Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis

Carlos Busso, *Student Member, IEEE*, Zhigang Deng, *Member, IEEE*, Michael Grimm, *Student Member, IEEE*, Ulrich Neumann, and Shrikanth Narayanan, *Senior Member, IEEE*

***Abstract*—Rigid head motion is a gesture that conveys important nonverbal information in human communication, and hence it needs to be appropriately modeled and included in realistic facial animations to effectively mimic human behaviors. In this paper, head motion sequences in expressive facial animations are analyzed in terms of their naturalness and emotional salience in perception. Statistical measures are derived from an audiovisual database, comprising synchronized facial gestures and speech, which revealed characteristic patterns in emotional head motion sequences. Head motion patterns with neutral speech significantly differ from head motion patterns with emotional speech in motion activation, range, and velocity. The results show that head motion provides discriminating information about emotional categories. An approach to synthesize emotional head motion sequences driven by prosodic features is presented, expanding upon our previous framework on head motion synthesis. This method naturally models the specific temporal dynamics of emotional head motion sequences by building hidden Markov models for each emotional category (sadness, happiness, anger, and neutral state). Human raters were asked to assess the naturalness and the emotional content of the facial animations. On average, the synthesized head motion sequences were perceived even more natural than the original head motion sequences. The results also show that head motion modifies the emotional perception of the facial animation especially in the valence and activation domain. These results suggest that appropriate head motion not only significantly improves the naturalness of the animation but can also be used to enhance the emotional content of the animation to effectively engage the users.**

***Index Terms*—Emotion, head motion, hidden Markov models (HMMs), prosody, talking avatars driven by speech.**

I. Introduction

In normal human–human interaction, gestures and speech are intricately coordinated to express and emphasize ideas and to provide suitable feedback. The tone and the intensity of speech, facial expressions, rigid head motion and hand movements are combined in a nontrivial manner, as they unfold in natural human communication. These interrelations need to be considered in the design of realistic human animation to effectively engage the users.

One important component of our body language that has received little attention compared to other nonverbal gestures is rigid head motion. Head motion is important not only to acknowledge active listening or replace verbal information (e.g., "nod"), but also for many interesting aspects of human communication. Munhall *et al.* showed that head motion improves the acoustic perception of the speech [1]. They also suggested that head motion helps to distinguish between interrogative and declarative statements. Hill and Johnston found that head motion also helps to recognize speaker identity [2]. Graf *et al.* proved that the timings of head motion and the prosodic structure of the text are consistent [3], suggesting that head motion is useful to segment the spoken content. In addition to that, we hypothesize that head motion provides useful information about the mood of the speaker, as suggested by [3]. We believe that people use specific head motion patterns to emphasize their affective states.

Given the importance of head motion in human communication, this aspect of nonverbal gestures should be properly included in an engaging talking avatar. The manner in which people move their head depends on several factors such as speaker styles and idiosyncrasies [2]. However, the production of speech seems to play a crucial role in the production of rigid head motion. Kuratate *et al.* [4] presented preliminary results about the close relation between head motion and acoustic prosody. They concluded, based on the strong correlation between these two streams of data ($r = 0.8$), that the production systems of the speech and head motion are internally linked. These results suggest that head motion can be estimated from prosodic features.

In our previous work, we presented a synthesis framework for rigid head motion sequences driven by prosodic features [5]. We modeled the problem as classification of discrete representations of head poses, instead of estimating mapping functions between the head motion and prosodic features, as in [3], [6].

Manuscript received January 22, 2006; revised June 29, 2006. This work was supported in part by funds from the National Science Foundation (NSF) (through the Integrated Media Systems Center, an NSF Engineering Research Center, Cooperative Agreement No. EEC-9529152 and a CAREER award), the Department of the Army, and a MURI award from the Office of Naval Research. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies. This work was performed when the authors were with Integrated Media Systems Center, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael Davies.

C. Busso, U. Neumann, and S. Narayanan are with the Integrated Media Systems Center, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: busso@usc.edu; shri@sipi.usc.edu).

Z. Deng was with the Integrated Media Systems Center, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089 USA. He is now with the Department of Computer Science, University of Houston, Houston, TX 77004 USA.

