A Statistical Quality Model for Data-Driven Speech Animation

Xiaohan Ma, *Student Member*, *IEEE*, and Zhigang Deng, *Senior Member*, *IEEE*

Abstract—In recent years, data-driven speech animation approaches have achieved significant successes in terms of animation quality. However, how to automatically evaluate the realism of novel synthesized speech animations has been an important yet unsolved research problem. In this paper, we propose a novel statistical model (called SAQP) to automatically predict the quality of on-the-fly synthesized speech animations by various data-driven techniques. Its essential idea is to construct a phoneme-based, Speech Animation Trajectory Fitting (SATF) metric to describe speech animation synthesis errors and then build a statistical regression model to learn the association between the obtained SATF metric and the objective speech animation synthesis quality. Through delicately designed user studies, we evaluate the effectiveness and robustness of the proposed SAQP model. To the best of our knowledge, this work is the first-of-its-kind, quantitative quality model for data-driven speech animation. We believe it is the important first step to remove a critical technical barrier for applying data-driven speech animation techniques to numerous online or interactive talking avatar applications.

Index Terms—Facial animation, data-driven, visual speech animation, lip-sync, quality prediction, statistical models

1 Introduction

During the past several decades, plenty of research efforts have been focused to generate realistic speech animation given novel spoken or typed input [1]. In particular, state-of-the-art data-driven speech animation approaches have achieved significant successes in terms of animation quality. One of the main reasons is that these techniques heavily exploit the prior knowledge encoded in precollected training facial motion data sets, by concatenating presegmented motion samples (e.g., triphone or syllable-based motion subsequences) [2], [3], [4], [5], [6], [7], [8] or learning facial motion statistical models [9], [10], [11].

On the other side, all the above data-driven approaches are only empirically tested, that is, researchers first generate a small number of novel speech animations based on selected texts (or speech) and then evaluate the animations via tedious user studies. It is obvious that such user studies can only be used to evaluate offline synthesized speech animations, since it is infeasible to use offline user studies to assess the quality of on-the-fly synthesized speech animations. Therefore, *can we automatically predict the quality of dynamically synthesized speech animations without conducting actual user studies*?

In addition, a number of different data-driven speech animation algorithms are available these days (e.g., assuming all these algorithms run at the back end of an online or interactive talking avatar application), and their online performances could be varied depending on specific inputs. Here is a simple yet conceptual example: for a specific inputted sentence, maybe the first algorithm generates a better speech animation than the second one; while for another sentence input, this situation could be reversed, that is, the second algorithm could outperform the first one. Thus, an interesting question is: *Can we dynamically compare and determine which algorithm (among them) can synthesize the best speech animation for specific text or speech input*? Indeed, in spite of the practical importance of the automated evaluation of data-driven speech animations, no plausible solution has yet been proposed and validated to date.

Inspired by the above challenge, in this work, we propose a novel statistical model to automatically predict the quality of synthesized speech animations on-the-fly generated by various data-driven algorithms. Fig. 1 is a schematic illustration of the proposed approach. In the training stage, we construct a phoneme-based, *Speech Animation Trajectory Fitting (SATF)* metric to describe speech animation synthesis errors. Then, we build a statistical regression model, called the *Speech Animation Quality Prediction (SAQP)* model, to learn the association between the obtained SATF metric and the ground-truth speech animation synthesis quality. At the end, given any new text/voice input (i.e., a phoneme sequence with timing information) and a given training facial motion data set, the constructed SAQP model can be used to predict the quality of on-the-fly synthesized data-driven speech animations. We also conduct delicately designed user studies to evaluate the effectiveness and robustness of our SAQP model.

To the best of our knowledge, this work is the first-of-its-kind, automated quality model for data-driven speech animation. We believe it is the important first step to remove a critical technical barrier for applying wealthy

- *X. Ma is with the Computer Graphics Lab, Department of Computer Science, University of Houston, PGH 309, 4800 Calhoun Road, Houston, TX 77204-3010. E-mail: xiaohan@cs.uh.edu.*
- *Z. Deng is with the Department of Computer Science, University of Houston, PGH 501, 4800 Calhoun Road, Houston, TX 77204-3010. E-mail: zdeng@cs.uh.edu.*

Manuscript received 6 Aug. 2011; revised 22 Dec. 2011; accepted 29 Jan. 2012; published online 17 Feb. 2012.
Recommended for acceptance by R. Boulic.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCG-2011-08-0179.
Digital Object Identifier no. 10.1109/TVCG.2012.67.

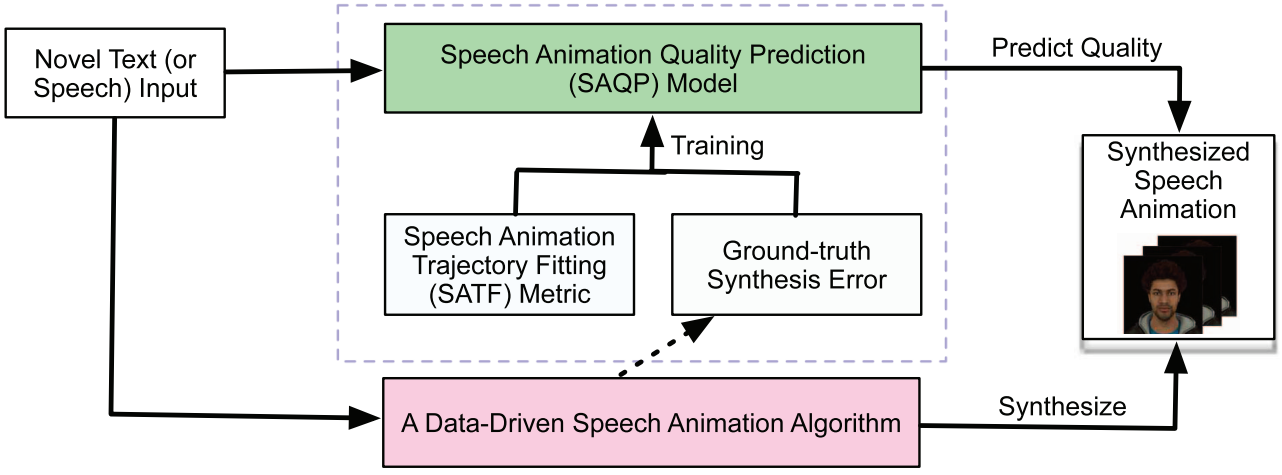

Fig. 1. The introduced speech animation quality prediction model can automatically predict the quality of synthetic speech animations that are generated by a data-driven algorithm based on novel text or speech input.

data-driven speech animation techniques to numerous online or interactive talking avatar applications.

It is noteworthy that, although various data-driven speech animation approaches had been proposed, in general, they can be roughly classified into the following two categories: *sample-based* and *learning-based*. As such, we select representative algorithms from each of the two categories to evaluate our SAQP model. From existing sample-based speech animation approaches [2], [3], [4], [5], [6], [7], we choose the *Anime-Graph* approach [5] and the *eFASE* approach [6]. Among current learning-based approaches [9], [10], [11], we extend the classical Multidimensional Morphable Model (MMM) proposed by Ezzat et al. [10] from 2D to 3D for 3D lip-sync generation. Section 4 encloses more technical and implementation details of the three chosen approaches. We believe that a methodology successfully applied to the above three chosen data-driven approaches [5], [6], [10] can be soundly generalized to other data-driven speech animation algorithms.

The remainder of this paper is organized as follows. Section 2 briefly reviews recent related research efforts. Section 3 describes how we collect and process a facial motion data set used in this work. Section 5 details how we construct the proposed SAQP statistical model. Section 6 presents the application of our SAQP model and user study results. Finally, discussion and concluding remarks are given in Section 7.

2 Related Work

Researchers have conducted extensive research efforts on facial animation including face modeling [12], [13], [14], [15], [16], [17], deformation [18], [19], [20], [21], [22], and expression transferring and editing [23], [24], [25], [26], [27]. Comprehensively reviewing these efforts is beyond the scope of this paper (interested readers can refer to the recent facial animation survey by Deng and Noh [28]). Here, we only briefly review recent efforts most related to this work.

2.1 Speech Animation

Traditional speech animation approaches typically require users to design visemes (i.e., key mouth shapes), and then empirical smooth functions [29], [30], [31] or coarticulation rules [32] are used to synthesize novel speech animations. For example, in the early era, a linear prediction model is employed to generate lip-sync animations given novel sound track [29]. In the Cohen-Massaro coarticulation model [30] and its various extensions [31], a viseme shape is defined via hand-crafted dominance functions in terms of certain facial measurements such as the width and height of the mouth. In this way, final mouth shapes at animation time are determined as the weighted sum of dominance values.

In recent years, a large variety of data-driven speech animation approaches utilize a precollected facial motion data set to produce realistic lip-sync animations corresponding to new inputted texts [2], [9], [10], [3], [4], [5], [6], [7], [11]. For example, Brand [9] learns Hidden Markov Models (HMMs) from aligned video and audio tracks through an entropy minimization learning algorithm. In his approach, facial motions can be directly synthesized by inputting audio track into the learned HMMs. Ezzat et al. [10] learn a multidimensional morphable model from a set of mouth prototype images (corresponding to basic visemes) and then generate facial motion trajectories in the constructed MMM space for any desired utterance.

Meanwhile, Bregler et al. [2] proposed the concept of recombining triphone segments, extracted from precollected video footages, to generate new speech animations. Along this line, different semantic speech-motion segments such as syllable motions [3] and multiphoneme motions [4], [5], [6], [7], [8] have also been explored. Also, various search strategies including greedy search [5] and constraint-guided dynamic programming [6], [8] are adapted in these algorithms. The above data-driven facial animation approaches often focus on the accuracy or efficacy of their generation algorithms, while little attention has been paid to automatically evaluate the quality of their synthesized speech animations.

2.2 Perceptual Approaches for Animation

In computer graphics and animation community, many approaches have been proposed to determine the visual quality of an image [33], [34] or a clip of animation [35], [36]. Quality metrics or heuristics have also been developed to measure or predict the fidelity of images and rendering for character animations in recent years. For instance,

Hodgins et al. [37] explored the perceptual effect of different geometric models for character animation. O'Sullivan et al. [38] evaluated the visual fidelity of physically based animations. Nonetheless, the above approaches still need time-consuming and delicately designed subjective user studies.

