Real-Time Speech-Driven Face Animation with Expressions Using Neural Networks

Pengyu Hong, Zhen Wen, and Thomas S. Huang, *Fellow, IEEE*

Abstract— **A real-time speech-driven synthetic talking face provides an effective multimodal communication interface in distributed collaboration environments. Nonverbal gestures such as facial expressions are important to human communication and should be considered by speech-driven face animation systems. In this paper, we present a framework that systematically addresses facial deformation modelling, automatic facial motion analysis, and real-time speech-driven face animation with expression using neural networks. Based on this framework, we learn a quantitative visual representation of the facial deformations, called the Motion Units (MUs). An facial deformation can be approximated by a linear combination of the MUs weighted by MU parameters (MUPs). We develop an MU-based facial motion tracking algorithm which is used to collect an audio-visual training database. Then, we construct a real-time audio-to-MUP mapping by training a set of neural networks using the collected audio-visual training database. The quantitative evaluation of the mapping shows the effectiveness of the proposed approach. Using the proposed method, we develop the functionality of real-time speech-driven face animation with expressions for the iFACE system [1]. Experimental results show that the synthetic expressive talking face of the iFACE system is comparable with a real face in terms of the effectiveness of their influences on bimodal human emotion perception.**

Keywords— **Real-time speech-driven talking face with expressions, facial deformation modelling, facial motion analysis and synthesis, neural networks.**

I. Introduction

Synthetic talking faces have been developed for applications such as email reader, web newscaster, virtual friend, computer agent, and so on [2], [3], [4], [5], [6], [7]. Research shows that a synthetic talking face can help people understand the associated speech in noisy environments [8]. It also helps people react more positively in interactive services [9]. Real-time speech-driven synthetic talking face, as a computer-aided human-human interface, provides an effective and efficient multimodal communication channel for "face-to-face" communication in distributed collaboration environments [10], [11], [12], [13], [14].

However, up to date, few real-time speech-driven face animation systems have considered synthesizing facial expressions. Facial expression can strengthen or weaken the sense of the corresponding speech. It helps attract the attention of the listener. More importantly, it is the best way to visually express emotion [15]. Therefore, a real-time speech driven facial animation system which synthesizes facial expressions will be more effective in terms of delivering visual and emotional information.

It would be ideal if the computer can accurately recognize the emotions from speech and use the recognition results to synthesize facial expressions. However, research has shown that it is very difficult to recognize emotion from speech. It is shown that emotion recognition by either human or computer from speech is much more difficult than the recognition of words or sentences. In Scherer's study, the voice of 14 professional actors is used [16]. Scherer's experimental results showed that human ability to recognize emotions from purely vocal stimuli is around 60%. Dellaert et al. [17] compared the performances of different classification algorithms on a speech database, which contain four emotion categories (happy, sad, anger, and fear) and 50 short sentences per category spoken by 5 speakers. The highest recognition rate achieved by those algorithms is 79.5%. Petrushin [18] compared human and computer recognition of emotions from speech and reported around 65% recognition rate for both cases. Recently, Petrushin [19] reported that human subjects can recognize five emotions (normal, happy, angry, sad, and afraid) with the average accuracy of 63.5%.

On the other hand, research has showed that recognizing facial expressions using visual information alone is much easier. Recent works on automatic facial expression recognition by computer use optical flow, appearance-based models, or local parametric motion models to extract the information about facial features [20], [21], [22], [23], [24]. The extracted information is inputted into classification algorithms for facial expressions recognition. A recognition rate as high as 98% was achieved by Essa and Pentland on five expressions (smile, surprise, anger, disgust and raise brow) [24]. Therefore, visual information is more effective than audio information in terms of conveying informatino related to the emotional states. This is because that the voice and facial expressions may not convey the same emotional information in many situations. For example, the subject may speak calmly while smiling. On the other hand, the subject may speak emotionally without noticeable facial expressions.

There are works on synthesizing facial expressions directly from speech [25], [26]. However, being aware of the above difficulties, we assume that the emotional state of the user is known and concentrate on developing the methodology and techniques for synthesizing expressive talking faces given the speech signals and emotional states. The user is required to decide the facial expression of his/her avatar. For example, the user can tell the computer which expres-

P. Hong is with the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. E-mail: hong@ifp.uiuc.edu.

Z. Wen is with the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA, E-mail: zhenwen@ifp.uiuc.edu

T. S. Huang is with the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA, E-mail: huang@ifp.uiuc.edu

sion should be added to the synthetic talking face by hitting the corresponding button in the interface of the face animation system or the corresponding key in the key board. This will give the user more freedom to remotely control his/her avatar-based appearance.

The rest of the paper is organized as following. In Section II, we review the related work. We then introduce an integrated framework for face model, facial motion analysis and synthesis in Section III. The integrated framework systematically addresses three related issues: (1) Learning a quantitative visual representation for facial deformation modelling and face animation; (2) Automatic facial motion analysis; and (3) Speech to facial coarticulation modelling. Section IV discusses learning a quantitative visual representation, called the Motion Units (MUs), from a set of labelled facial deformation data. In Section V, we propose an MU-based facial motion analysis algorithm which is used to analyze the facial movements of speakers. In Section VI, we discuss how to train a real-time audio-to-visual mapping with expression using neural networks. The experimental results are provided in Section VII. Finally, the paper is concluded with summary and discussions in Section VIII.

II. Previous Work

The core of speech-driven face animation is the audio-to-visual mapping which maps the audio information to the visual information representing the facial movements. The visual information is decided by the way that a face is modelled. To achieve natural face animation, the audio-to-visual mapping should be learned from a large audio-visual training database of real facial movements and the corresponding speech streams. A robust automatic facial motion analysis algorithm is required to effectively and efficiently collect a large set of audio-visual training data from real human subjects. In this section, we review previous works on face modelling, facial motion analysis, and real-time speech-driven face animation.

A. Face Modelling

One main goal of face modelling is to develop a facial deformation control model that deforms the facial surface spatially. Human faces are commonly modelled as free-form geometric mesh models [27], [28], [29], [30], [31], parametric geometric mesh models [32], [33], [34], or physics-based models [35], [36], [37]. Different face models have different facial deformation control model and result in different visual features used in the audio-visual training database.

A free-form face model has an explicit control model, which consists of a set of control points. The user can manually adjust the control points to manipulate the facial surface. Once the coordinates of the control points are decided, the remaining vertices of the face model are deformed by interpolation using B-spline functions [27], radial basis functions [28], [29], the combination of affine functions and radial basis functions [30], or rational functions [31]. It is straight forward to manipulate the facial surface using free-form face models. However, little research has been done in a systematic way to address how to choose the control points, how to choose the interpolation functions, how to adjust control points, and what are the correlations among those control points.

Parametric face models use a set of parameters to decide the shapes of the face surface. The coordinates of the vertices on the face models are calculated by a set of predefined functions whose variables are those parameters. The difficulty of this kind of approach is how to design those functions. Usually, they are designed manually and subjectively. Therefore, those functions may not well represent the characteristics of natural facial deformations.

Physics-based models simulate facial skin, tissue, and muscles by multi-layer dense meshes. Facial surface deformation is triggered by the contractions of the synthetic facial muscles. The muscle forces are propagated through the skin layer and finally deform the facial surface. The simulation procedure solves a set of dynamics equations. This kind of approach can achieve very realistic animation results. However, the physical models are sophisticated and computationally complicated. In addition, how to decide the values of a large set of parameters in a physics-based face model is an art.

B. Facial Motion Analysis

It is well known that tracking facial motions based on the low-level facial image features (e.g., edges or facial feature points) alone is not robust. Model-based facial motion tracking algorithms achieve more robust results by using some high-level knowledge models [38], [39], [24], [40]. Those high-level models correspond to the facial deformation control models and encode information about possible facial deformations. The tracking algorithms first extract the control information by combining the low-level image information, which is obtained by low-level image processing (e.g., edge detection, skin/lip color segmentation, template matching, optical flow calculation, etc.), and the high-level knowledge models. The control information is used to deform the face model. The deformation results are the tracking results of the current time stamp and are used by the tracking algorithms in the next step.