M. Grimm was with the Integrated Media Systems Center, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089 USA. He is now with the Institut für Nachrichtentechnik (INT), Universität Karlsruhe (TH), 12–6131 Karlsruhe, Germany.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2006.885910

hidden Markov models (HMMs) were used to learn the temporal relation between the dynamics of head motion sequences and the prosodic features. The HMMs were used to generate quantized head motion sequences, which were smoothed using first-order Markov models (bi-gram) and spherical cubic interpolation. Notice that prosodic features predominantly describe the source of speech rather than the vocal tract. Therefore, this head motion synthesis system is independent of the specific lexical content of what is spoken, reducing the size of the database needed to train the models. In addition to that, prosodic features contain important clues about the affective state of the speakers. Consequently, the proposed model can be naturally extended to include emotional content of the head motion sequence, by building HMMs appropriate for each emotion, instead of generic models.

In this paper, we address three fundamental questions. 1) How important is rigid head motion for natural facial animation? 2) Do head motions change our emotional perception? 3) Can emotional and natural head motion be synthesized only by prosodic features? To answer these questions, the temporal behavior of head motion sequences extracted from our audiovisual database were analyzed for three different emotions: happiness, anger, sadness, and the neutral state. The results show that the dynamic of head motion with neutral speech significantly differs from the dynamics of head motion with emotional speech. These results suggest that emotional models need to be included to synthesize head motion sequences that effectively reflect these characteristics. Following this direction, an extension of the head motion synthesis method, originally proposed in [5], is presented. The approach described in the present paper includes emotional models that learn the temporal dynamics of the real emotional head motion sequences. To investigate whether rigid head motion affects our perception of the emotion, we synthesized facial animation with deliberate mismatches between the emotional speech and the emotional head motion sequence. Human raters were asked to assess the emotional content and the naturalness of the animations. In addition, animations without head motion were also included in the evaluation. Our results indicate that head motion significantly improves the naturalness perception in the facial animation. They also show that head motion changes the emotional content perceived from the animation, especially in the valence and activation domain. Therefore, head motion can be appropriately and advantageously included in the facial animation to emphasize the emotional content of the talking avatars.

The paper is organized as follows: Section II motivates the use of audiovisual information to synthesize expressive facial animations. Section III describes the audiovisual database, the head pose representation and the acoustic features used in the paper. Section IV presents statistical measures of head motion displayed during expressive speech. Section V describes the multimodal framework, based on HMMs, to synthesize realistic head motion sequences. Section VI summarizes the facial animation techniques used to generate the expressive talking avatars. Section VII presents and discusses the subjective evaluations employed to measure the emotional and naturalness perception under different expressive head motion sequences. Finally, Section VIII gives the concluding remarks and our future research direction.

II. Emotion Analysis

For engaging talking avatars, special attention needs to be given to include emotional capability in the virtual characters. Importantly, Picard has underscored that emotions play a crucial rule in rational decision making, in perception, and in human interaction [7]. Therefore, applications such as virtual teachers, animated films, and new human–machine interfaces can be significantly improved by designing control mechanisms to animate the character to properly convey the desired emotion. Human beings are especially good at not only inferring the affective state of other people, even if emotional clues are subtly expressed, but also in recognizing nongenuine gestures, which challenges the designs of these control systems.

The production mechanisms of gestures and speech are internally linked in our brain. Cassell *et al.* mentioned that they are not only strongly connected, but also systematically synchronized in different scales (phonemes–words–phrases–sentences) [8]. They suggested that hands gestures, facial expressions, head motion, and eye gaze occur at the same time as speech, and they convey similar information as the acoustic signal. Similar observations were mentioned by Kettebekov *et al.* [9]. They studied *deictic* hand gestures (e.g., pointing) and the prosodics of the speech in the context of gesture recognition. They concluded that there is a multimodal coarticulation of gestures and speech, which are loosely coupled.

From an emotional expression point of view, in communication, it has been observed that human beings jointly modify gestures and speech to express emotions. Therefore, a more complete human–computer interaction system should include details of the emotional modulation of gestures and speech.

In sum, all these findings suggest that the control system to animate virtual human-like characters needs to be closely related and synchronized with the information provided by the acoustic signal. This is especially important if a believable talking avatar conveying specific emotion is desired. Following this direction, Cassell *et al.* proposed a rule-based system to generate facial expressions, hand gestures, and spoken intonation, which were properly synchronized according to rules [8]. Other talking avatars that take into consideration the relation between speech and gestures to control the animation were presented in [10]–[13].

Given that head motion also presents similar close temporal relation with speech [3], [4], [14], this paper proposes to use HMMs to jointly model these streams of data. As shown in our previous work [5], HMMs provide a suitable framework to capture the temporal relation between speech and head motion sequences.