Exploiting human perception and psychophysical insight for graphics and animation applications has attracted a lot of attention in recent years [39]. For example, researchers conducted extensive psychophysical experiments to evaluate the perceptual quality of animated facial expressions or talking heads [40], [41], [42], [43], [44]. Geiger et al. [41] performed two types (implicit and explicit) of perceptual discrimination tasks to tell whether a video-realistic speech animation clip is real or synthetic. Cosker et al. [42] evaluated the behavioral quality of synthetic talking heads based on the "McGurk Effect" test. In addition, Wallraven et al. conducted a series of psychophysical experiments to study the perceptual quality of animated facial expressions [44] or stylized facial expression representations [40].

Ma et al. [45] quantitatively analyze how human perception is affected by audio-head motion characteristics of talking avatars, specifically, quantifying the correlation between perceptual user ratings (obtained via user study) and joint audio-head motion features as well as head motion patterns in the frequency-domain. Recently, Deng and Ma [46] proposed a computational perceptual metric to evaluate synthetic facial expressions such as expression type and scale. However, the same methodology cannot be straightforwardly applied to the case of evaluating synthetic speech animations. Arguably, one of the main reasons is that audio-motion synchronization and speech coarticulation pose additional technical challenges to the task.

3 Experimental Data Set

To quantify data-driven speech animation, we need to experiment with a collection of facial motion data (as the training facial motion data set). In this work, we acquired a facial motion capture data set for this purpose. Specifically, natural speaking of a chosen native English female speaker was recorded, with a 120 Hz sampling rate, by an optical motion capture system.

As pointed out in the existing literature [47], talking might affect the entire facial regions (e.g., the cheek moves when mouth is opened). Thus, we captured facial motion of the whole face instead of the lip region only. A total of 95 facial markers were put on the face and head (90 markers on the face and 5 markers on the head, refer to Fig. 2). The captured subject was directed to speak a custom phoneme-balanced corpus consisting of 237 sentences. In addition, the facial marker motions and aligned acoustic speech were recorded simultaneously. Subsequently, we removed the 3D rigid head motion of each frame by computing a rotation matrix based on the markers on the head, and performed phoneme-alignments (i.e., align each phoneme with its corresponding motion capture subsequence) using the Festival system [48].

3.1 Region-Based Reduced Facial Motion Representation

In order to train our statistical quality model, we need to transform the original, high-dimensional facial motion data to a compact, low-dimensional representation. In this work, we choose to transform the original 3D facial motion to a region-based, reduced representation. Its basic idea is to partition the whole face into different facial regions using a physically motivated facial segmentation scheme [49] and then apply Principal Component Analysis (PCA) to motions of the markers in each region, separately. The region-based PCA representation encodes more intuitive correspondences between PCA eigen-vectors and localized facial movements than applying a single PCA to the whole face [50]. Fig. 2 shows the facial region segmentation result in this work.

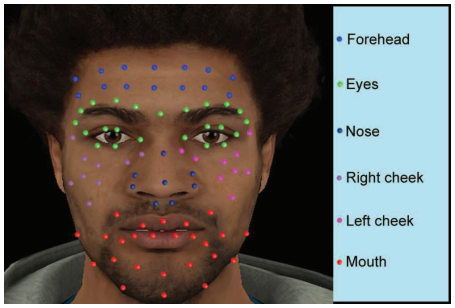

Fig. 2. Illustration of the facial region segmentation result in this work. Different marker colors represent distinct regions.

Based on the above facial region segmentation, we obtain the following six facial regions: forehead, eye, the left cheek, the right cheek, mouth, and nose. For each facial region, we apply Robust Principal Component Analysis (RPCA) [51] to reduce its dimension and construct a truncated PCA space. Though a part-based PCA parameterization of facial motions is essentially linear, we choose PCA parameterization over other forms of nonlinear deformer-based facial motion parameterization schemes (e.g., face rigging [27]) due to its efficiency and characteristic of localized facial movement mapping [50], [46]. In this work, to retain more than 95 percent of the motion variations, the retained dimension is 4 for the forehead region, 4 for the eye region, 3 for the left cheek region, 3 for the right cheek region, 8 for the mouth region, and 5 for the nose region. In this way, we can transform any facial motion capture frame into a region-based, reduced representation. In follow-up sections, our model mainly deals with this reduced facial motion representation.

4 Selected Data-Driven Algorithms for Evaluation

To increase the readability of this paper, we briefly describe the three data-driven speech animation approaches chosen in this work: the Anime-Graph approach [5], the eFASE approach [6], and 3D extension of the MMM-based approach [10], as follows: for more technical details of the three approaches, please refer to their original publications [5], [6], [10].

It is noteworthy that this work is only focused on speech animation, and emotion is not its main focus. In addition, among the three chosen algorithms, the MMM-based approach [10] cannot deal with "emotional visual speech" while the other two [5], [6] can deal with it. Therefore, to

this end, in our implementations we intentionally did not utilize the emotion part in the Anime-Graph and eFASE approaches [5], [6] in order to have sound and fair evaluations on the three algorithms. We are aware that evaluating the quality of emotional visual speech would be an important future research topic to explore.

- **The anime-graph approach [5].** It first constructs a large set of interconnected anime-graphs from a precollected facial motion data set, and each anime-graph essentially encloses aligned facial motion, acoustic features, emotion label and phoneme information. During the synthesis process, given novel speech input, its algorithm searches for an optimal concatenation of anime-graphs using a greedy search strategy.
- **The eFASE approach [6].** At the offline data processing stage, it extracts facial motion nodes (or called viseme segments) from a precollected facial motion data set, and each motion node represents the facial motion of a phoneme. Then, its algorithm further clusters and organizes the motion nodes based on their phoneme labels. At runtime, its algorithm searches for a minimal cost path to optimally concatenate viseme segments through a constraint-based dynamic programming algorithm, and various constraints can be interactively specified by users.
- **3D extension of the MMM-based approach [10].** The original MMM-based approach first trains a multidimensional morphable model from a small set of 2D talking face video clips. At runtime, a novel trajectory (i.e., facial motion and texture parameters) in the MMM space is optimized based on novel phoneme sequence input. The original MMM-based approach is only for 2D talking face generation. In this work, we straightforwardly extend it from 2D to 3D for 3D lip-sync generation.

5 Speech Animation Quality Model

In this section, we describe how we construct a phoneme-based, *Speech Animation Quality Prediction* model for data-driven speech animation. Assuming an inputted phoneme sequence to a data-driven synthesis algorithm, Ψ, is $\mathbf{P} = \{\mathbf{P_1}, \mathbf{P_2}, \ldots, \mathbf{P_T}\}$ and the used facial motion data set is denoted as $\mathbf{M}$, we aim to construct an SAQP model $\mathbf{Q} = F_\Psi(\mathbf{P}, \mathbf{M})$ that can automatically predict the quality of synthesized speech animation corresponding to $\mathbf{P}$, based on the training facial motion data set $\mathbf{M}$, by using the data-driven speech animation algorithm Ψ.

We use the following main steps to construct the SAQP model (refer to Fig. 1).

1. As described in follow-up Section 5.1, we first construct a *Speech Animation Trajectory Fitting* metric (denoted as T) to quantify the speech animation synthesis trajectory fitting error, $\mathbf{E}_{pm}$, of any inputted $\mathbf{P}$ based on the given $\mathbf{M}$. The SATF metric does not depend on any specific speech animation synthesis algorithm since it is computed solely based on the inputted $\mathbf{P}$ and the used training data set $\mathbf{M}$. Hence, $\mathbf{E}_{pm} = T(\mathbf{P}, \mathbf{M})$.
2. We also compute the ground-truth speech animation synthesis quality $\mathbf{Q}_{pm}$ of the inputted $\mathbf{P}$ based on the given $\mathbf{M}$ as the Root Mean-Squared-Error (RMSE) between the prerecorded facial motions (i.e., ground-truth) and synthetic motions by the data-driven algorithm Ψ.
3. Then, as described in Section 5.2, given a small number (n) of training samples randomly selected from the prerecorded facial motion data set, we learn a *Gaussian Process Regression (GPR)* model, $\mathbf{F}_\Psi$, to connect the SATF metric, $\mathbf{E} = \{\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_{pm}, \ldots, \mathbf{E}_n\}$, and the ground-truth speech animation synthesis quality, $\mathbf{Q} = \{\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_{pm}, \ldots, \mathbf{Q}_n\}$. In other words, $\mathbf{Q} = \mathbf{F}_\Psi(\mathbf{E})$.
4. Finally, $\mathbf{F}_\Psi$ can be employed to predict the speech animation synthesis quality, $\mathbf{Q}'$, of any new inputted phoneme sequence, P', based on a new facial motion training data set, $\mathbf{M}'$, that is, $\mathbf{Q}' = \mathbf{F}_\Psi(T(\mathbf{P}', \mathbf{M}'))$. Note that $\mathbf{M}'$ is expected to be captured on the same subject as the one in $\mathbf{M}$, with the same facial marker layout.

5.1 Speech Animation Trajectory Fitting Metric

Our Speech Animation Trajectory Fitting metric is built on the concept of speech animation synthesis trajectory fitting. Specifically, we modify the trajectory fitting technique proposed by Ezzat et al. [10] to compute the SATF metric due to the following reasons: 1) it does not depend on any specific speech animation synthesis algorithm, and 2) it is phoneme-based so that phoneme contexts can be naturally exploited in the SATF metric.

The goal of speech animation trajectory fitting is to quantify the synthesis error of an inputted phoneme sequence $\mathbf{P}$ based on the given $\mathbf{M}$ (the training facial motion data set). $\mathbf{P}$ is a stream of phonemes $\{\mathbf{P}_i\}$ that represent the phonetic data of the utterance. Since the audio and facial motion are aligned, we have all the region-based PCA coefficients/parameters (Section 3.1) for any particular phoneme. Inspired by the Ezzat et al.'s work [10], we represent the facial motion characteristics of each phoneme $\mathbf{P}_i$ mathematically as a multidimensional Gaussian with mean and variance of all the facial motion frames.