The final tracking results will be greatly degraded if a biased high-level knowledge model is used in this loop. To be faithful to the real facial deformations, the high-level knowledge models should be learned from labelled real facial deformations.

C. Real-Time Speech-Driven Face Animation

Some approaches train the audio-to-visual mapping using hidden Markov models (HMMs) [41], [42], [25], which have relative long time delay. Some approaches attempt to generate lip shapes in real-time using only one audio frame. Those approaches use vector quantization [43], affine transformation [44], Gaussian mixture model [45], or artificial neural networks [46], [47] in the audio-to-visual mapping.

Vector quantization [43] is a classification-based audio-to-visual conversion approach. The audio features are classified into one of a number of classes. Each class is then

mapped to a corresponding visual output. Though it is computationally efficient, the vector quantization approach often leads to discontinuous mapping results. The affine transformation approach [44] maps the audio feature to the visual feature by simple linear matrix operations. The Gaussian mixture approach [45] models the joint probability distribution of the audio-visual vectors as a Gaussian mixture. Each Gaussian mixture component generates a linear estimation for a visual feature given an audio feature. The estimations of all the mixture components are then weighted to produce the final visual estimation. The Gaussian mixture approach produces smoother results than the vector quantization approach does. Morishima and Harashima [46] trained a three layer neural network to map the LPC Cepstrum coefficients of each speech segment to the mouth-shape parameters for five vowels. Kshirsagar and Magnenat-Thalmann [47] also trained a three-layer neural network to classify each speech segment into vowels. The average energy of the speech segment is then used to modulate the lip shape of the recognized vowel.

However, those approaches in [43], [44], [45], [46], [47] do not consider the audio contextual information, which is very important for modelling mouth coarticulation due to speech producing. Many other approaches also train neural networks for audio-to-visual mapping while taking into account the audio contextual information. Massaro et al. [48] trained multilayer perceptrons (MLP) to map the LPC cepstral parameters of speech signals to face animation parameters. They modelled the mouth coarticulation by considering the audio context of eleven consecutive audio frames (five backward, current, and five forward frames). Another way to model the audio context is to use time delay neural networks (TDNNs) model, which uses ordinary time delays to perform temporal processing. Lavagetto [49] and Curinga et al. [50] train TDNNs to map the LPC cepstral coefficients of speech signals to lip animation parameters. Nevertheless, the neural networks used in [48], [49], [50] have a large number of hidden units in order to handle large vocabulary, which results in high computational complexity during the training phrase.

The above speech-driven face animation approaches mainly focus on how to train the audio-to-visual mappings. They do not consider the problem of facial deformations. A sound audio-to-visual mapping may not lead to sound speech-driven face animation results if an inappropriate visual representation is used for modelling facial deformations. Moreover, most of them can not synthesize facial expressions.

III. The Integrated Framework

It has been shown above that speech-driven face animation is closely related to facial deformation modelling and facial motion analysis. Here, we present an integrated framework that systematically addresses face modelling, facial motion analysis, and audio-to-visual mapping (see Figure 1). The framework provides a systematic guideline for building a speech-driven synthetic talking face.

First, a quantitative representation of facial deformations, called the Motion Units (MUs), is learned from a set of labelled real facial deformations. It is assumed that any facial deformation can be approximated by a linear combination of MUs weighted by the MU parameters (MUPs). MUs can be used not only for face animation but also as the high-level knowledge model in facial motion tracking. MUs and the MUPs form a facial deformation control model. A MU-based face model can be animated by adjusting the MUPs. Second, a robust MU-based facial motion tracking algorithm is presented to analyze facial image sequences. The tracking results are represented as MUP sequences. Finally, a set of facial motion tracking results and the corresponding speech streams are collected as the audio-visual training data. The audio-visual database is used to train a real-time audio-to-MUP mapping using neural networks.

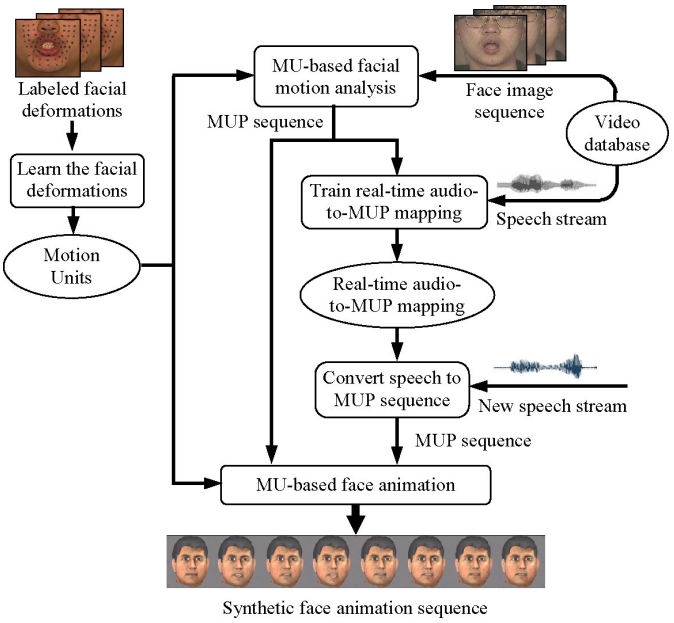


Fig. 1. An integrated framework for face modelling, facial motion analysis and synthesis.

IV. Motion Units

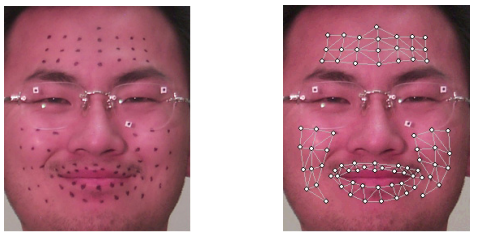
MU is inspired by the Action Units of the Facial Action Coding System (FACS), proposed by Ekman and Friesian [51]. FACS is designed by observing stop-motion video and considered to be the most popular visual representation for facial expression recognition. An Action Unit corresponds to an independent motion of the face. However, Action Units do not provide quantitative temporal and spatial information required by face animation. To utilize FACS, researchers need to manually design Action Units for their models [24], [52].

A. Learning MUs

MUs are learned from a set of labelled facial deformation data and serve as the basic information unit, which links the components of the framework together. MUs define the facial deformation manifold. Each MU represent an axis in the manifold. For computational simplicity, we assume that a facial deformation can be approximated by

a linear combination of the MUs and apply principal component analysis (PCA) [53] to learning the MUs. PCA is a popular tool for modelling facial shape, deformation, and appearance [54], [39], [55], [56]. PCA captures the second-order statistics of the data by assumeing the data has a Gaussian distribution. We categorize the MUs into the Utterance MUs and the Expression MUs. Each expression has a corresponding set of Expression MUs. The Utterance MUs capture the characteristics of the facial deformations caused by speech production. The Expression MUs capture the residual information which is mainly due to facial expressions and beyond the modelling capacity of the Utterance MUs.

We collect the facial deformation data of a speaking subject with and without expressions. We mark some points in the face of the subject (see Figure 2 (a)). The facial deformations are represented by the deformation of the markers. Twenty-two points are marked on the forehead of the subject. Twenty-four makers are put in the cheeks. Thirty markers are placed around the lips. The number of the markers decides the representation capacity of the MUs. More markers enable the MUs to encode more information. Currently, we only deal with 2D facial deformations. The same method can be applied to 3D facial deformations when 3D facial deformations are available. A mesh model is created according to those markers (see Figure 2 (b)). The mesh model that corresponding to the neutral face is used as the mesh model in the MU-based facial motion tracking algorithm, which will be described in Section V. The subject is asked to wear a pair of glasses, where three additional markers are placed.



(a) The markers. (b) The mesh model.

Fig. 2. The markers and the mesh model.

We tempt to include as great a variety of facial deformations as possible in the training data and capture the facial deformations of the subject while he is pronouncing all English phonemes with and without expressions. The video is digitized at 30 frame per second, which results in more than 1000 samples for each expression. The markers are automatically tracked by zero-mean normalized cross correlation template matching technique [57]. A graphic interactive interface is developed for the user to correct the positions of trackers when the template matching fails due to large face or facial motions. To compensate the global face motion, the tracking results are aligned by affine transformations so that the markers on the glasses are coincident for all the data samples. After aligning the data, we calculate the deformations of the markers with respect to the positions of the markers in the neutral face.