III. AudioVisual Database

The audiovisual database used in this research was collected from an actress with 102 markers attached to her face (left of Fig. 1). She was asked to repeat a custom-made, phoneme-balance corpus four times, expressing different emotions, respectively (neutral state, sadness, happiness, and anger). A VICON motion capture system with three cameras (middle of Fig. 1) was used to track the 3-D position of each marker. The sampling rate was set to 120 frames per second. The acoustic signal was

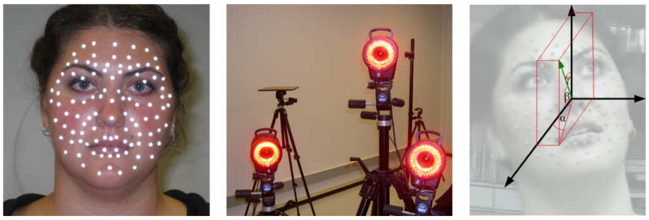


Fig. 1. Audiovisual database collection. The left figure shows the facial marker layout, the middle figure shows the facial motion capture system, and the right figure shows the head motion features extraction.

simultaneously recorded by the system, using a close talking SHURE microphone working at 48 kHz. In total, 640 sentences were used in this work. The actress did not receive any instruction about how to move her head.

After the data were collected, the 3-D Euler angles, which were used to represent the rigid head poses, were computed. First, all the markers' positions were translated to make the nose marker the center of the coordinate system. Then, a neutral head pose was selected as the reference frame, M_{ref} (102 × 3 matrix). For each frame, a matrix M_t was created, using the same marker order as the reference. Following that, the singular value decomposition (SVD) UDV^T of the matrix $M_t^T \cdot M_{\text{ref}}$ was calculated. The product VU^T gives the rotational matrix, R_t, used to spatially align the reference and the frame head poses [15]. Finally, the 3-D Euler angles x_t were computed from this matrix (right of Fig. 1)

$$M_t^T \cdot M_{\text{ref}} = UDV^T \tag{1}$$

$$R_t = VU^T. \tag{2}$$

In previous work, head motion has been modeled with six *degrees of freedom* (DOF), corresponding to head rotation (3 DOF) and translation (3 DOF) [14], [16]. However, for practical reasons, in this paper we consider only head rotation. As discussed in Section V, the space spanned by the head motion features is split using vector quantization. For a constant quantization error, the number of clusters needed to span the head motion space increases as the dimension of the feature vector increases. Since an HMM is built for each head pose cluster, it is preferred to model head motion with only a 3-D feature vector, thereby decreasing the number of HMMs. Furthermore, since most of the avatar applications require close-view of the face, translation effects are considerably less important than the effects of head rotation. Thus, the 3 DOF of head translation are not considered here, reducing the number of required HMM models and the expected quantization errors.

The acoustic prosodic features were extracted with the Praat speech processing software [17]. The analysis window was set to 25 ms with an overlap of 8.3 ms, producing 60 frames per second. The pitch (F0) and the rms energy and their first and second derivatives were used as prosodic features. The pitch was smoothed to remove any spurious spikes, and interpolated to avoid zeros in the unvoiced regions of the speech, by using the corresponding options provided by the Praat software [17].

TABLE I
STATISTICS OF RIGID HEAD MOTION

	Neu	Sad	Hap	Ang
	Canonical correlation Analysis			
	0.74	0.74	0.71	0.69
	Motion Coefficient [°]			
α	3.32	4.76	6.41	5.56
β	0.88	3.23	2.60	3.67
γ	0.81	2.20	2.32	2.69
	Range [°]			
α	9.54	13.71	17.74	16.05
β	2.31	8.29	6.14	9.06
γ	2.27	6.52	6.67	8.21
	Velocity Magnitude [°/sample]			
Mean	.08	0.11	0.15	0.18
Std	.07	0.10	0.13	0.15
	Discriminant Analysis			
Neu	0.92	0.02	0.04	0.02
Sad	0.15	0.61	0.11	0.13
Hap	0.14	0.09	0.59	0.18
Ang	0.14	0.11	0.25	0.50

IV. HEAD MOTION CHARACTERISTICS IN EXPRESSIVE SPEECH

To investigate head motion in expressive speech, the audiovisual data were separated according to the four emotions. Different statistical measurements were computed to quantify the patterns in rigid head motion during expressive utterances.