A *viseme* is defined as the visual representation of a phoneme (i.e., the motion subsequence of a phoneme). In this work, the means and variances of all the visemes in $\mathbf{P}$ are computed based on a given data set $\mathbf{M}$ as follows: the middle frame of a viseme, precisely, the region-based PCA representation of the middle frame (Section 3.1), is chosen as the viseme's *representative sample*. Since the mean and variance of each viseme depend on its phoneme context (i.e., its preceding/following phonemes), its means and variances are computed by only considering its representative samples with the same phoneme context.

In this work, two basic phoneme contexts are considered: *triphone* and *diphone*. To compute the mean and variance of a phoneme $\mathbf{P_i}$, we first find and compute the mean and variance of all the visemes of $\mathbf{P_i}$ with the triphone context $(\mathbf{P_{i-1}}, \mathbf{P_i}, \mathbf{P_{i+1}})$ in $\mathbf{M}$; otherwise, consider its diphone context $(\mathbf{P_i}, \mathbf{P_{i+1}})$; and, the last alternative would be its flat mean and variance (considering all the representative samples of $\mathbf{P_i}$ regardless their phoneme contexts).

Basically, the trajectory fitting problem can be mathematically formed as a regularization problem [52]. As described in the work of [10], for a given phoneme sequence $\mathbf{P}$, the following objective function (1), called the SATF metric in this work, is minimized to find the best fitting facial motion sequence, $\mathbf{y}$.

$$\begin{aligned}\mathbf{E}_{pm} = (\mathbf{y}-\mathbf{U})^T\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D}(\mathbf{y}-\mathbf{U}) \\ + \lambda\mathbf{y}^T\mathbf{W}^T\mathbf{W}\mathbf{y}.\end{aligned} \tag{1}$$

Here, $\mathbf{D}$ is the time duration matrix for all the phonemes in $\mathbf{P}$; $\mathbf{W}$ is a difference operator (i.e., a matrix) used for smoothing the fitted result, $\mathbf{y}$; $\mathbf{U}$ and $\mathbf{V}$ are the mean and variance matrices estimated from the facial motion sequences of phonemes, and the coefficient λ balances the tradeoff between the fitting errors of individual phonemes and the smoothness of phoneme transitions. How to construct $\mathbf{U}$, $\mathbf{V}$, $\mathbf{D}$, and $\mathbf{W}$ and how to solve for $\mathbf{y}$ are detailed in follow-up Section 5.1.1. Finally, we can obtain $\mathbf{E}_{pm}$ by plugging $\mathbf{y}$ into the above (1).

Although there is no explicit dynamic modeling of coarticulation effect in the phoneme sequence, the $\mathbf{V}$ (variances) matrix implicitly generates coarticulation effect. As reported in the work of [10], a small variance indicates that the synthesized trajectory for a particular phoneme is more likely to only pass its own parameter space, and thus this phoneme has little coarticulation effect. Meanwhile, a large variance indicates that the synthesized trajectory for the phoneme is more likely to pass through the neighboring phonemes' parameter spaces, and thus the coarticulation effect of this phoneme can be modeled.

5.1.1 Computing Means and Variances of Visemes through Trajectory Fitting

Assuming $\mathbf{P}$ is the inputted phoneme sequence and $\mathbf{M}$ is a given facial motion data set, the computed mean and variance elements of all the visemes in $\mathbf{P}$ are diagonally packed into vector $\mathbf{U}$ and matrix $\mathbf{V}$, respectively. $\mathbf{U}$ is a vertical concatenation of the mean vector $\mathbf{U}_i$ for $\mathbf{P}_i$ in phoneme sequence, and each $\mathbf{U}_i$ vector is a concatenation of the 27 region-based PCA coefficients (described in Section 3). Here, K is the total number of the used PCA coefficients ($K = 27$ in this work).

$$\mathbf{U} = \begin{bmatrix}\mathbf{U}_1\\ \mathbf{U}_2\\ .\\ .\\ \mathbf{U}_i\\ .\\ .\\ \mathbf{U}_T\end{bmatrix} \qquad \mathbf{U}_i = \begin{bmatrix}\mathbf{U}_i^1\\ \mathbf{U}_i^2\\ .\\ .\\ \mathbf{U}_i^K\end{bmatrix}. \tag{2}$$

$\mathbf{V}$ is a diagonal concatenation of the variance matrix $\mathbf{V}_i$ (for $\mathbf{P}_i$ in the phoneme sequence $\mathbf{P}$). Each $\mathbf{V}_i$ matrix is $K \times K$.

$$\mathbf{V} = \begin{bmatrix}\mathbf{V}_1 & . & . & . & .\\ . & \mathbf{V}_2 & . & . & .\\ . & . & . & . & .\\ . & . & \mathbf{V}_i & . & .\\ . & . & . & . & .\\ . & . & . & . & \mathbf{V}_T\end{bmatrix}, \tag{3}$$

$$\mathbf{V}_i = \begin{bmatrix}\mathbf{V}_i^1 & . & . & .\\ . & \mathbf{V}_i^2 & . & .\\ . & . & . & .\\ . & . & . & \mathbf{V}_i^K\end{bmatrix}.$$

After $\mathbf{U}$ and $\mathbf{V}$ are obtained, we compute the fitted result, $\mathbf{y}$, by minimizing (1) (i.e., setting $\partial\mathbf{E}_{pm}/\partial\mathbf{y} = 0$). At the end, we obtain the following equation:

$$(\mathbf{D}^T\,\mathbf{V}^{-1}\,\mathbf{D} + \lambda\,\mathbf{W}^T\,\mathbf{W})\mathbf{y} = \mathbf{D}^T\,\mathbf{V}^{-1}\mathbf{D}\,\mathbf{U}. \tag{4}$$

Since all the frames are represented as region-based PCA coefficients; therefore, $\mathbf{y}$ is a vertical concatenation of the region-based PCA coefficient vector $\mathbf{y}_i$ at each time step (refer to (5)).

$$\mathbf{y} = \begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ .\\ .\\ \mathbf{y}_i\\ .\\ .\\ \mathbf{y}_T\end{bmatrix} \qquad \mathbf{y}_i = \begin{bmatrix}\mathbf{y}_i^1\\ \mathbf{y}_i^2\\ .\\ .\\ .\\ \mathbf{y}_i^K\end{bmatrix}. \tag{5}$$

The time duration matrix $\mathbf{D}$ is a diagonal duration-weighted matrix that emphasizes shorter phonemes and de-emphasizes longer ones [10]. The purpose of the duration-weighted matrix is to relieve the impact of extremely long phonemes. In this work, we use the following form for each diagonal element of $\mathbf{D}_i$ in the time duration matrix

$$\mathbf{D} : \sqrt{1 - \frac{\mathbf{D}_{\mathbf{P}_i}}{\mathbf{D}_{\mathbf{P}}}}$$

(Here, $\mathbf{D}_{\mathbf{P}_i}$ denotes the duration of $\mathbf{P}_i$ in time frames, and $\mathbf{D}_{\mathbf{P}}$ denotes the length of the entire utterance $\mathbf{P}$ in time frames), and each $\mathbf{D}_i$ is $K \times K$. Thus, $\mathbf{D}$ is formulated as follows:

$$\begin{aligned}\mathbf{D} &= \begin{bmatrix}\mathbf{D}_1 & . & . & . & .\\ . & \mathbf{D}_2 & . & . & .\\ . & . & . & . & .\\ . & . & \mathbf{D}_i & . & .\\ . & . & . & . & .\\ . & . & . & . & \mathbf{D}_T\end{bmatrix},\\ \mathbf{D}_i &= \begin{bmatrix}\sqrt{1-\frac{\mathbf{D}_{\mathbf{P}_i}}{\mathbf{D}_{\mathbf{P}}}} & . & . & .\\ . & \sqrt{1-\frac{\mathbf{D}_{\mathbf{P}_i}}{\mathbf{D}_{\mathbf{P}}}} & . & .\\ . & . & . & .\\ . & . & . & .\\ . & . & . & \sqrt{1-\frac{\mathbf{D}_{\mathbf{P}_i}}{\mathbf{D}_{\mathbf{P}}}}\end{bmatrix}.\end{aligned} \tag{6}$$

The matrix $\mathbf{W}$ is the first order difference operator [53] (Section 5.1.2 describes how we determine the optimal order of $\mathbf{W}$ is 1), and it is used to smooth the fitted result $\mathbf{y}$. Each $\mathbf{I}$ is a $K \times K$ identity matrix

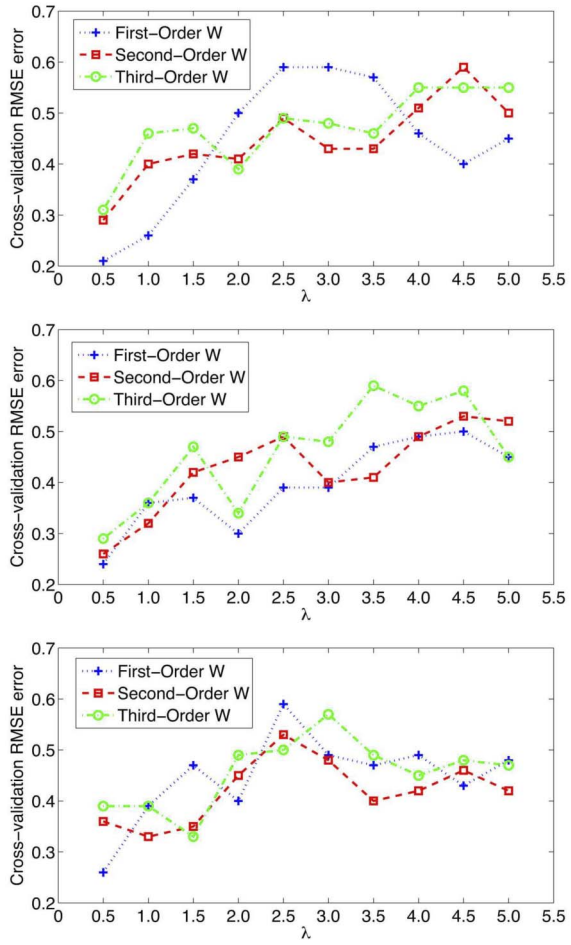

Fig. 3. Illustration of how the cross-validation RMSE error is changed when λ and the order of $\mathbf{W}$ are varied: the anime-graph approach [5] (Top), the eFASE approach [6] (Middle), and the 3D extension of the MMM-based approach [10] (Bottom).

$$\mathbf{W} = \begin{bmatrix} -\mathbf{I} & \mathbf{I} & . & . & . & . \\ . & -\mathbf{I} & \mathbf{I} & . & . & . \\ . & . & -\mathbf{I} & \mathbf{I} & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & -\mathbf{I} & \mathbf{I} \end{bmatrix}. \quad (7)$$

The coefficient λ in (4) balances the tradeoff between the fitting errors of individual phonemes and the smoothness of phoneme transitions. In this work, we experimentally set it to 0.5. How to determine the optimal λ as well as the order of $\mathbf{W}$ is described in follow-up Section 5.1.2.