The deformations of the markers at each time frame are concatenated to form a vector. We use $D_0 = \{\overrightarrow{d}_{0i}\}_{i=1}^{N_0}$ to denote the facial deformation vector set without expressions and use $D_k = \{\overrightarrow{d}_{ki}\}_{i=1}^{N_k}$ $(1 \le k \le K)$ to denote the facial deformation vector set with the kth expression. First, D_0 is used to learn the Utterance MUs M_0. We obtain $\overrightarrow{m}_{00} = E[\overrightarrow{d}_{0i}]$ and $\Lambda_0 = E[(\overrightarrow{d}_{0i} - \overrightarrow{m}_{00})(\overrightarrow{d}_{0i} - \overrightarrow{m}_{00})^T]$. The eigenvectors and eigenvalues of Λ_0 are calculated. The first A_0 (in our case, $A_0 = 7$) significant eigenvectors $\{\overrightarrow{m}_{0a}\}_{a=1}^{A_0}$ which correspond to the largest A_0 eigenvalues, are selected. They account for 97.56% of the facial deformation variation in D_0. The Utterance MUs are denoted as $M_0 = \{\overrightarrow{m}_{0a}\}_{a=0}^{A_0}$.

We then calculate the Expression MUs for each expression as following. For each $D_k = \{\overrightarrow{d}_{ki}\}_{i=1}^{N_k}$, we calculate $R_k = \{\overrightarrow{r}_{ki}\}_{i=1}^{N_k}$ so that

$$\overrightarrow{r}_{ki} = \overrightarrow{d}_{ki} - \sum_{j=1}^{A_0} \overrightarrow{d}_{ki}^T \overrightarrow{m}_{0j} \overrightarrow{m}_{0j} \tag{1}$$

$\overrightarrow{r}_{ki}$ is the residual information that beyond the modelling capability of M_0. We then apply PCA to R_k and obtain the kth Expression MU set $M_k = \{\overrightarrow{m}_{0a}\}_{a=0}^{A_k}$, where $\overrightarrow{m}_{k0} = E[\overrightarrow{r}_{ki}]$ and $\{\overrightarrow{m}_{kb}\}_{b=1}^{A_k}$ are the first A_k significant eigenvectors of the covariance matrix of R_k. We find $A_k = 2$ $(1 \le k \le K)$ is able to capture at least 98.13% residual information of the collected data.

B. MU and Face Animation

MUs have some nice properties. First, MUs are learned from real data and encode the characteristics of real facial deformations. Second, the way that MUs are calculated considers the correlation between the deformations of the facial points represented by the markers. Third, the number of the MUs is much smaller that of the vertices on the face model. Only a few parameters need to be adjusted in order to animate the face model. It only requires very low bandwidth to transmit the those parameters over the networks. A facial deformation $\overrightarrow{d}$ can be calculated by linearly combining MUs

$$\overrightarrow{d} = \sum_{k=0}^{K} \alpha_k \left(\sum_{i=1}^{A_k} c_{ki} \overrightarrow{m}_{ki} + \overrightarrow{m}_{k0}\right) \tag{2}$$

where

- $\alpha_0 = 1$ is a constant.
- $\alpha_k = 1$ $(K \ge k \ge 1)$ if and only if the expression state is k. Otherwise $\alpha_k = 0$.[1]
- $\{c_{0i}\}_{i=1}^{A_0}$ is the Utterance MUP (UMUP) set.
- $\{c_{ki}\}_{i=1}^{A_k}$ $(1 \le k \le K)$ is the Expression MUP (EMUP) set of M_k.

It can be easily shown that MU-based face animation technique is compatible with the linear keyframe-based face animation technique, which is widely used. This is

[1]We assume that the face can only be in one expression state at any time.

very important from the industrial point of view. The linear keyframe-based face animation technique animates the face model by interpolating among a set of keyframes, say $\{\vec{\kappa}_i\}_{i=1}^{P}$, where P is the number of the keyframes. Since the face models used in different face animation system may be different, we can establish the correspondence at the semantic level defined by the keyframes. We can find a set of training samples $\{\vec{\kappa}'_i\}_{i=1}^{P}$ in the training set of MUs so that $\vec{\kappa}'_i$ semantically corresponds to $\vec{\kappa}_i$ for $1 \le i \le P$. We can then use $\{\vec{\kappa}'_i\}_{i=1}^{P}$ to derive the following expressions.

A facial shape $\vec{s}$ can be represented as a weighted combination of the keyframes as $\vec{s} = \sum_{i=1}^{P} b_i \vec{\kappa}'_i$, where $\{b_i\}_{i=1}^{P}$ is the keyframe parameter set. $\vec{s}$ can also be represented by MUs as

$$\vec{s} = \sum_{k=0}^{K} \alpha_k \Big(\sum_{i=1}^{A_k} c_{ki}\vec{m}_{ki} + \vec{m}_{k0}\Big) + \vec{s}_0 \tag{3}$$

where $\vec{s}_0$ the shape of neutral face.

The conversion between the MUPs and the keyframe parameters can be achieved by

$$\begin{aligned}
\vec{c}_{0j} &= \vec{m}_{0j}^{T}[\mathcal{K}\vec{b} - \vec{s}_0 - \vec{m}_{00}] \\
\vec{c}_{kj} &= \alpha_k \vec{m}_{kj}^{T}[\mathcal{K}\vec{b} - \vec{s}_0 - \sum_{i=1}^{A_0} c_{0i}\vec{m}_{0i} - \vec{m}_{00} - \vec{m}_{k0}] \\
\vec{b} &= (\mathcal{K}^T\mathcal{K})^{-1}\mathcal{K}^T\Big(\sum_{k=0}^{K}\alpha_k\Big(\sum_{i=1}^{A_k} c_{ki}\vec{m}_{ki} + \vec{m}_{k0}\Big) + \vec{s}_0\Big)
\end{aligned} \tag{4}$$

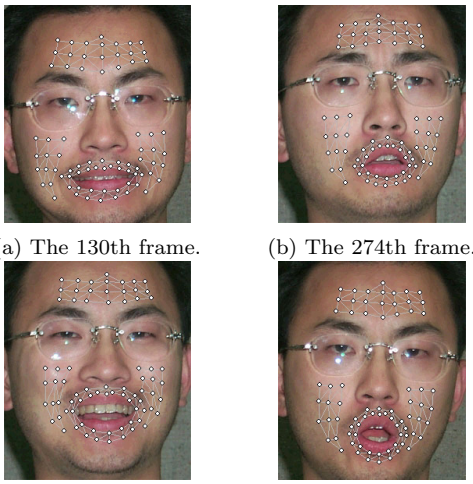
where $\vec{b} = [b_1 \dots b_P]^T$ and $\mathcal{K} = [\vec{\kappa}'_1 \cdots \vec{\kappa}'_P]$.

The conversion from MUPs to the keyframe parameters provides a method for normalizing the facial deformations of different subjects. In other words, the facial deformations of different subjects are normalized at the semantic level. Potentially, this method could benefit research on computer lip-reading and expression recognition.

V. MU-based Facial Motion Analysis

MUs can be used as the high-level knowledge model to guide facial motion tracking. We assume that the expression state k of the subject is known. This is reasonable because we are more interested in using the tracking algorithm to collect the training data for speech-driven face animation research. We also assume an affine motion model, which is a good approximation when the size of the object is relative much smaller than the distance between the object and the camera and the face only undergoes relative small global 3D motion. The tracking procedure consists of two steps. First, at the low-level image processing step, we calculated the facial shape in the next image by tracking each facial point separately using the approach proposed in [58]. The results are usually very noisy and denoted as $\vec{\zeta}^{(t)}$. We then constrain that the facial deformation should be in the manifold defined by MUs.

Mathematically, the tracking problem can be formulated as a minimization problem



(a) The 130th frame. (b) The 274th frame. (c) The 411th frame. (d) The 496th frame.