Canonical correlation analysis (CCA) was applied to the audiovisual data to validate the close relation between the rigid head motions and the acoustic prosodic features. CCA provides a scale-invariant optimal linear framework to measure the correlation between two streams of data with equal or different dimensionality. The basic idea is to project the features into a common space in which Pearson's correlation can be computed. The first part of Table I shows these results. One-way *analysis of variance* (ANOVA) evaluation indicates that there are significant differences between emotional categories ($F[3, 640]$, $p = 0.00013$). Multiple comparison tests also show that the CCA average of neutral head motion sequences is different from the CCA mean of sad ($p = 0.001$) and angry ($p = 0.001$) head motion sequences. Since the average of the first-order canonical correlation in each emotion is over $r = 0.69$, it can be inferred that head motion and speech prosody are strongly linked. Consequently, meaningful information can be extracted from prosodic features to synthesize the rigid head motion.

To measure the motion activity of head motion in each of the three Euler angles, we estimated a *motion coefficient* Ψ which is defined as the standard deviation of the sentence-level mean-removed signal

$$\Psi = \sqrt{\frac{1}{N \cdot T} \sum_{u=1}^{N} \sum_{t=1}^{T} \left(x_t^u - \overline{\mu^u}\right)^2} \tag{3}$$

where T is the number of frames, N is the number of utterances, and $\overline{\mu^u}$ is the mean of the sentence u. The results shown in Table I suggest that the head motion activity displayed when the speaker is under emotional states (sadness, happiness, or anger) is much higher than the activity displayed under neutral speech. Furthermore, it can be observed that head motion activity related to sad emotion is slightly lower than the activity for happy or angry. As an aside, it is interesting to note that similar trends with respect to emotional state have been observed in articulatory data of tongue and jaw movement [18].

Table I also shows the average ranges of the three Euler angles that define the head poses. The results indicate that during emotional utterances the head is moved over a wider range than in normal speech, which is consistent with the results of the motion coefficient analysis.

The velocity of head motion was also computed. The average and the standard deviation of the head motion velocity magnitude is presented in Table I. The results indicate that the head motion velocities for happy and angry sequences are about two times greater than that of neutral sequences. The velocities of sad head motion sequences are also greater than that of neutral head motion, but smaller than that of happy and angry sequences. In terms of variability, the standard deviation results reveal a similar trend. These results suggest that emotional head motion sequences present different temporal behavior than those of neutral condition.

To analyze how distinct the patterns of rigid head motion for emotional sentences are, a discriminant analysis was applied to the data. The mean, standard deviation, range, maximum, and minimum of the Euler angles computed at the sentence-level were used as features. Fisher classification was implemented with leave-one-out cross validation method. Table I shows the results. On average, the recognition rate just with head motion features was 65.5%. Notice that the emotional class with lower performance (anger) is correctly classified with an accuracy higher than 50% (chance is 25%). These results suggest that there are distinguishable emotional characteristics in rigid head motion. Also, the high recognition rate of neutral state implies that global patterns of head motion in normal speech are completely different from the patterns displayed under an emotional state.

These results suggest that people intentionally use head motion to express specific emotion patterns. Therefore, to synthesize expressive head motion sequences, suitable models for each emotion need to be built.

V. Rigid Head Motion Synthesis

The framework used in this work to synthesize realistic head motion sequences builds upon the approach presented in our previous publication [5]. This section presents the extension of this method.

The proposed speech-driven head motion sequence generator uses HMMs because they provide a suitable framework to jointly model the temporal relation between prosodic features and head motion. Instead of estimating a mapping function [3], [6], or designing rules according to the lexical content of the speech [8], or finding similar samples in the training data [16], we model the problem as classification of discrete representations of head poses which are obtained by the use of vector quantization. The *Linde–Buzo–Gray vector quantization* (LBG-VQ) technique [19] is used to compute K Voronoi cells in the 3-D Euler angle space. The clusters are represented with their mean vector U_i and covariance matrix Σ_i, with $i = 1, \ldots, K$. For each of these clusters V_i, an HMM is built to generate the most likely head motion sequence, given the observations O, which correspond to the prosodic features. The number of HMMs that need to be trained is given by the number of clusters (K) used to represent the head poses.

Two smoothing techniques are used to produce continuous head pose sequences. The first smoothing technique is imposed in the decoding step of the HMMs, by constraining the transition between clusters. The second smoothing technique is applied during synthesis, by using spherical cubic interpolation to avoid breaking of the discrete representation. More details of these smoothing techniques are given in Sections V-A and B, respectively.