5.1.2 Determining Optimal Parameters

In the above $\mathbf{W}$, higher orders of smoothness are formed by repeatedly multiplying $\mathbf{W}$ with itself: second order $\mathbf{W}^T\mathbf{W}^T\mathbf{W}\mathbf{W}$, third order $\mathbf{W}^T\mathbf{W}^T\mathbf{W}^T\mathbf{W}\mathbf{W}\mathbf{W}$, and so on. Fig. 3 shows how the cross-validation Root Mean Squared Error is changed when the order of $\mathbf{W}$ and the coefficient λ in Eq. (1) are varied (also refer to Section 5.3). As shown in Fig. 3, for all the three chosen approaches [5], [6], [10], the optimal order of $\mathbf{W}$ is 1 (that is, the optimal $\mathbf{W}$ is a matrix corresponding to the first-order difference operator) and the optimal λ is 0.5, thus we empirically set λ to 0.5 in this work.

TABLE 1
The Phoneme Contexts of the Six Selected Test Sentences and Their Average (per Phoneme) Fitting Errors Computed by Our Introduced SATF Metric

Sentence #	Triphone Percentage	Diphone Percentage	Monophone Percentage	Average Fitting Error
1	30%	53%	17%	0.134
2	34%	48%	18%	0.180
3	37%	45%	18%	0.146
4	39%	42%	19%	0.153
5	41%	41%	18%	0.163
6	44%	40%	16%	0.155

In order to understand how the combination of triphones, diphones, and monophones in the inputted phoneme sequence $\mathbf{P}$ could affect the trajectory fitting error of the constructed SATF metric, we randomly select and analyze six test sentences. Table 1 shows the detailed phoneme contexts of the six selected test sequences and their average (per phoneme) fitting errors computed by our introduced SATF metric (i.e., $\mathbf{E}_{pm}/T$). As shown in this table, since the percentages of combined triphones and diphones in the six test sentences are close, their resultant fitting errors are numerically close accordingly. Also, the direct correlation between triphone/diphone combination and the resultant fitting error is not straightforward. However, quantitatively analyzing and modeling this issue could be an interesting future study.

5.2 Statistical Quality Prediction Model

We split the prerecorded facial motion data set (described in Section 3) into a training subset, $\mathbf{M}_t$ (80 percent, 189 sentences), and a test/validation subset, $\mathbf{M}_v$ (20 percent, 48 sentences). Then, we use $\mathbf{M}_t$ to train the SAQP statistical model that can predict speech animation synthesis quality based on the SATF metric. We detail its modeling procedure in this section.

The construction of our SAQP model is based on the cross-validation mechanism. As illustrated in Fig. 4, we first split the whole training motion data set ($\mathbf{M}_t$) into tenfolds, and each fold has less than 20 sentences. In this phase, we

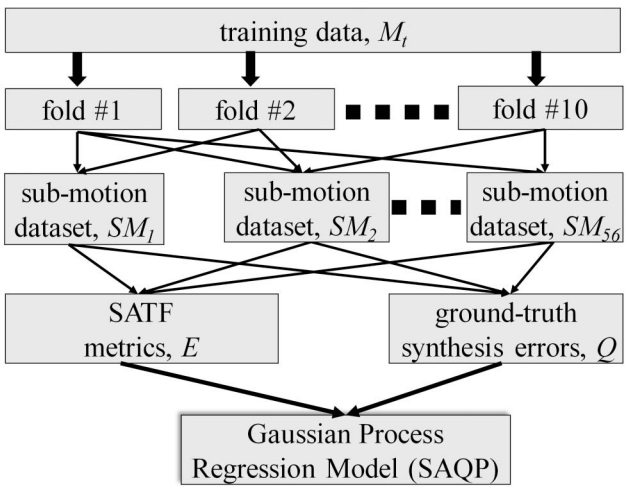

Fig. 4. Illustration of the SAQP modeling process.

keep the split of $\mathbf{M}_t$ in a phoneme-balanced manner. In other words, each of the folds covers facial motions for all the phonemes used in this study. Subsequently, we randomly pick and combine k folds ($1 \leq k \leq 9$) into a submotion data set, $\mathbf{SM}_j$. In this work, a total of 56 submotion data sets are constructed. Then, for each of the submotion data sets, $\mathbf{SM}_j$, we use $\mathbf{M}_t - \mathbf{SM}_j$ as the training data to compute the SATF metric for each sentence in $\mathbf{SM_j}$ based on Eq. (1). In other words, if the ith sentence (i.e., phoneme sequence) in $\mathbf{SM_j}$ is denoted as $\mathbf{SM}_j^i$, and its corresponding SATF metric is denoted as $\mathbf{E}_j^i$, then $\mathbf{E}_j^i = T(\mathbf{SM_j^i}, \mathbf{M}_t - \mathbf{SM}_j)$. In this way, we obtain a total of 484 SATF metrics. For the sake of a clear explanation, we use $\{\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_n\}$ $(\mathrm{n} = 484)$ to denote these obtained SATF metrics.

Meanwhile, for each of the above 484 synthesis tasks such as $\mathbf{E_j^i}$, we also know its corresponding prerecorded (ground-truth) facial motion in $\mathbf{M}_t$. As such, we can compute its ground-truth synthesis error, $\mathbf{Q}_j^i$, by calculating the RMSE error between the synthesized facial motion and the prerecorded ground truth.

The used RMSE error is determined by computing the differences of both the positions and velocities between the synthesized facial motion and the prerecorded ground truth. In this work, we choose a set of facial markers over the whole facial geometry to compute the RMSE error due to the following main reasons: 1) 3D facial meshes of various persons or even the same person typically have different topologies (i.e., the number of vertices) such as the widely used multiresolution mesh representation. Thus, if the RMSE error is computed based on the difference of all the deformed facial mesh vertices, it will directly depend on the used facial mesh representation (as an additional factor to the SAQP model). By contrast, the marker-based RMSE error is independent to the used facial mesh. 2) Computing the marker-based RMSE error is much more efficient than computing all the mesh vertices based RMSE error. In addition, marker-driven facial deformation (e.g., the thin-shell linear deformation model [54], [21]) has been proven to be effective to soundly produce realistic and high-fidelity facial deformations. In this work, we also choose the thin-shell linear deformation model to deform the face mesh based on the displacements of a set of markers.

Specifically, to compute the RMSE error in this work, we consider not only the first-order effect (positions of facial markers), but also the second-order effect (velocities of facial markers) towards and away from the targets. Equation (8) gives the formula to compute the RMSE error $\mathbf{Q}_j^i$ as the ground-truth synthesis error.

$$\mathbf{Q}_j^i = \lambda_1 \sqrt{\frac{\sum_t \left(\omega(\mathbf{X}_t^S - \mathbf{X}_t^G)\right)^2}{F}} + \lambda_2 \sqrt{\frac{\sum_t \left(\omega(\dot{\mathbf{X}}_t^S - \dot{\mathbf{X}}_t^G)\right)^2}{F}}, \tag{8}$$

Here F denotes the total frame number of the ith facial motion sequence in the training data set; $\mathbf{X}_t^S$ and $\mathbf{X}_t^G$ denote the marker positions of the tth frame of the synthesized facial motion and the prerecorded ground truth motion, respectively; $\dot{\mathbf{X}}_t$ denotes the marker velocities of the tth facial motion frame; ω denotes the weights/importances of facial markers; and λ_1 and λ_2 are the weights to balance the position part and the velocity part. In this work, we empirically set λ_1 and λ_2 to 0.64 and 0.36, respectively.

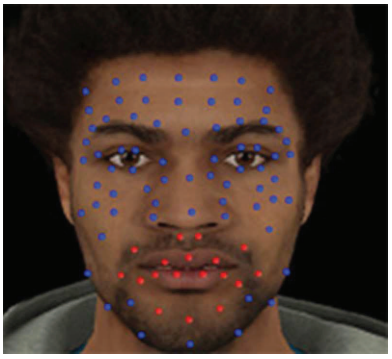

Fig. 5. The used facial marker partition scheme for weight assignment.

When computing the above $\mathbf{Q}_j^i$, we apply a higher weight to the facial markers in the mouth region than those in other facial regions. Its underlying rationale is that when perceiving lip-sync animations, humans tend to put more attention/emphasis on the mouth region than other facial regions. As such, we assign a higher weight, ω_m, to the markers in the mouth region, while assign 1 as the default weight to the other facial markers. In Section 5.3, we will describe how to determine the optimal ω_m via cross-validation. Fig. 5 illustrate how the facial markers are partitioned to two categories (the mouth region and the other) for weight assignment.

Finally, we use a linear mapping to transform the computed RMSE error $\mathbf{Q}_j^i$ to the range of 1 to 5 in order to make it consistent with the five-point Likert scale (employed in the user study described in Section 6), where 1 represents the worst quality and 5 represents the best quality. This linear mapping can be constructed straightforwardly: assuming the largest synthesis error in our data set is ξ, then the mapped value of $\mathbf{Q}_j^i$ is $5 - 4 * (\mathbf{Q}_j^i/\xi)$.