Fig. 3. Typical tracking results.

$$\begin{aligned}
(\mathbb{C}^*, \vec{\beta}^*) = \arg\min_{\mathbb{C},\vec{\beta}} \| \vec{\zeta}^{(t)} - T_{\vec{\beta}}\Big(\sum_{k=0}^{K}\alpha_k\Big(\sum_{i=1}^{A_k} c_{ki}\vec{m}_{ki} \\
+ \vec{m}_{k0}\Big) + \vec{s}_0\Big)\|^2
\end{aligned} \tag{5}$$

where $\mathbb{C} = \{c_{ki}\}$ and $T_{\vec{\beta}}(\bullet)$ is the transformation function whose parameter $\vec{\beta}$ describes the global affine motion (2D rotation, scaling and translation) of the face. The readers are asked to refer to Appendix A for the details about solving eq. (5).

The MU-based facial motion tracking algorithm requires that the face be in the neutral state and face the camera in the first image frame so that the mesh model can be fit to the neutral face. The mesh model has two vertices corresponding to two mouth corners. Two mouth corners are manually selected in the facial image. The mesh model is fit to the face by translation, scaling and rotation. The task of tracking is to track those facial points represented by the vertices of the mesh. Figure 3 shows some typical tracking results in an image sequence. Currently, the tracking algorithm can only track 2D facial motion. However, if a 3D facial deformation training data is available, we can learn 3D MUs. Substituting 3D MUs into Eq. 5, we can track 3D face and facial motion.

VI. Real-Time Audio-to-MUP Mapping Using Neural Networks

Audio-visual training data are required to train the real-time audio-to-MUP mapping. We collect the audio-visual training data in the following way. A subject is asked to read a text corpus with and without expressions. We video tape the speaking subject and digitize the video at 30 frame per second. The sampling rate of the audio is 44.1 kHz. The MU-based facial motion tracking algorithm in Section V is used to analyze the facial image sequence of the video. The tracking results that are represented as MUP sequences and used as the visual feature vectors of

the AV database. We calculate ten Mel-frequency cepstrum coefficients (MFCC) [59] of each audio frame as the audio feature vector. We collect an audio-visual database without expression $\Psi_0 = \{\langle \overrightarrow{a}_{0i}, \overrightarrow{v}_{0i}\rangle\}_{i=1}^{H_0}$, where $\overrightarrow{a}_{0i}$ is the audio feature vector and $\overrightarrow{v}_{0i}$ is the visual feature vector. For each expression k, we collect an audio-visual database as $\Psi_0 = \{\langle \overrightarrow{a}_{ki}, \overrightarrow{v}_{ki}\rangle\}_{i=1}^{H_k}$, where $\overrightarrow{a}_{ki}$ is the audio feature vector and $\overrightarrow{v}_{ki}$ is the visual feature vector.

We use a method that is similar to boosting to train a set of neural networks for real-time audio-to-MUP mapping. First, Ψ_0 is used to train a set of MLPs, say $\{\Xi_{0i}\}$, as the real-time audio-to-UMUP mapping. For each Ψ_k, the trained $\{\Xi_{0i}\}$ is first used to estimate the UMUP component of $\overrightarrow{v}_{ki}$ given the corresponding audio features $\overrightarrow{a}_{ki}$. An MLP, say Υ_k, is then trained for each Ψ_k to map the estimated UMUPs to $\overrightarrow{v}_{ki}$, which includes the final UMUPs and the EMUPs.

A. Audio-to-UMUP Mapping

An MLP is a universal nonlinear function approximator and has been successfully used to train audio-to-visual mapping [13], [48]. The best results were reported by Massaro el al. [48]. They trained only one MLP for audio-to-visual mapping while considering the audio context of eleven consecutive audio frames. Hence, the MLP used in [48] has large number of hidden units (The best results are achieved by using an MLP with 600 hidden units). We found that it is in practice difficult to use just one MLP to handle the whole audio-to-visual mapping due to the large searching space and high computational complexity in the training phrase. We divided Ψ_0 into 44 subsets according to the audio feature vector $\overrightarrow{a}_{0i}$.[2] The audio features of subset are modelled by a Gaussian mixture. Each audio-visual sample $\langle \overrightarrow{a}_{0i}, \overrightarrow{v}_{0i}\rangle$ is classified into one of the 44 subsets whose Gaussian mixture gives the highest score for $\overrightarrow{a}_{0i}$.

A three-layer perceptron Ξ_{0i} is trained to perform audio-to-visual mapping using each subset. The input of Ξ_{0i} is the audio feature vectors taken at seven consecutive time frames (3 backward, current and 3 forward time windows). Those 3 backward and 3 forward audio frames are the context of the current audio frame. Therefore, the delay between the input and the output is about 100 ms. In the estimation stage, an audio feature vector $\overrightarrow{a}$ is first classified into one of the 44 subsets using those Gaussian mixtures. The corresponding MLP is selected to estimate the visual feature given $\overrightarrow{a}$ and the its contextual information. In our experiments, the maximum number of the hidden units used in $\{\Xi_{0i}\}_{i=1}^{44}$is only 25 and the minimum number of the hidden units is 15. Therefore, both training and estimation have very low computational complexity.

B. Audio-to-UMUP+EMUP Mapping

A straightforward way to build the mapping for speech-driven expressive talking face is to retrain a new set of MLPs for each Ψ_k. This problem can be greatly simplified by taking advantage of the correlation between the facial deformations without expressions and facial deformations with expressions that account for the same speech content. The mapping can then be divided into two steps. The first step maps the speech to the UMUPs, which represent the facial deformation caused by producing speech. The second step maps the estimated UMUP of the first step to the final UMUPs and EMUPs. The function of the second step is to add expression information to the results of the first step. Therefore, we can reuse $\{\Xi_{0i}\}_{i=1}^{44}$ that are trained in the previous subsection and train an MLP to perform the task of the second step for each Ψ_k.

The trained $\{\Xi_{0i}\}_{i=1}^{44}$ is used to estimate the UMUPs for each audio feature vector $\overrightarrow{a}_{ki}$ in Ψ_k. Of course, the estimation results will not be accurate and do not contain expression information. An MLP Υ_k with one hidden layer is further trained to map the estimated UMUPs to the visual feature vector $\overrightarrow{v}_{ki}$ of $\overrightarrow{a}_{ki}$. In our experiments, the number of the hidden units of the Υ_k is only thirty. Therefore, this approach is computationally very efficient.

VII. Experimental Results

A. Numeric Evaluation

We collect the audio-visual training database by recording the front view of a speaking subject with and without expressions. Currently, we only examine two expressions: smile and sad. One hundred sentences are selected from the text corpus of the DARPA TIMIT speech database. Both the audio and video are digitized at thirty frame per second, which results in 19563 audio-visual training data samples. The sampling rate of the audio is 44.1 kHz. The MU-based facial motion tracking algorithm in Section V is used to analyze the facial image sequence of the video. The tracking results are represented as MUP sequences and used as the visual feature vector in the audio-visual database. Ten MFCCs are calculated for each audio frame as the audio features. Eighty percent of the data is randomly selected for training. The rest is used for testing.

We reconstruct the estimated displacements of the facial feature points using MUs and the estimated MUPs. We divide the displacement (both the ground truth and the estimated results) of each facial feature point by the its maximum absolute displacement in the collected audio-visual database so that the displacement is normalized to [-1.0, 1.0]. To evaluate the performance, we calculate the Pearson product-moment correlation coefficients (R), the average standard deviations, and the mean square errors (MSEs) using the normalized data. The Pearson product-moment correlation coefficient measures how good the global match between the shapes of two signal sequences is. It is calculated as

$$R = \frac{trace(Cov_{\overrightarrow{d}\,\overrightarrow{d}'})}{\sqrt{trace(Cov_{\overrightarrow{d}\,\overrightarrow{d}})trace(Cov_{\overrightarrow{d}'\,\overrightarrow{d}'})}} \tag{6}$$

where $\overrightarrow{d}$ is the normalized ground truth, $\overrightarrow{d}'$ is the normalized estimated result, and

[2]The reason of choosing 44 classes is that we use a phoneme symbol set that consists of 44 phonemes.