In our previous work, we proposed the use of generic (i.e., emotion-independent) models to generate head motion sequences [5]. As shown in Section IV, the dynamics and the patterns of head motion sequences under emotional states are significantly different. Therefore, these generic models do not reflect the specific emotional behaviors. In this paper, the technique is extended to include emotion-dependent HMMs. Instead of using generic models for the whole data, we proposed building HMMs for each emotional category to incorporate in the models the emotional patterns of rigid head motion.

A. Learning Relations Between Prosodic Features and Head Motion

To synthesize realistic head motion, our approach searches for the sequences of discrete head poses that maximize the posterior probability of the cluster models $V = (V_{i_1}^{t}, V_{i_2}^{t+1}, \ldots,)$, given the observations $O = (o^t, o^{t+1}, \ldots,)$

$$\arg \max_{i_1, i_2, \ldots} P\left(V_{i_1}^{t}, V_{i_2}^{t+1}, \ldots, |O\right). \tag{4}$$

This posterior probability is computed according to Bayes rule as

$$P(V|O) = \frac{P(O|V) \cdot P(V)}{P(O)}. \tag{5}$$

$P(O)$ is the probability of the observation which does not depend on the cluster models. Therefore, it can be considered as a constant. $P(O|V)$ corresponds to the likelihood distribution of the observation, given the cluster models. This probability is modeled as a first-order Markov process, with S states. Hence, the probability description includes only the current and previous state, which significantly simplifies the problem. For each of the S states, a mixture of M Gaussians is used to estimate the distribution of the observations. The use of mixtures of Gaussians models the many-to-many mapping of head motion and prosodic features. Under this formulation, the estimation of the likelihood is reduced to computing the parameters of the HMMs, which can be estimated using standard methods such as forward-backward and Baum–Welch reestimation algorithms [20], [21].

$P(V)$ in (5) corresponds to the prior probability of the cluster models. This probability is used as a first smoothing technique to guarantee valid transition between the discrete head poses. A first-order state machine is built to learn the transition probabilities of the clusters, by using bi-gram models (similar to bi-gram language models [20]).The transition between clusters are learned from the training data. In the decoding step of the HMMs, these bi-gram models are used to penalize or reward

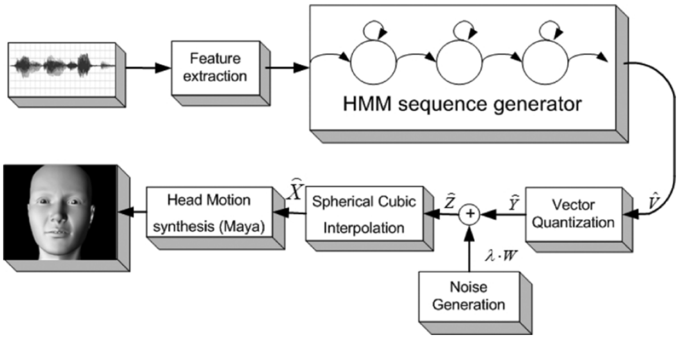


Fig. 2. Head motion synthesis framework.

transitions between discrete head poses according to their occurrences in the training database. As our results suggest, the transition between clusters is also emotion-dependent. Therefore, this prior probability is separately trained for each emotion category.

Notice that in the training procedure the segmentation of the acoustic signal is obtained from the vector quantization step. Therefore, the HMMs were initialized with this known segmentation, avoiding the use of forced alignment, as it is usually done in speech recognition to align phonemes with the speech features.

B. Generating Realistic Head Motion Sequences

Fig. 2 describes the proposed framework to synthesize head motion sequences. Using the acoustic prosodic features as input, the HMMs, which were previously trained as described in Section V-A, generate the most likely head pose sequences $\widehat{V} = (\widehat{V}_{i_1}^{t}, \widehat{V}_{i_2}^{t+1}, \ldots,)$, according to (4). After the sequence $\widehat{V}$ is obtained, the means of the clusters are used to form a 3-D sequence, $\widehat{Y} = (U_{i_1}^{t}, U_{i_2}^{t+1} \ldots)$, which is the first approximation of the head motion.