5.2.1 Gaussian Process SAQP Model

Given the above obtained SATF metrics, $\mathbf{E}$ $(= \{\mathbf{E}_1, \ldots, \mathbf{E}_i, \ldots, \mathbf{E}_n\})$, and their corresponding ground-truth synthesis errors, $\mathbf{Q}$ $(= \{\mathbf{Q}_1, \ldots, \mathbf{Q}_i, \ldots, \mathbf{Q}_n\})$, we train a Gaussian Process Regression model to learn the mapping from $\mathbf{E}$ to $\mathbf{Q}$. We choose the GPR model for this learning due to the following reasons: first, the GPR model is nonparametric, so it does not require extensive manual efforts for parameter tuning to achieve good training results. Second, the GPR model is context-dependent; hence, it is capable of automatically and robustly handling SATF metrics with different characteristics.

Mathematically, a GPR model is characterized by its hyperparameter vector θ that includes a characteristic length-scale parameter θ_1 and a signal magnitude parameter θ_2. Training a GPR model is to learn the hyperparameter vector θ. In this work, we learn the hyperparameter vector θ by optimizing the following marginal log-likelihood function (9).

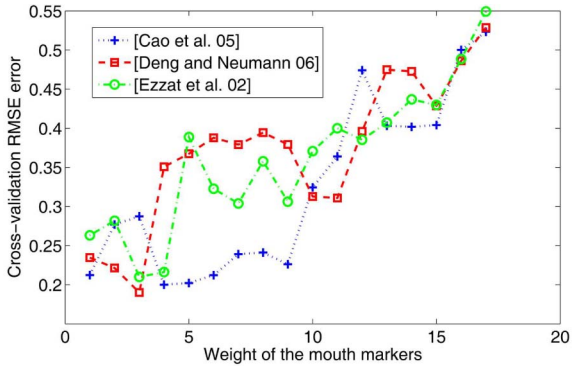

Fig. 6. Plotting of how the cross-validation RMSE is changed when ω_m is increased.

$$L_{GP} = -\log P(\mathbf{Q}|\mathbf{E},\theta) = \frac{1}{2}\log|\mathbf{K}+\sigma^2\mathbf{I}| + \frac{1}{2}\mathbf{Q}^T(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{Q} + \frac{N}{2}\log 2\pi. \quad (9)$$

Here L_{GP} is the negative log-posterior of the model, σ^2 is the variance of noise (0.014 in this work), and $\mathbf{K}$ is a used kernel function. In this work, the used kernel function is an ARD covariance function [55] (refer to (10))

$$\mathbf{K}_{i,j} = k(\mathbf{E}_i,\mathbf{E}_j) = \theta_2 \exp\left(-\frac{1}{2{\theta_1}^2}\|\mathbf{E}_i-\mathbf{E}_j\|^2\right). \quad (10)$$

Then, Rasmussen's minimization algorithm [56] is chosen to optimize L_{GP} due to its efficiency. The maximum number of iterations is experimentally set to 1,024. After θ is optimally solved, the trained GPR model yields a likelihood function for any predicted output. Concretely, for any new SATF metric (i.e., a scalar value), $\mathbf{e}$, we can compute a probabilistic distribution of its predicted synthesis quality error, $\mathbf{q}$. In addition, we can evaluate the negative log probability of the predicted output. This log-likelihood function is shown in (11)

$$L_S = -\log P(\mathbf{q}|\mathbf{e},\theta) = \frac{1}{2}\log(2\pi(V(\mathbf{e})+\sigma^2)) + \frac{\|\mathbf{q}-U(\mathbf{e})\|^2}{2(V(\mathbf{e})+\sigma^2)}, \quad (11)$$

$$U(\mathbf{e}) = \kappa(\mathbf{e})^{\mathbf{T}}(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{Q},$$
$$V(\mathbf{e}) = k(\mathbf{e},\mathbf{e}) - \kappa(\mathbf{e})^{\mathbf{T}}(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\kappa(\mathbf{e})^T.$$

Here $\kappa(\mathbf{e})$ is a vector in which the ith entry is $k(\mathbf{e},\mathbf{E}_i)$, function U returns the mean of the posterior distribution of the learned model given new input $\mathbf{e}$, and function V returns the variance of the learned posterior distribution. In sum, we just need to minimize L_S to obtain the predicted output $\mathbf{q}$.

5.3 SAQP Model Cross-Validation

We validate the accuracy of our SAQP model by applying it to three chosen data-driven speech animation synthesis algorithms [5], [6], [10]. Specifically, in the model training step, each of the three chosen algorithms is used to generate its own set of ground-truth synthesis errors $\{\mathbf{Q}_1,\mathbf{Q}_2,\ldots,\mathbf{Q}_n\}$ ($n = 484$, refer to Section 5.2), though the same SATF metrics

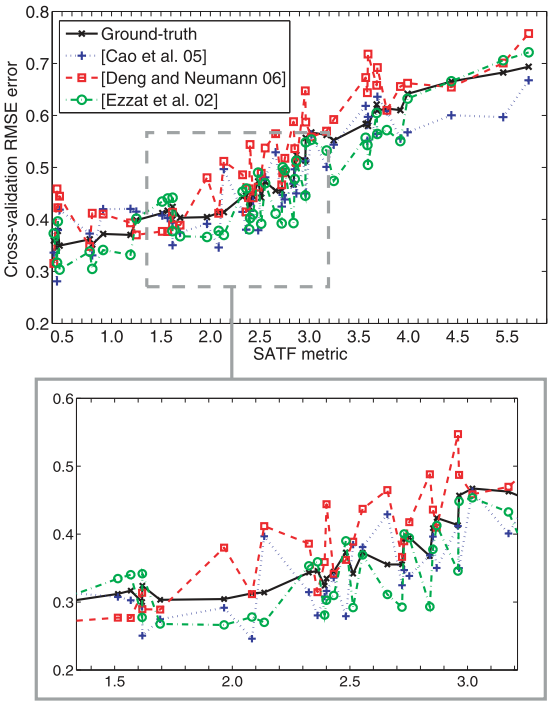

Fig. 7. Plotting of the SATF metrics $\mathbf{E}_v$ versus the cross-validation RMSE errors $\mathbf{Q}_v$ by the three chosen approaches and the ground-truth RMSE errors. The bottom panel is a zoomed version of a selected portion in the top panel.

$\{\mathbf{E}_1,\mathbf{E}_2,\ldots,\mathbf{E}_n\}$ are used. We train three different SAQP models for the three algorithms, respectively.

As mentioned in Section 5.2, the facial motion subset $\mathbf{M}_v$ (total 48 sentences) is specifically retained for cross-validation purpose. Therefore, in the validation step, we first compute the SATF metrics of the 48 validation sentences, $\{\mathbf{E}_{v_1},\mathbf{E}_{v_2},\ldots,\mathbf{E}_{v_{48}}\}$. Meanwhile, each of the three chosen algorithms [5], [6], [10] is used to compute its own set of speech animation synthesis errors (ground-truth) for the 48 validation sentences, denoted as $\{\mathbf{Q}_{v_1},\mathbf{Q}_{v_2},\ldots,\mathbf{Q}_{v_{48}}\}$. Finally, we use the trained GPR model to predict the speech animation synthesis qualities based on the inputted $\mathbf{E}_{v_i}$ ($1 \le i \le 48$), that is, $\hat{\mathbf{Q}}_{v_i} = F_\Psi(\mathbf{E}_{v_i})$.

Our cross-validation procedure consists two steps: the first step determines the optimal ω_m (i.e., the weight for facial markers in the mouth region, refer to (8)) via cross-validation, and the second step performs the cross-validation comparison over the selected test sentences by using the determined optimal ω_m. The reason is that the learned GPR hyperparameters in the SAQP model depends on the chosen value of the ω_m parameter, and thus it is necessary to determine the optimal ω_m in the first step.

Fig. 6 shows how the cross-validation RMSE error is changed when ω_m varies. In this figure, X-axis denotes ω_m, and Y-axis denotes the averaged cross-validation error (total 48 cross-validation sentences). As clearly shown in Fig. 6, the optimal ω_m for both [6] and [10] is 3, and the optimal ω_m for [5] is 4. The constructed SAQP model in the remaining writing uses the optimal values of ω_m.

Fig. 7 plots the SATF metrics $\mathbf{E}_v$ versus the cross-validation RMSE errors $\mathbf{Q}_v$ by the three chosen approaches and the ground-truth RMSE errors of the cross-validation

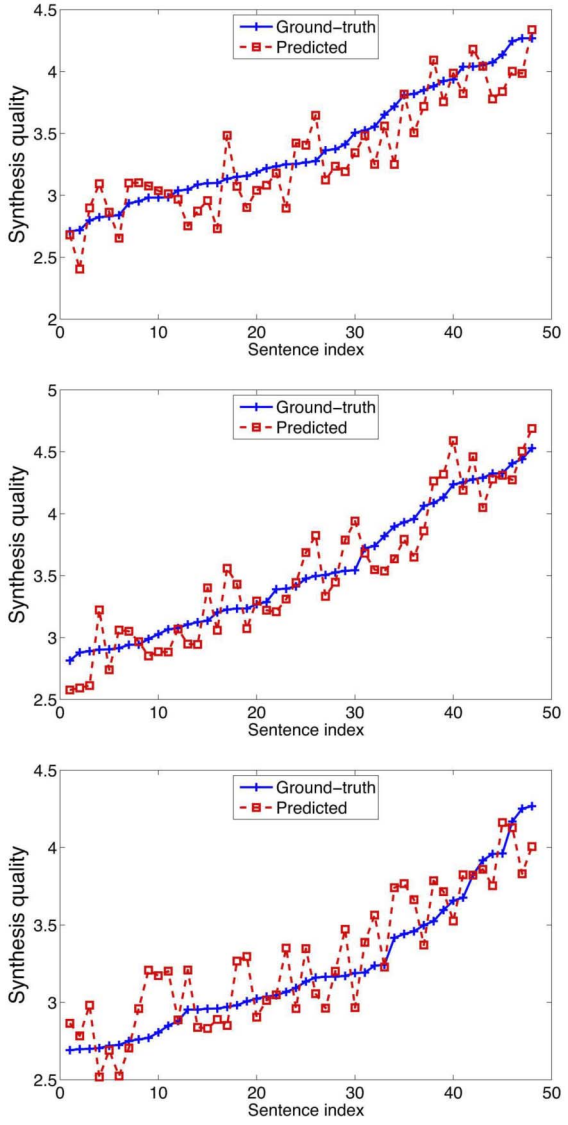

Fig. 8. Cross-validation comparison results of all the 48 sentences in our validation data set: the Anime-Graph-based approach [5] (Top), the eFASE approach [6] (Middle), and 3D extension of the MMM-based approach [10] (Bottom).

data set. From this figure, we can observe that the overall trend of the three approaches is approximately close to the ground-truth, although noticeable fluctuations exist.