$$\overrightarrow{\mu}_{\overrightarrow{d}} = E[\overrightarrow{d}]$$
$$\overrightarrow{\mu}_{\overrightarrow{d}'} = E[\overrightarrow{d}']$$
$$Cov_{\overrightarrow{d}\,\overrightarrow{d}'} = E((\overrightarrow{d} - \overrightarrow{\mu}_{\overrightarrow{d}})(\overrightarrow{d}' - \overrightarrow{\mu}_{\overrightarrow{d}'})^T)$$
$$Cov_{\overrightarrow{d}\,\overrightarrow{d}} = E((\overrightarrow{d} - \overrightarrow{\mu}_{\overrightarrow{d}})(\overrightarrow{d} - \overrightarrow{\mu}_{\overrightarrow{d}'})^T)$$
$$Cov_{\overrightarrow{d}',\overrightarrow{d}'} = E((\overrightarrow{s}' - \overrightarrow{\mu}_{\overrightarrow{d}})(\overrightarrow{d}' - \overrightarrow{\mu}_{\overrightarrow{d}'})^T)$$

We also calculate the the average standard deviations

$$\nu_{\overrightarrow{d}} = \frac{\sum_{c=1}^{\gamma}(Cov_{\overrightarrow{d}\,\overrightarrow{d}}[c][c])^{1/2}}{\gamma}$$
$$\nu_{\overrightarrow{d}'} = \frac{\sum_{c=1}^{\gamma}(Cov_{\overrightarrow{d}',\overrightarrow{d}'}[c][c])^{1/2}}{\gamma} \quad (7)$$

where γ is the dimension of $\overrightarrow{d}$. The MSEs are calculated by

$$MSE = E\left[\frac{\|\overrightarrow{d} - \overrightarrow{d}'\|^2}{\gamma}\right] \quad (8)$$

The results are shown in Table I and II. The "Neutral" column shows the results of Ψ_0. The "Smile" column shows the results of the audio-visual data with smile expression. The "Sad" column shows the results of the audio-visual data with sad expression.

TABLE I
THE NUMERIC EVALUATION RESULTS OF THE TRAINING SET.

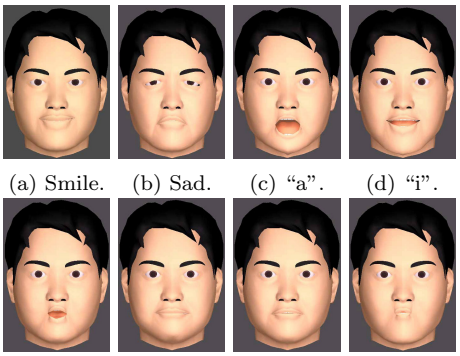
	Training set		
	Neutral	Smile	Sad
R	0.980	0.972	0.977
$\nu_{\overrightarrow{d}}$	0.197	0.205	0.209
$\nu_{\overrightarrow{d}'}$	0.179	0.192	0.197
MSE	0.0025	0.0031	0.0033

TABLE II
THE NUMERIC EVALUATION RESULTS OF THE TESTING SET.

	Testing set		
	Neutral	Smile	Sad
R	0.968	0.945	0.942
$\nu_{\overrightarrow{d}}$	0.196	0.208	0.213
$\nu_{\overrightarrow{d}'}$	0.184	0.202	0.208
MSE	0.0029	0.0034	0.0037

B. Generate Face Animation Sequence

We have developed a face modelling and animation system, called the iFACE system [1]. The iFACE system uses a generic geometric face mesh model and uses the linear keyframe technique to animation the face model. The technique described in Section IV-B is used to convert the estimated MUPs into the keyframe parameters. It enables the iFACE system to utilize MUs without undergoing large modification. Currently, eight keyframes are used. They are smile, sad, and six visemes, which correspond to six phonemes "i", "a", "o", "f", "u", and "m", respectively. The keyframes are shown using the generic geometric face model in Figure 4.



(a) Smile. (b) Sad. (c) "a". (d) "i".
(e) "o". (f) "m". (g) "f". (h) "u".

Fig. 4. The keyframes that are used for the conversion between MUPs and the keyframe parameters.

Given the Cyberware^{TM} scanner data of an individual, the user can use the iFACE system to interactively customize the generic model for that individual by clicking some facial feature points. The facial texture can be mapped onto the customized model to achieve realistic appearance (see Figure 5).

(a) Dr. Russell L. Storms. (b) The actress.

Fig. 5. The customized face models.

Figure 6 shows a speech-driven face animation sequence generated by the iFACE system using the method described in this paper. The speech content of the animation sequence is: "Dialog is an essential element." Figure 7 shows three typical frames in three animation sequences with or without expressions. Those three frames in Figure 7 share the same speech context while baring different expressions.

Fig. 6. An example of real-time speech-driven face animation.

C. Human Emotion Perception Study

Since the end user of the real-time speech-driven synthetic talking face are human beings, it is necessary to carry out human perception study on the synthetic talking face. For convenience, we will use synthetic talking face or synthetic face to denote our real-time speech-driven synthetic

(a) Neutral. (b) Smile. (c) Sad.

Fig. 7. Typical face animation frames.

talking face in the rest of the paper. Here, we design the following experiments to compare the influence of the synthetic talking face on human emotion perception with that of the real face. The experimental results can help the user with how to use the synthetic talking face to deliver the intended visual information.

We video tape a speaking subject, whose audio-visual data is used in Section VII-A. The subject is asked to calmly read three sentences without expression, with smile expression, or with sad expression. Hence, the audio tracks do not convey any emotional information. The content of the first sentence is "It is normal.", which contains neutral information. The content of the second sentence is "It is good.", which contains positive information. The content of the third sentence is "It is bad.", which contains negative information. The audio tracks are used to generate three sets of face animation sequences. All three audio tracks are used in each set of animation sequence. The first set is generated without expression. The second set is generated with smile expression. The third set is generated with sad expression. Twenty untrained human subjects, who never used our system before, participate the experiments.

The first experiment investigates human emotion perception based on either the visual stimuli alone or the audio stimuli alone. The subjects are first asked to recognize the expressions of both the real face and the synthetic talking face and infer their emotional states based on the animation sequences without audio. All subjects correctly recognized the expressions of both the synthetic face and the real face. Therefore, our synthetic talking face is capable to accurately deliver facial expression information. The emotional inference results in terms of the number of the subjects are shown in Table III. The "S" columns show the results using the synthetic talking face. The "R" columns show the results using the real face. As shown, the effectiveness of the synthetic talking face is comparable with that of the real face.

TABLE III

Emotion inference based on the animation sequences without audio.

		Facial Expression					
		Neutral		Smile		Sad	
		S	R	S	R	S	R
Emotion	Neutral	20	20	3	2	0	0
	Happy	0	0	17	18	0	0
	Sad	0	0	0	0	20	20

The subjects are then asked to listen to the audio and decide the emotional state of the speaker. Each subject listens to each audio only once. Note that the audio tracks are produced without emotions. Hence, the subjects try to infer the emotion from the content of the speech tracks. The results in terms of the number of the subjects are shown in Table IV.

TABLE IV

Emotion inference based on the audio.

		Audio 1	Audio 2	Audio 3
Emotion	Neutral	20	7	6
	Happy	0	13	0
	Sad	0	0	14

The second and third experiments are designed to compare the influence of synthetic face on bimodal human emotion perception and that of the real face. In the second experiment, the subjects are asked to infer the emotional state while observing the synthetic talking face and listening to the audio tracks. In the third experiment, the subjects are asked to infer the emotional state while observing the real face and listening to the same audio tracks. We divide the subjects into two groups. Each of them has ten subjects. One group first participates the second experiment and then participates the third experiment. The other group first participates the third experiment and then participates the second experiment. The results are then combined and compared in Table V, VI, and VII. The "S" columns in Table V, VI, and VII show the results using the synthetic talking face. The "R" columns in Table V, VI, and VII show the results using the real face.

TABLE V

Bimodal emotion inference using audio track 1.

		Facial Expression					
		Neutral		Smile		Sad	
		S	R	S	R	S	R
Emotion	Neutral	20	20	4	2	1	0
	Happy	0	0	16	18	0	0
	Sad	0	0	0	0	19	20

TABLE VI

Bimodal emotion inference using audio track 2.

