figure shows, there is a trend that the difference between the estimated value and the ground truth decreases while the number of adaptation words increases. When the set of adaptation data is small, the selected words will be critical to the result of model adjustment. With inappropriate extra data, the performance of adapted model could be worse, even than using the original user-dependent model. When the number of adaptation words expands to about 20, the effect of applying PLR for model adaptation will be affirmatively positive.

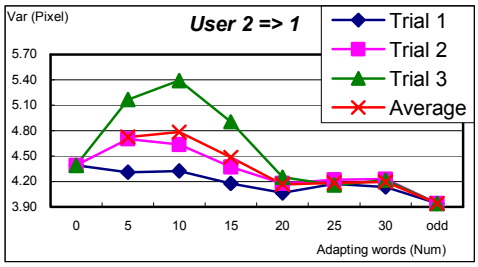

Figure 3. One result of model adaptation with different number of adaptation words

5. CONCLUSION

We have proposed a new adaptation algorithm using partial-linear-regression. The PLR method can be used in updating a part of the mean vector in Gaussian mixture model, keeping the corresponding relationship unchanged. This is due to that the precise visual data of a new user can not be obtained easily, and we may only collect the audio information in the adaptation procedure. As the experimental result in Table 1 shows, we can derive a more adequate model for the new user via the PLR adaptation algorithm, rather than a time-consuming re-training task. The set of adaptation data plays a very important role when it is small and randomly selected. The adjusted model could outperform the original one only if the words were chosen appropriately. How to choose more efficient adaptation data is an important issue and this is still under investigation, although it is obvious that if the more adaptation data is used, the better performance there will be.

6. REFERENCE

[1] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication," *Proc. IEEE*, vol. 86, no. 5, pp. 837-852, May 1998.

[2] Anastaskos T., McDonough J., Schwartz R. and Makhoul J., "A Compact Model for Speaker Adaptive Training", *ICSLP*, 1996.

[3] J. L. Gauvain, and C. H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Transactions on Speech and Audio Processing*, 1994.

[4] C.J. Leggetter, P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language, Vol. 9, pp. 171–185, 1995.

[5] M. Tonomura, T. Kosaka, and S. Matunaga, "Speaker Adaptation Based on Transfer Vector Field Smoothing Using Maximum a Posteriori Probability Estimation", *ICASSP-95*, Vol. 1, pp. 688–691, 1995.

[6]]M. Gales and P. Woodland, "Variance Compensation Within the MLLR Frame Work," *Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR.242*, Feb. 1996.