Fig. 8 shows cross-validation comparison results ($\mathbf{Q}_{v_i}$ versus $\hat{\mathbf{Q}}_{v_i}$) of the retained 48 validation sentences (the above optimally determined ω_m is used). As shown in this figure, we can observe that the predicted speech animation synthesis qualities of the three chosen approaches [5], [6], [10] are measurably close to the ground-truth, considering the intrinsic difficulty of this problem. Note that in Fig. 8, in order to make the plotting easier to understand, we intentionally rearrange the animation clip indexes by the ground-truth synthesis quality in the ascending order.

6 Results and Evaluations

In this work, we conducted a user study to evaluate the effectiveness of our proposed SAQP model. In the effectiveness user study, we choose three recent representative data-driven speech animation approaches [5], [6], [10] to evaluate its effectiveness (refer to Section 4). In the above Section 5.3, we performed cross-validation on the three approaches. However, it still remains unclear whether the proposed SAQP model can be soundly accurate and robust when arbitrary texts from real-world applications or voices from different subjects (i.e., different from the subjects whose facial motion data are acquired for SAQP model training, refer to Section 3) are inputted. In particular, in many online or interactive applications, it is technically infeasible to acquire the ground-truth speech motions in advance (e.g., via preplanned facial motion acquisition). As such, we focus on quantifying the performance of the SAQP model in these scenarios.

Fig. 9. Four selected frames in one synthetic speech animation clip used in our user study.

6.1 Effectiveness User Study

We first randomly extracted 17 sentences from CNN News, Yahoo News, and Internet speech as the test sentences, and then recorded the voice of a male human subject when he spoke the chosen 17 sentences. Note that this male subject is different from the motion capture female subject in Section 3. After that, we used the Festival system [48] to extract their corresponding phoneme sequences with timing information from the recorded voices. Based on the obtained 17 phoneme sequences, each of the three chosen data-driven speech animation approaches [5], [6], [10] was used to generate 17 speech animation clips (in other words, a total of 51 synthetic speech animation clips with aligned audio) by using the same facial motion data set as detailed in Section 3. In our user study, the resolution of all the synthetic speech animation clips is 640×480. Fig. 9 shows four randomly selected frames in one synthetic clip. To suppress the potential influences of other visual factors on user perception, eye gaze in these clips stays still (looking straight ahead), and there is no head movements in the clips. For animation results, please refer to the enclosed demo video.

We conducted a user study on the above 51 clips. A total of 16 student volunteers participated in the study and they were specifically instructed to only rate the "lip-sync quality" (that is, not other visual factors such as rendering, eye movements, and head movements) of the clips one by one. We used a five-point Likert scale, where 1 represents "extremely poor," and 5 represents "realistic like a real human." In particular, the participants were allowed to assign any real number (not restricted to integer numbers, e.g., he/she can give a 4.3 rating) between 1.0 and 5.0 as their rating. To counter balance the display order of the visual stimuli in the study, the animation clips were displayed in a random order for each participant.

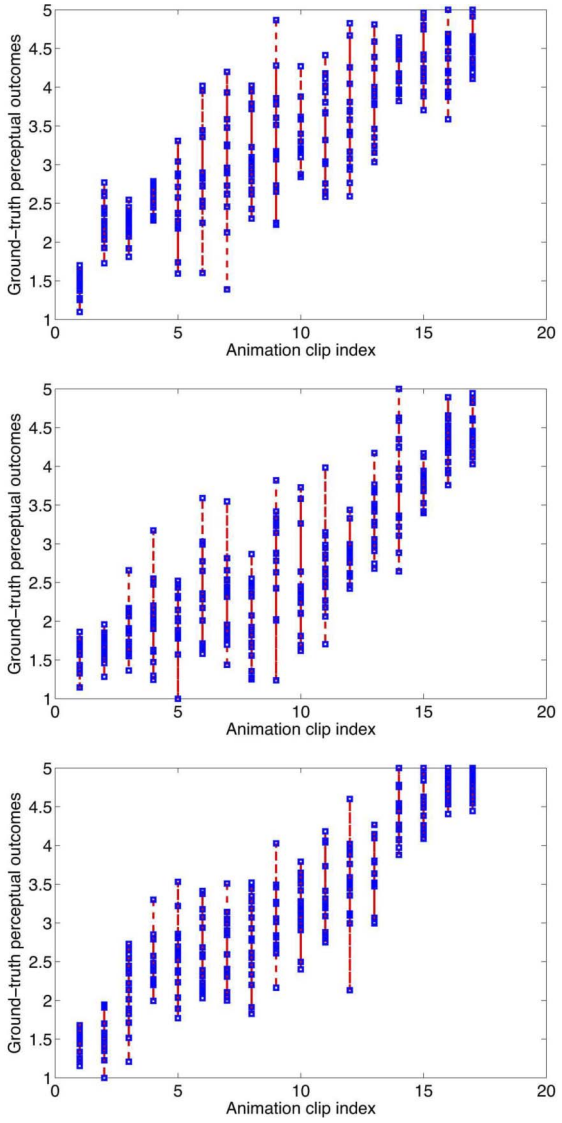

Fig. 10. The obtained user ratings (the ground-truth perceptual outcomes) of the 51 synthetic speech animation clips: the anime-graph approach [5] (Top), the eFASE approach [6] (Middle), and the 3D extension of the MMM-based approach [10] (Bottom). To better visualize the user ratings, we rearrange the animation clip indexes by the averaged ground-truth perceptual outcomes in the ascending order.

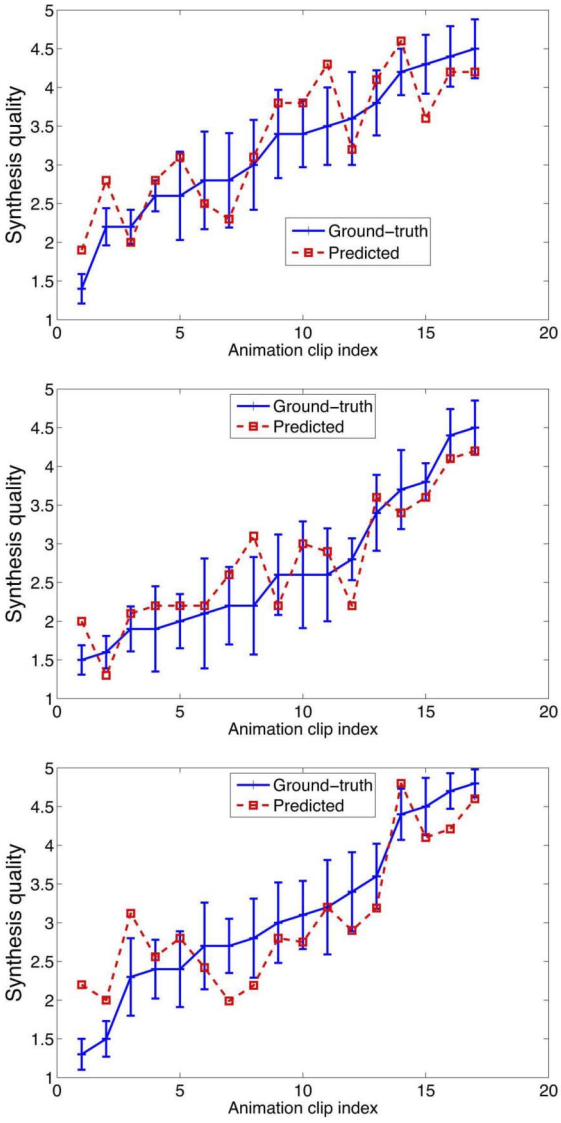

Fig. 11. Comparisons between the ground-truth perceptual outcomes (the averaged user ratings) and the predicted perceptual outcomes by our SAQP model: the anime-graph approach [5] (Top), the eFASE approach [6] (Middle), and the 3D extension of the MMM-based approach [10] (Bottom). To better visualize the results, we rearrange the animation clip indexes by the ground-truth perceptual outcomes in the ascending order. The blue lines in the figure denote the standard deviations.

6.2 Analysis of User Study Results

We analyze the user study results using two different ways. First, we statistically analyze the obtained user ratings to check the rating consistency. Second, we perform comparison analysis on the user-rated (i.e., ground-truth) outcomes and the algorithm outcomes predicted by our SAQP model in order to evaluate its accuracy and robustness.

Statistics of the ground-truth (user-rated) perceptual outcomes. The averaged user ratings of the $51(=17*3)$ synthetic clips are called the *ground-truth perceptual outcomes* in this work. As shown in Fig. 10, we observe that despite some outliers, most of the participants had certain agreements on the perceptual ratings of all the synthetic clips. In particular, such a rating consistency is more obvious at those lowly rated and highly rated clips.

We also computed the standard deviations of the obtained ground-truth perceptual outcomes. Fig. 11 plots the standard deviations as blue error-bars. We can observe that, only 8 (i.e., 23.53 percent) out of the used 51 clips have their standard deviations larger than 0.5, and only 1 out of 51 has its standard deviation larger than 0.7. This indicates majority of the obtained user ratings are consistent, to a certain extent.