		Facial Expression					
		Neutral		Smile		Sad	
		S	R	S	R	S	R
Emotion	Neutral	16	17	2	0	0	0
	Happy	4	3	18	20	0	0
	Sad	0	0	0	0	12	15
	Not sure	0	0	0	0	8	5

We can see the face movements (either synthetic or real) and the content of the audio tracks jointly influence the

TABLE VII
BIMODAL EMOTION INFERENCE USING AUDIO TRACK 3.

		Facial Expression					
		Neutral		Smile		Sad	
		S	R	S	R	S	R
Emotion	Neutral	13	14	14	13	0	0
	Happy	0	0	2	4	0	0
	Sad	7	6	0	0	20	20
	Not sure	0	0	4	3	0	0

decisions of the subjects. Let's take the first audio track as an example. Although the first audio track only contains neutral information, sixteen subjects think the emotional state is happy if the expression of the synthetic talking face is smile. And nineteen subjects classify the emotional state into sad if the expression of the synthetic face is sad. The influence of sad expression is slightly stronger than that of smile expression. This may be because the subjects see smile expression more frequently than sad expression in the daily life. Therefore, the subjects react more strongly when they see sad expression.

If the audio tracks and the facial represent the same kind of information, the human perception on the information will be enhanced. For example, when the associated facial expression of the audio track 2 is smile, nearly all subjects say that the emotional state is happy (see Table VI). The numbers of the subjects who agree with happy emotion are higher than those using visual stimuli alone (see Table III) or audio information alone (see Table IV).

However, it will confuse human subjects if the facial expressions and the audio tracks represent opposite information. For example, many subjects are confused when they listen to an audio track, which contains positive information, and observe a facial expression, which represents negative information. An example is shown in the seventh and eighth columns of Table VI. The audio track conveys positive information while the facial expression is sad. Eight subjects report that they are confused if the synthetic talking face with sad expression is shown. The number of the confused subjects reduces to five if the real face is used. This difference is mainly due to the fact that the subjects are still able to tell the synthetic talking face from the real face. When confusion happens, the subjects tend to think that the expression of the synthetic face is not the original expression associating with the audio. Therefore, when the visual information conflicts with the audio information, the real face is more persuasive than this version of synthetic face. In other words, the synthetic face is less capable of conveying fake emotion information in this kind of situation.

Overall, the experimental results show that our real-time speech-driven synthetic talking face successfully affects human emotion perception. The effectiveness of the synthetic face is comparable with that of the real face though it is slightly weaker.

VIII. Summary and Discussions

This paper presents an integrated framework for systematically building a real-time speech driven talking face for an individual. To handle the non-Gaussianity of facial deformation distribution, we assume that the facial deformation space can be represented by a hierarchical linear manifold described by the Utterance MUs and the Expression MUs. The Utterance MUs define the manifold representing the facial deformations caused by speech production. The Expression MUs capture the residual information which is mainly due to facial expressions and beyond the modelling capacity of the Utterance MUs. PCA is applied to learning both the Utterance MUs and the Expression MUs.

The MU-based face animation technique animates a face model by adjusting the parameters of the MUs. We also show that the MU-base face animation technique is compatible with the linear keyframe-based face animation technique. In fact, this provides a method to normalize the facial deformations of different people at the semantic level.

We propose an MU-based facial motion analysis algorithm that explains the facial deformations into UMUPs and EMUPs. The algorithm is used to obtain the visual information for an audio-visual database. We train a set of MLPs for real-time speech-driven face animation with expressions using the collected audio-visual database. The audio-to-visual mapping consists of two steps. The first step maps the audio features to UMUPs. The second step maps the estimated UMUPs calculated by the first step to the final UMUPs and the EMUPs. To evaluate the mapping, we calculate the normalized MSEs and the Pearson product-moment correlation coefficients between the ground truth and the estimated results. The Pearson product-moment correlation coefficients of the training set are 0.98 for no expression, 0.972 for smile expression, and 0.977 for sad expression, respectively. The Pearson coefficients of the testing set are 0.968 for no expression, 0.945 for smile expression, and 0.942 for sad expression, respectively. The normalized MSEs of the training set are 0.0025 for no expression, 0.0031 for smile expression, and 0.0033 for sad expression, respectively. The normalized MSEs of the testing set are 0.0029 for no expression, 0.0034 for smile expression, and 0.0037 for sad expression, respectively.

Using the proposed method, we develop the function of real-time speech-driven face animation with expressions for the iFACE system. The iFACE system is then used in the bimodal human emotion perception study. We generate three sets of face animation sequences for three audio tracks, which convey neutral, positive, and negative information respectively. Each set of the face animation sequences consist of three sequences, which contain no expression, smile, and sad, respectively. Human subjects are asked to infer the emotion states from the face animation sequences or the videos of the real face while listening to the corresponding audio tracks. The experiment results show that our synthetic talking face effectively contribute to the bimodal human emotion perception and its effects are comparable with a real talking face. To extensively evaluate our method, future work on the bimodal human

emotion perception study using the iFACE system will use a larger subject set and more audio/visual stimuli data .

Acknowledgments

This research is supported partially by USA Army Research Laboratory under Cooperative Agreement No. DAAL01-96-2-0003, and partially by National Science Foundation Grant IIS-00-85980.

Appendix A

Here, we show how to solve the following minimization problem

$$(\mathbb{C}^*, \overrightarrow{\beta}^*) = \arg\min_{\mathbb{C}, \overrightarrow{\beta}} || \overrightarrow{\zeta}^{(t)} - T_{\overrightarrow{\beta}}\big(\sum_{k=0}^{K} \alpha_k (\sum_{i=1}^{A_k} c_{ki} \overrightarrow{m}_{ki} + \overrightarrow{m}_{k0}) + \overrightarrow{s}_0 \big) ||^2$$

We first define the notations which will be used to derive the results:

a. We track N facial feature points.

b. $\overrightarrow{\zeta}^{(t)} = [x_1^{(t)}\ y_1^{(t)}\ \ldots\ x_N^{(t)}\ y_N^{(t)}]^T$, where $\langle x_n^{(t)}, y_n^{(t)} \rangle$ is the coordinate of facial feature point n in the image plane at time t ($t > 0$) and N is the number of the facial feature points.

c. $\overrightarrow{\beta} = [\beta_1\ \beta_2\ \beta_3\ \beta_4\ \beta_5\ \beta_6]^T$, where β_1, β_2, β_3, and β_4 describe 2D rotation and scaling, and β_5 and β_6 describe 2D translation.

d. $\overrightarrow{m}_{ki} = [m_{11}^{ki}\ m_{12}^{ki}\ \ldots\ m_{N1}^{ki}\ m_{N2}^{ki}]^T$, where $\langle m_{n1}^{ki}, m_{n2}^{ki} \rangle$ denotes the deformation information of the facial feature point n, which is encoded by $\overrightarrow{m}_{ki}$.

e. $\overrightarrow{s}_0 = [x_1^0\ y_1^0\ \ldots\ x_N^0\ y_N^0]^T$, where $\langle x_n^0, y_n^0 \rangle$ is the coordinate of facial feature point n in the neutral position.

We can then write down

$$\begin{aligned} &\Big\| \overrightarrow{\zeta}^{(t)} - T_{\overrightarrow{\beta}}\big(\textstyle\sum_{k=0}^{K} \alpha_k (\sum_{i=1}^{A_k} c_{ki} \overrightarrow{m}_{ki} + \overrightarrow{m}_{k0}) + \overrightarrow{s}_0 \big) \Big\|^2 \\ &= \sum_{n=1}^{N} \left\| \begin{bmatrix} x_n^{(t)} \\ y_n^{(t)} \end{bmatrix} - \begin{bmatrix} \beta_1 & \beta_2 & \beta_5 \\ \beta_3 & \beta_4 & \beta_6 \end{bmatrix} \begin{bmatrix} x_n \\ y_n \\ 1 \end{bmatrix} \right\|^2 \end{aligned} \tag{9}$$

where

$$x_n = \sum_{k=0}^{K} \alpha_k (\sum_{i=1}^{A_k} c_{ki} m_{n1}^{ki} + m_{n1}^{k0}) + x_n^0 \tag{10}$$

and

$$y_n = \sum_{k=0}^{K} \alpha_k (\sum_{i=1}^{A_k} c_{ki} m_{n2}^{ki} + m_{n2}^{k0}) + y_n^0 \tag{11}$$