Comparison analysis of the user study results. In this writing, the quality outcomes computed/predicted by our SAQP model (Section 5.2) are called the *predicted perceptual outcomes*. Fig. 11 plots the comparison between the ground-truth and the predicted perceptual outcomes. As shown in this figure, we can observe that regardless which of the three algorithms [5], [6], [10] is used, at most cases, the predicted perceptual outcomes computed by the SAQP model are sufficiently close to the ground-truth perceptual outcomes (user ratings).

TABLE 2
Summary of the Quantitative Prediction Errors by the Proposed SAQP Model

	[Cao et al. 05]	[Deng and Neumann 06]	[Ezzat et al. 02]
RMSE	0.44	0.39	0.48
CCA	0.86	0.90	0.87

We also computed the quantitative errors of our predictions. Table 2 shows the computed RMSE errors. As shown in the table, for the three chosen approaches [5], [6], [10], the quantitative errors of our SAQP model are reasonably small: RMSE for [5] is 0.44, RMSE for [6] is 0.39, and RMSE for [10] is 0.48. We also used Canonical Correlation Analysis (CCA) [57] to measure the correlation between the ground-truth perceptual outcomes and the predicted perceptual outcomes. The computed CCA coefficients: $r_1 = 0.86$ for [5], $r_2 = 0.90$ for [6], and $r_3 = 0.87$ for [10], are reasonably close to 1.0 (perfectly linear). This shows there is an approximately linear correlation between the ground-truth perceptual outcomes and the predicted outcomes by our SAQP model.

In sum, through the quantitative analysis of the prediction errors by our SAQP model (i.e., RMSE and CCA), our user study results showed that the proposed SAQP model is able to soundly predict the qualities of synthetic data-driven speech animations, and the algorithm predictions are measurably close to the ground-truth user ratings. We believe, with relatively minor modifications, the proposed SAQP model can be generalized and used as a quantified quality predictor for other existing data-driven speech animation approaches such as the work of [3], [4], [5], [11], [7].

7 Discussion and Conclusions

In this paper, we introduce a novel speech animation quality prediction model that can robustly predict the quality of synthetic speech animations dynamically generated by data-driven approaches. Its core element is a trained statistical regression model that bridges the Speech Animation Fitting Trajectory (SATF) metric with the ground-truth synthesis measure.

To the best of our knowledge, this work is the first reported, automated, quantitative quality predictor for data-driven speech animation approaches. In particular, at runtime it does not need to conduct offline, costly, and tedious user studies. Our user study results showed that the SAQP model is able to soundly predict the synthesis quality of data-driven speech animation approaches and the predictions are reasonably close to the ground-truth user ratings. Moreover, we also believe with straightforward modifications or extensions, our SAQP model can be plausibly generalized and used as a quantified quality predictor for other existing data-driven speech animation approaches [3], [4], [11], [7].

As in general automatically predicting the visual quality of synthetic animations (in particular, facial and character animations) is a challenging problem, our current work has a number of limitations, described as follows:

- First, the accuracy of the current SAQP model still needs to be further improved. For example, as shown in Fig. 11, the predictions by the SAQP model are less accurate when the ground-truth perceptual qualities of synthetic speech animations are low.
- Second, since the used SATF metric utilizes phoneme contexts, a reasonably large training data set would be needed to construct and train a well-behaviored SAQP model. For instance, if the training facial motion data set cannot provide a good coverage of various diphones and triphones, then the constructed SAQP model might not be able to make accurate predictions when the inputted phoneme sequence contains a significant portion of uncovered diphone or triphone contexts.
- Third, in our current work, the RMSE error is used to measure the ground-truth synthesis error (quality) in the SAQP model construction step (refer to Section 5.2). However, we are aware that the RMSE error may not be the ideal metric to represent ground-truth perceptual outcomes of those synthetic speech animation clips. Ideally, extensive subjective user studies need to be conducted to consistently rate the hundreds (about 500 in this work) of visual speech animation clips to obtain the true perceptual outcomes, and if such obtained user ratings (not the computed RMSE errors in the current work) are used to train the proposed SAQP model, we anticipate that the prediction accuracy of the SAQP model could be significantly improved. Nevertheless, in reality, such large-scale user studies are often impractical. Therefore, our current work selects the RMSE error as the economical alternative to the user-rated perceptual outcome.
- Fourth, if our proposed SAQP model is trained based on one subject's facial motion data set, without model retraining, it cannot be directly used as a quality predictor for data-driven synthetic speech animations that are based on another different subject's facial motion data set. The main reasons include: 1) in reality, it is practically difficult to put the identical facial marker layout for different motion capture subjects; 2) even if the facial marker layouts of different motion capture subjects are identical, different subjects typically have distinct facial geometries (that is, the transformed region-based PCA representation based on one subject cannot be directly used to describe another subject) and have idiosyncrasies of mouth movements even when speaking the same utterance (that is, different persons typically have distinct sets of visemes and expressions). Therefore, both the region-based PCA representation and the computed ground-truth synthesis error essentially depend on the idiosyncrasy of the particular subject in the training data set, while both of them are used in the SAQP model training step (Section 5.2). In addition, the optimized SAQP model parameters (e.g., those optimized in Section 5) might be also data set-specific, to a certain extent.

Another potential application of the proposed SAQP model is to on-the-fly compare and evaluate the online

performance of various data-driven speech animation algorithms. For example, assuming a number of different data-driven speech animation algorithms run at the back end, and then based on the SAQP predictions, an online system (e.g., taking arbitrary text or speech input from users and generate a corresponding talking avatar) can automatically pick and display the best animation clip among all the clips generated by those back end algorithms.

Along a similar direction, our proposed SAQP model could also be potentially used to compare different data-driven speech animation approaches in a systematic manner. For example, researchers can input the sentences in some widely used corpora (e.g., the UPenn LDC Corpora [58]) to some existing data-driven speech animation approaches and then quantitatively compare the quality of the synthetic speech animations by those approaches, based on our SAQP model. In this way, those different data-driven approaches can be systematically compared and analyzed.

In the future, we plan to evaluate the SAQP model in real-world online applications such as online news virtual presenters, and thus we will be able to further quantify and improve the current model such as dynamically relearning the statistical model based on online user feedback. Second, as the future work, we are interested in exploring the direction of expanding and generalizing the current model. For example, it could be extended to predict the quality of synthetic speech animations generated not only by data-driven techniques, but also by any other speech animation approaches such as traditional yet well-liked key viseme based or blendshape animation techniques. In addition, eyebrow movement could be an important factor to the realism of synthetic speech animation, and we plan to explore how to effectively incorporate the eyebrow movement as well as emotional visual speech into statistical quality prediction models in the future.

ACKNOWLEDGMENTS

This work is supported in part by NSF IIS-0914965, Texas NHARP 003652-0058-2007, and generous research gifts from Google and Nokia. The authors also would like to thank the RocketBox Libraries for providing the used high-quality 3D avatar. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the agencies.

REFERENCES

[1] B. Theobald, S. Fagel, G. Bailly, and F. Elisei, "Visual Speech Synthesis Challenge," *Proc. Interspeech,* pp. 1875-1878, 2008.

[2] C. Bregler, M. Covell, and M. Slaney, "Video Rewrite: Driving Visual Speech with Audio," *Proc. ACM SIGGRAPH,* pp. 353-360, 1997.

[3] S. Kshirsagar and N.M. Thalmann, "Visyllable Based Speech Animation," *Computer Graphics Forum,* vol. 22, no. 3, pp. 631-639, 2003.

[4] J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise, "Accurate Visible Speech Synthesis Based on Concatenating Variable Length Motion Capture Data," *IEEE Trans. Visualization and Computer Graphics,* vol. 12, no. 2, pp. 266-276, Mar./Apr. 2006.

[5] Y. Cao, W.C. Tien, P. Faloutsos, and F. Pighin, "Expressive Speech-Driven Facial Animation," *ACM Trans. Graphics,* vol. 24, no. 4, pp. 1283-1302, 2005.

[6] Z. Deng and U. Neumann, "eFASE: Expressive Facial Animation Synthesis and Editing with Phoneme-Isomap Controls," *Proc. ACM SIGGRAPH/Eurographics Symp. Computer Animation (SCA '06),* pp. 251-260, 2006.

[7] K. Wampler, D. Sasaki, L. Zhang, and Z. Popović, "Dynamic, Expressive Speech Animation from a Single Mesh," *Proc. ACM SIGGRAPH/Eurographics Symp. Computer Animation (SCA '07),* pp. 53-62, 2007.

[8] Z. Deng and U. Neumann, "Expressive Speech Animation Synthesis with Phoneme-Level Control," *Computer Graphics Forum,* vol. 27, no. 8, pp. 2096-2113, 2008.

[9] M. Brand, "Voice Puppetry," *Proc. ACM SIGGRAPH,* pp. 21-28, 1999.

[10] T. Ezzat, G. Geiger, and T. Poggio, "Trainable Videorealistic Speech Animation," *Proc. ACM SIGGRAPH,* pp. 388-398, 2002.

[11] I.-J. Kim and H.-S. Ko, "3D Lip-Synch Generation with Data-Faithful Machine Learning," *Computer Graphics Forum,* vol. 26, no. 3, pp. 295-301, 2007.

[12] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D.H. Salesin, "Synthesizing Realistic Facial Expressions from Photographs," *Proc. ACM SIGGRAPH,* vol. 32, pp. 75-84, 1998.

[13] V. Blanz and T. Vetter, "A Morphable Model for the Synthesis of 3D Faces," *Proc. ACM SIGGRAPH,* pp. 187-194, 1999.

[14] Y. Lee, D. Terzopoulos, and K. Waters, "Realistic Modeling for Facial Animation," *Proc. ACM SIGGRAPH,* pp. 55-62, 1995.

[15] T. Weyrich, W. Matusik, H. Pfister, B. Bickel, C. Donner, C. Tu, J. McAndless, J. Lee, A. Ngan, H.W. Jensen, and M. Gross, "Analysis of Human Faces Using a Measurement-Based Skin Reflectance Model," *ACM Trans. Graphics,* vol. 25, no. 3, pp. 1013-1024, 2006.

[16] W.-C. Ma, A. Jones, J.-Y. Chiang, T. Hawkins, S. Frederiksen, P. Peers, M. Vukovic, M. Ouhyoung, and P. Debevec, "Facial Performance Synthesis Using Deformation-Driven Polynomial Displacement Maps," *ACM Trans. Graphics,* vol. 27, no. 5, pp. 1-10, 2008.

[17] T. Weise, H. Li, L. Van Gool, and M. Pauly, "Face/Off: Live Facial Puppetry," *Proc. ACM SIGGRAPH/Eurographics Symp. Computer Animation (SCA '09),* pp. 7-16, 2009.

[18] K. Singh and E. Fiume, "Wires: A Geometric Deformation Technique," *Proc. ACM SIGGRAPH,* pp. 405-414, 1998.

[19] L. Zhang, N. Snavely, B. Curless, and S.M. Seitz, "Spacetime Faces: High-Resolution Capture for Modeling and Animation," *ACM Trans. Graphics,* vol. 23, no. 3, pp. 548-558, 2004.

[20] E. Sifakis, I. Neverov, and R. Fedkiw, "Automatic Determination of Facial Muscle Activations from Sparse Motion Capture Marker Data," *ACM Trans. Graphics,* vol. 24, no. 3, pp. 417-425, 2005.

[21] B. Bickel, M. Botsch, R. Angst, W. Matusik, M. Otaduy, H. Pfister, and M. Gross, "Multi-Scale Capture of Facial Geometry and Motion," *ACM Trans. Graphics,* vol. 26, no. 3, p. 33, 2007.

[22] W.-W. Feng, B.-U. Kim, and Y. Yu, "Real-Time Data Driven Deformation Using Kernel Canonical Correlation Analysis," *Proc. ACM SIGGRAPH,* pp. 91:1-91:9, 2008.

[23] L. Williams, "Performance-Driven Facial Animation," *Proc. ACM SIGGRAPH,* pp. 235-242, 1990.

[24] J.-Y. Noh and U. Neumann, "Expression Cloning," *Proc. ACM SIGGRAPH,* pp. 277-288, 2001.

[25] R.W. Sumner and J. Popović, "Deformation Transfer for Triangle Meshes," *ACM Trans. Graphics,* vol. 23, no. 3, pp. 399-405, 2004.

[26] X. Ma, B.H. Le, and Z. Deng, "Style Learning and Transferring for Facial Animation Editing," *Proc. ACM SIGGRAPH/Eurographics Symp. Computer Animation (SCA),* pp. 114-123, Aug. 2009.

[27] H. Li, T. Weise, and M. Pauly, "Example-Based Facial Rigging," *Proc. ACM SIGGRAPH,* pp. 32:1-32:6, 2010.

[28] Z. Deng and J. Noh, "Computer Facial Animation: A Survey," *Data-Driven 3D Facial Animation.* Spring-Verlag Press, 2007.

[29] J.P. Lewis, "Automated Lip-Sync: Background and Techniques," *J. Visualization and Computer Animation,* vol. 2, pp. 118-122, 1991.

[30] M. Cohen and D. Massaro, "Modeling Co-Articulation in Synthetic Visual Speech," *Model and Technique in Computer Animation,* pp. 139-156, 1993.

[31] S.A. King and R.E. Parent, "Creating Speech-Synchronized Animation," *IEEE Trans. Visualization and Computer Graphics,* vol. 11, no. 3, pp. 341-352, May/June 2005.

[32] C. Pelachaud, "Communication and Coarticulation in Facial Animation," PhD thesis, Univ. of Pennsylvania, 1991.

[33] S. Daly, "The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity," *Proc. SPIE,* vol. 1666, pp. 179-206, 1993.

[34] M. Ramasubramanian, S.N. Pattanaik, and D.P. Greenberg, "A Perceptually Based Physical Error Metric for Realistic Image Synthesis," *Proc. ACM SIGGRAPH,* pp. 73-82, 1999.
[35] H. Yee, S. Pattanaik, and D.P. Greenberg, "Spatiotemporal Sensitivity and Visual Attention for Efficient Rendering of Dynamic Environments," *ACM Trans. Graphics,* vol. 20, no. 1, pp. 39-65, 2001.
[36] K. Myszkowski, T. Tawara, H. Akamine, and H.-P. Seidel, "Perception-Guided Global Illumination Solution for Animation Rendering," *Proc. ACM SIGGRAPH,* pp. 221-230, 2001.
[37] J.K. Hodgins, J.F. O'Brien, and J. Tumblin, "Perception of Human Motion with Different Geometric Models," *IEEE Trans. Visualization and Computer Graphics,* vol. 4, no. 4, pp. 307-316, Oct-Dec. 1998.
[38] C. O'Sullivan, J. Dingliana, T. Giang, and M.K. Kaiser, "Evaluating the Visual Fidelity of Physically Based Animations," *ACM Trans. Graphics,* vol. 22, no. 3, pp. 527-536, 2003.
[39] C. O'Sullivan, S. Howlett, Y. Morvan, R. McDonnell, and K. O'Conor, "Perceptually Adaptive Graphics," *Proc. Eurographics State-of-the-Art Report (STAR),* pp. 141-164, 2004.
[40] C. Wallraven, J. Fischer, D.W. Cunningham, D. Bartz, and H.H. Bülthoff, "The Evaluation of Stylized Facial Expressions," *Proc. Third Symp. Applied Perception in Graphics and Visualization (APGV '06),* pp. 85-92, 2006.
[41] G. Geiger, T. Ezzat, and T. Poggio, "Perceptual Evaluation of Video-Realistic Speech," *MIT-AI-Memo 2003-003,* Feb. 2003.
[42] D. Cosker, S. Paddock, D. Marshall, P.L. Rosin, and S. Rushton, "Toward Perceptually Realistic Talking Heads: Models, Methods, and McGurk," *ACM Trans. Applied Perception,* vol. 2, no. 3, pp. 270-285, 2005.
[43] A. Schwaninger, S. Schumacher, H. Bülthoff, and C. Wallraven, "Using 3D Computer Graphics for Perception: The Role of Local and Global Information in Face Processing," *Proc. Fourth Symp. Applied Perception in Graphics and Visualization (APGV '07),* pp. 19-26, 2007.
[44] C. Wallraven, M. Breidt, D.W. Cunningham, and H.H. Bulthoff, "Evaluating the Perceptual Realism of Animated Facial Expressions," *ACM Trans. Applied Perception,* vol. 4, no. 4, pp. 1-20, Jan. 2008.
[45] X. Ma, B.H. Le, and Z. Deng, "Perceptual Analysis of Talking Avatar Head Movements: A Quantitative Perspective," *Proc. ACM SIGCHI Int'l Conf. Human Factors in Computing Systems (CHI),* pp. 2699-2702, May 2011.
[46] Z. Deng and X. Ma, "Perceptually Guided Expressive Facial Animation," *Proc. ACM SIGGRAPHEG Symp. Computer Animation (SCA '08),* pp. 67-76, July 2008.
[47] H.P. Graf, E. Cosatto, and T. Ezzat, "Face Analysis for the Synthesis of Photo-Realistic Talking Heads," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 189-195, 2000.
[48] Festival, "The Festival Speech Synthesis System," http://www.cstr.ed.ac.uk/projects/festival/, 2004.
[49] P. Joshi, W.C. Tien, M. Desbrun, and F. Pighin, "Learning Controls for Blend Shape Based Realistic Facial Animation," *Proc. ACM SIGGRAPHEG Symp. Computer Animation (SCA '03),* pp. 187-192, 2003.
[50] Q. Li and Z. Deng, "Orthogonal Blendshape Based Editing System for Facial Motion Capture Data," *IEEE Computer Graphics and Applications,* vol. 28, no. 6, pp. 76-82, Nov./Dec. 2008.
[51] F.D. la Torre and M.J. Black, "Robust Principal Component Analysis for Computer Vision," *Proc. IEEE Int'l Conf. Computer Vision,* vol. 1, pp. 362-369, 2001.
[52] F. Girosi, M. Jones, and T. Poggio, "Priors Stabilizers and Basis Functions: From Regularization to Radial, Tensor and Additive Splines," technical report, 1993.
[53] G. Wahba, "Mathematics of Computation," *Spline Models for Observational Data,* vol. 57, SIAM, 1991.
[54] M. Botsch and O. Sorkine, "On Linear Variational Surface Deformation Methods," *IEEE Trans. Visualization and Computer Graphics,* vol. 14, no. 1, pp. 213-230, Jan./Feb. 2008.
[55] C.K.I. Williams and C.E. Rasmussen, "Gaussian Processes for Regression," *Advances in Neural Information Processing Systems 8,* pp. 514-520, MIT press, 1996.
[56] C.E. Rasmussen, "Minimize Function," http://www.kyb.tuebingen.mpg.de/bs/people/carl/, 2006.
[57] K.V. Mardia, J.T. Kent, and J.M. Bibby, *Multivariate Analysis.* Academic Press, 1979.
[58] M.P. Marcus, M.A. Marcinkiewicz, and B. Santorini, "Building a Large Annotated Corpus of English: The Penn Treebank," *Computational Linguistics,* vol. 19, pp. 313-330, June 1993.

Xiaohan Ma received the BS and MS degrees in computer science from Zhejiang University, China, in 2005 and 2007, respectively. Currently, he is working toward the PhD degree at the Department of Computer Science at the University of Houston (UH) under the supervision of Prof. Zhigang Deng. His research interests include computer graphics, computer animation, HCI, and GPU computing. He is a student member of the IEEE.

Zhigang Deng received the BS degree in mathematics from Xiamen University, China, and the MS degree in computer science from Peking University, China, and the PhD degree in computer science at the Department of Computer Science at the University of Southern California, Los Angeles, in 2006. Currently, he is working as an assistant professor of computer science at the University of Houston (UH). His research interests include computer graphics, computer animation, virtual human modeling and animation, human computer interaction, and visual-haptic interfacing. He is a senior member of the IEEE and the IEEE Computer Society and a member of the ACM and ACM SIGGRAPH.