Note that $\alpha_k = 1$ ($k > 0$) if and only if the facial is in the expression state k. We also constrain that the face can only be in one expression state at any time. Without losing generality, we can assume $\alpha_k = 0$ for $k > 1$. Eq. (9) can be then rewritten as:

$$\| B \overrightarrow{v} - \overrightarrow{\zeta}^{(t)} \|^2 \tag{12}$$

where

$$\begin{aligned} B &= [H_0\ B_1\ \ldots\ B_{A_0}\ E_1\ \ldots\ E_{A_1}] \\ H_0 &= [H_{01}\ H_{02}] \\ H_{01} &= \begin{bmatrix} m_{11}^{00} + m_{11}^{10} + x_1^0 & m_{12}^{00} + m_{12}^{10} + y_1^0 & 1 \\ 0 & 0 & 0 \\ \ldots & \ldots & \ldots \\ m_{N1}^{00} + m_{N1}^{10} + x_N^0 & m_{N2}^{00} + m_{N2}^{10} + y_N^0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \\ H_{02} &= \begin{bmatrix} 0 & 0 & 0 \\ m_{11}^{00} + m_{11}^{10} + x_1^0 & m_{12}^{00} + m_{12}^{10} + y_1^0 & 1 \\ \ldots & \ldots & \ldots \\ 0 & 0 & 0 \\ m_{N1}^{00} + m_{N1}^{10} + x_N^0 & m_{N2}^{00} + m_{N2}^{10} + y_N^0 & 1 \end{bmatrix} \\ B_i &= \begin{bmatrix} m_{11}^{0i} & m_{12}^{0i} & 0 & 0 \\ 0 & 0 & m_{11}^{0i} & m_{12}^{0i} \\ \ldots & \ldots & \ldots & \ldots \\ m_{N1}^{0i} & m_{N2}^{0i} & 0 & 0 \\ 0 & 0 & m_{N1}^{0i} & m_{N2}^{0i} \end{bmatrix} \quad (A_0 \geq i \geq 1) \\ E_i &= \begin{bmatrix} m_{11}^{1i} & m_{12}^{1i} & 0 & 0 \\ 0 & 0 & m_{11}^{1i} & m_{12}^{1i} \\ \ldots & \ldots & \ldots & \ldots \\ m_{N1}^{1i} & m_{N2}^{1i} & 0 & 0 \\ 0 & 0 & m_{N1}^{1i} & m_{N2}^{1i} \end{bmatrix} \quad (A_1 \geq i \geq 1) \\ \overrightarrow{v} &= [\overrightarrow{\phi}\ \overrightarrow{\psi}_1\ \ldots\ \overrightarrow{\psi}_{A_0}\ \overrightarrow{\xi}_1\ \ldots\ \overrightarrow{\xi}_{A_1}]^T \\ \overrightarrow{\phi} &= [\beta_1\ \beta_2\ \beta_5\ \beta_3\ \beta_4\ \beta_6] \\ \overrightarrow{\psi}_i &= [\beta_1 c_{0i}\ \beta_2 c_{0i}\ \beta_3 c_{0i}\ \beta_4 c_{0i}]\ (A_0 \geq i \geq 1) \\ \overrightarrow{\xi}_i &= [\beta_1 c_{1i}\ \beta_2 c_{1i}\ \beta_3 c_{1i}\ \beta_4 c_{1i}]\ (A_1 \geq i \geq 1) \end{aligned}$$

We can use a least square estimator to solve $\overrightarrow{v}$ from eq. (12). It is easy to recover $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ form $\overrightarrow{v}$, and then calculate $\{c_{0i}\}_{i=1}^{A_0}$ and $\{c_{1i}\}_{i=1}^{A_1}$.

References

[1] P. Hong, Z. Wen, and T. S. Huang, "iface: a 3d synthetic talking face," *International Journal of Image and Graphics*, vol. 1, no. 1, pp. 1 – 8, 2001.

[2] J. Cassell et al., "Animated conversation: Rule-based generation of facial display, gesture and spoken intonation for multiple conversational agents," in *Proc. SIGGRAPH*, 1994, pp. 413–420.

[3] Ananova Limited, *The virtual newscaster*, http://www.ananova.com.

[4] Haptek Corporate, *VirtualFriend*, http://www.haptek.com/.

[5] Inc. LifeFX, *Facemail*, http://www.lifefx.com/.

[6] K. Nagao and A. Takeuchi, "Speech dialogue with facial displays," in *Proc. 32nd Annual Meeting of the Asso. for Computational Linguistics (ACL-94)*, 1994, pp. 102–109.

[7] K. Waters, J. M. Rehg, et al., "Visual sensing of humans for active public interfaces," Tech. Rep. CRL 96-5, Cambridge Research Lab, 1996.

[8] D. W. Massaro, *Speech perception by ear and eye: A paradigm for psychological inquiry*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.

[9] I. Pandzic, J. Ostermann, and D. Millen, "User evaluation: Synthetic talking faces for interactive services," *The Visual Computer*, vol. 15, no. 7/8, pp. 330–340, 1999.

[10] K. Aizawa and T. S. Huang, "Model-based image coding," *Proc. IEEE*, vol. 83, pp. 259–271, Aug. 1995.

[11] W. H. Leung et al., "Networked intelligent collaborative environment (netice)," in *IEEE Intl. Conf. on Multimedia and Expo*, Jul. 2000.

[12] P. Hong, *An Integrated Framework for Face Modeling, Facial Motion Analysis and Synthesis*, Ph.D. thesis, University of Illinois at Urbana-Champaign, Urbana, IL 61801, 2001.

[13] S. Morishima, "Real-time talking head driven by voice and its application to communication and entertainment," in *Proceedings of the International Conference on Auditory-Visual Speech Processing*, Terrigal, Australia, Dec. 1998.

[14] S. Morishima and T. Yotsukura, "Face-to-face communicative avatar driven by voice," in *IEEE International Conference on Image Processing*, Kobe, Japan, 1999.

[15] T. Capin, I. Pandzic, N. M. Thalmann, and D. Thalmann, *Avatars in Networked Virtual Environments*, John Wiley & Sons, 1999.

[16] K. R. Petrushin, "Adding the affective dimension: A new look in speech analysis and synthesis," in *Proc. International Conference on Spoken Language Processing 1996*, Philadelphia, PA, USA, Oct. 1996, pp. 1808–1811.

[17] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc International Conf. on Spoken Language Processing*, Philadelphia, PA, USA, 1996, pp. 1970–1973.

[18] V. A. Petrushin, "How well can people and computers recognize emotions in speech?," in *Papers from the 1998 AAAI Fall Symposium*, 1998, pp. 141–145.

[19] V. A. Petrushin, "Emotion recognition in speech signals: experimental study, development, and application," in *Proc. International Conference on Spoken Language Processing*, 2000.

[20] K. Mase, "Recognition of facial expression from optical flow," *ICICE Transactions*, vol. E74, pp. 3474–3483, Oct. 1991.

[21] A. Lanitis, C. J. Taylor, and T. F. Cootes, "A unified approach to coding and interpreting face images," in *Proc. International Conference on Computer Vision*, 1995, pp. 368–373.

[22] M. J. Black and Y. Yacoob, "Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion," in *Proc. International Conference on Computer Vision*, 1995, pp. 371–384.

[23] Y. Yacoob and L. Davis, "Recognizing human facial expressions from long image sequences using optical flow," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, pp. 636–642, Jan. 1996.

[24] I. A. Essa and A. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 10, no. 7, pp. 757 – 763, July 1997.

[25] M. Brand, "Voice puppetry," in *Proc. SIGGRAPH*, 1999.

[26] Y. Li, F. Yu, Y. Xu, E. Chang, and H. Shum, "Speech-driven cartoon animation with emotions," in *Proc. ACM Multimedia*, Ottawa, Canada, Oct. 2001.

[27] M. Nahas, H. Huitric, and M. Saintourens, "Animation of a b-spline figure," *The Visual Computer*, vol. 3, pp. 272–276, 1988.

[28] G. M. Nielson, "Scattered data modeling," *IEEE Computer Graphics and Applications*, vol. 13, no. 1, pp. 60–70, 1993.

[29] L. Williams, "Performance-driven facial animation," *Computer Graphics*, vol. 21, no. 2, pp. 235–242, 1990.

[30] F. Pighin et al., "Synthesizing realistic facial expressions from photographs," in *Proc. SIGGRAPH '98*, 1998.

[31] P. Kalra, A. Mangili, N. Magnenat Thalmann, and D. Thalmann, "Simulation of facial muscle actions based on rational free form deformations," in *Proc. Eurographics'92*, 1992, pp. 59–69.

[32] F. I. Parke, *A parametric model of human faces*, Ph.D. thesis, University of Utah, 1974.

[33] F. I. Parke, "A parameterized model for facial animation," *IEEE Computer Graphics and Applications*, vol. 2, no. 9, pp. 61–70, 1982.

[34] A. Pearce, B. Wyvill, G. Wyvill, and D. Hill, "Speech and expression: A computer solution to face animation," in *Graphics Interface*, 1986.

[35] K. Waters, "A muscle model for animating three-dimensional facial expressions," *Computer Graphics*, vol. 21, no. 4, pp. 17 – 24, Jul. 1987.

[36] D. Terzopoulos and K. Waters, "Techniques for realistic facial modeling and animation," in *Computer Animation*. Springer-Verlag, Tokyo, Japan, 1991.

[37] Y. C. Lee, D. Terzopoulos, and K. Waters, "Realistic modeling for facial animation," in *Proc. SIGGRAPH*, 1995, pp. 55–62.

[38] H. Li, P. Roivainen, and R. Forchheimer, "3-D motion estimation in model-based facial image coding," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 545–555, 1993.

[39] T. F. Cootes, C. J. Taylor, et al., "Active shape models - their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, Jan. 1995.

[40] D. DeCarlo and D. Mataxas, "Optical flow constraints on deformable models with applications to face tracking," *International Journal of Computer Vision*, vol. 38, no. 2, pp. 99–127, 2000.

[41] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communications," *Proc. IEEE, Special Issue on Multimedia Signal Processing*, vol. 86, no. 5, pp. 837–852, May. 1998.

[42] F. J. Huang and T. Chen, "Real-time lip-synch face animation driven by human voice," in *IEEE Workshop on Multimedia Signal Processing*.

[43] S. Morishima, K. Aizawa, and H. Harashima, "An intelligent facial image coding driven by speech and phoneme," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Glasgow, UK, 1989.

[44] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1-2, pp. 23–43, 1998.

[45] R. Rao, T. Chen, and R. M. Mersereau, "Exploiting audio-visual correlation in coding of talking head sequences," *IEEE Trans. on Industrial Electronics*, vol. 45, no. 1, pp. 1522, 1998.

[46] S. Morishima and H. Harashima, "A media conversion from speech to facial image for intelligent man-machine interface," *IEEE J. Selected Areas in Communications*, vol. 4, pp. 594–599, 1991.

[47] S. Kshirsagar and N. Magnenat-Thalmann, "Lip synchronization using linear predictive analysis," in *Proceedings of IEEE International Conference on Multimedia and Expo*, New York, USA, 2000.

[48] D. W. Massaro, J. Beskow, M. M. Cohen, et al., "Picture my voice: Audio to visual speech synthesis using artificial neural networks," in *Proc. AVSP'99*, 1999.

[49] F. Lavagetto, "Converting speech into lip movements: A multimedia telephone for hard of hearing people," *IEEE Transactions on Rehabilitation Engineering*, vol. 3, no. 1, Mar. 1995.

[50] S. Curinga, F. Lavagetto, and F. Vignoli, "Lip movements synthesis using time-delay," in *Proc. EUSIPCO-96*, Trieste, 1996.

[51] P. Ekman and W. V. Friesen, *Facial action coding system*, Consulting Psychologists Press, Inc., Palo Alto, Calif., 1978.

[52] H. Tao and T. Huang, "Explanation-based facial motion tracking using a piecewise bezier volume deformation model," in *Proc. IEEE conference on Computer Vision and Pattern Recognition*, 1999.

[53] I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, 1986.

[54] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1994, pp. 669–672.

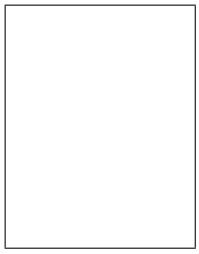
[55] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proc. 5th European Conference on Computer Vision*, 1998, vol. 2, pp. 484–498.

[56] P. Hong, X. Lin, and T. S. Huang, "Mouth motion learning and generating from observation," in *IEEE Workshop on Multimedia Signal Processing*, Los Angeles, California, Dec. 1998.

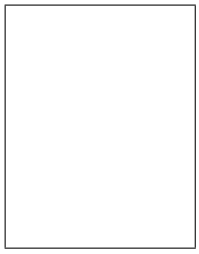
[57] O. Faugeras, *Three-Dimensional Computer Vision: a Geometric Viewpoint*, MIT Press, Palo Alto, Calif., 1993.

[58] J. Shi and C. Tomasi, "Good features to track," in *Proc. IEEE CVPR*, 1994, pp. 593–600.

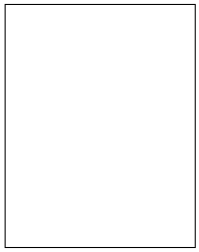
[59] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.



Pengyu Hong received the B. Engr. and M. Engr. degree, both in computer science, from Tsinghua University, Beijing, China, in 1995 and 1997, respectively. He received his Ph.D. degree in Computer Science from the University of Illinois at Urbana-Champaign, Urbana, Illinois, USA in 2001. His research interests include Human-Computer Interaction, Computer Assisted Human-Human Interaction, Computer Vision and Pattern Recognition, Machine Learning, and Multimedia Database. In 2000, he received the Ray Ozzie fellowship for his research work on face modelling, facial motion analysis and synthesis. His research interests include face modelling and animation, face and facial motion tracking, speech driven face animation, unsupervised temporal pattern extraction, unsupervised spatial pattern modelling, and multimedia database.



Zhen Wen received the B. Engr. degree from Tsinghua University, Beijing, China, and MS degree from University of Illinois at Urbana-Champaign, Urbana, Illinois, USA, both in computer science. Currently he is a Ph.D. student in the Department of Computer Science at University of Illinois at Urbana-Champaign. His research interests are face modelling, facial motion analysis and synthesis, image based modelling and rendering.



Thomas S. Huang Thomas S. Huang received his B.S. degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan, China; and his M.S. and Sc.D. degrees in Electrical Engineering from the Massachusetts Institute of Technology, Cambridge, Massachusetts. He was on the Faculty of the Department of Electrical Engineering at MIT from 1963 to 1973, and on the faculty of the School of Electrical Engineering and Director of its Laboratory for Information and Signal Processing at Purdue University from 1973 to 1980. In 1980, he joined the University of Illinois at Urbana-Champaign, where he is now William L. Everitt Distinguished Professor of Electrical and Computer Engineering, Research Professor at the Coordinated Science Laboratory, and Head of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology.

During his sabbatical leaves, Dr. Huang has worked at the MIT Lincoln Laboratory, the IBM Thomas J. Watson Research Center, and the Rheinishes Landes Museum in Bonn, West Germany, and held visiting professor positions at the Swiss Institutes of Technology in Zurich and Lausanne, University of Hannover in West Germany, INRS-Telecommunications of the University of Quebec in Montreal, Canada, and University of Tokyo, Japan. He has served as a consultant to numerous industrial firms and government agencies both in the U.S. and abroad.

Dr. Huang's professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals. He has published 12 books and over 300 papers in network theory, digital filtering, image processing, and computer vision. He is a Fellow of the International Association of Pattern Recognition, IEEE, and the Optical Society of American; and has received a Guggenheim Fellowship, an A.V. Humboldt Foundation Senior U.S. Scientist Award, and a fellowship from the Japan Association for the Promotion of Science. He received the IEEE Acoustics, Speech, and Signal Processing Society's Technical Achievement Award in 1987, and the Society Award in 1991. He is a Founding Editor of the International Journal Computer Vision, Graphics, and Image Processing, and Editor of the Springer Series in Information Sciences, published by Springer-Verlag.