In the next step, colored noise is added to the sequence $\widehat{Y}$, according to (6) (see Fig. 2). The purpose of this step is to compensate for the quantization error of the discrete representation of head poses. The noise is colored with the covariance matrix of the clusters Σ so as to distribute the noise in proportion to the error yielded during vector quantization. The parameter λ is included in (6) to attenuate, if desired, the level of noise used to blur the sequence $\widehat{Y}$ (e.g., $\lambda = 0.7$). Notice that this is an optional step that can be ignored by setting λ equal to zero. Fig. 3 shows an example of $\widehat{Z}$ (blue solid lines)

$$\widehat{Z}_i^t = \widehat{Y}_i^t + \lambda \cdot W(\Sigma_i). \tag{6}$$

As can be observed from Fig. 3, the head motion sequence $\widehat{Z}$ shows a break in the cluster transition even if colored noise is added or the number of clusters is increased. To avoid these discontinuities, a second smoothing technique is applied to this sequence which is based on spherical cubic interpolation [22]. With this technique, the 3-D Euler angles are interpolated in the unit sphere by using quaternion representation. This technique performs better than interpolating each Euler angle separately, which has been shown to produce jerky movements and undesired effects such as *Gimbal lock* [23].

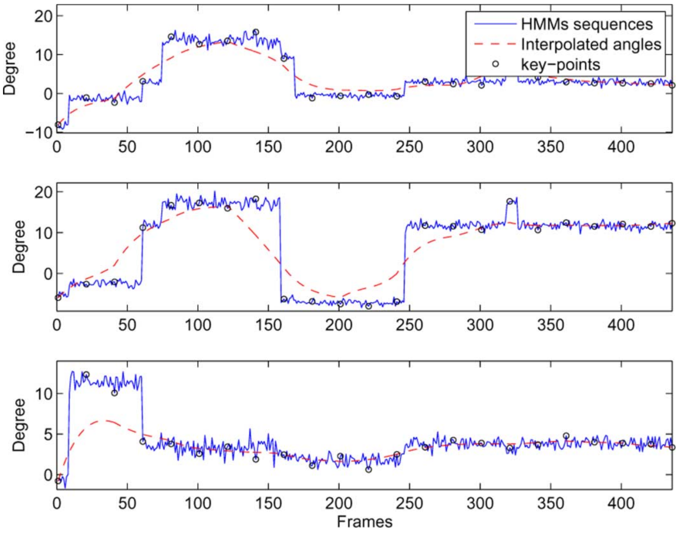


Fig. 3. Example of a synthesized head motion sequence. The figure shows the 3-D noisy signal $\widehat{Z}$ [equation (6)], with the key-points marked as a circle, and the 3-D interpolated signal $\widehat{X}$, used as head motion sequence.

In the interpolation step, the sequence $\widehat{Z}$ is down-sampled to six points per second to obtain equidistant frames. These frames are referred here as key-points and are marked as a circle in Fig. 3. These 3-D Euler angles points are then transformed into the quaternion representation [22]. Then, spherical cubic interpolation (SQUAD) is applied over these quaternion points. The SQUAD function builds upon the spherical linear interpolation, slerp. The functions slerp and SQUAD are defined by (7) and (8)

$$\begin{aligned} &\mathrm{slerp}(q_1, q_2, \mu) \\ &\quad = \frac{\sin(1-\mu)\theta}{\sin\theta} q_1 + \frac{\sin\mu\theta}{\sin\theta} q_2 \end{aligned} \tag{7}$$

$$\begin{aligned} &\mathrm{Squad}(q_1, q_2, q_3, q_4, \mu) \\ &\quad = \mathrm{slerp}\left(\mathrm{slerp}(q_1, q_4, \mu), \mathrm{slerp}(q_2, q_3, \mu), 2\mu(1-\mu)\right) \end{aligned} \tag{8}$$

where q_i are quaternions, $cos\theta = q_1 \cdot q_2$ and μ is a parameter that ranges between 0 and 1 and determines the frame position of the interpolated quaternion. Using these equations, the frames between key-points are interpolated by setting μ at the specific times to recover the original sample rate (120 frames per second). The final step in this smoothing technique is to transform the interpolated quaternions into the 3-D Euler angle representation.

Notice that colored noise is applied before the interpolation step. Therefore, the final sequence $\widehat{X}$ is a continuous and smooth head motion sequence without the jerky behavior of the noise. Fig. 3 shows the synthesized head motion sequence for one example sentence. The figure shows the 3-D noisy signal $\widehat{Z}$ (6), with the key-points marked as a circle, and the 3-D interpolated signal $\widehat{X}$, used here as head motion sequence.

Finally, for animation a blend shape face model composed of 46 blend shapes is used in this work (eye ball is controlled separately, as explained in Section VI). The head motion sequence $\widehat{X}$ is directly applied to the angle control parameters of the face model. The face modeling and rendering are done in Maya [24]. Details of the approach used to synthesize the face are given in Section VI.

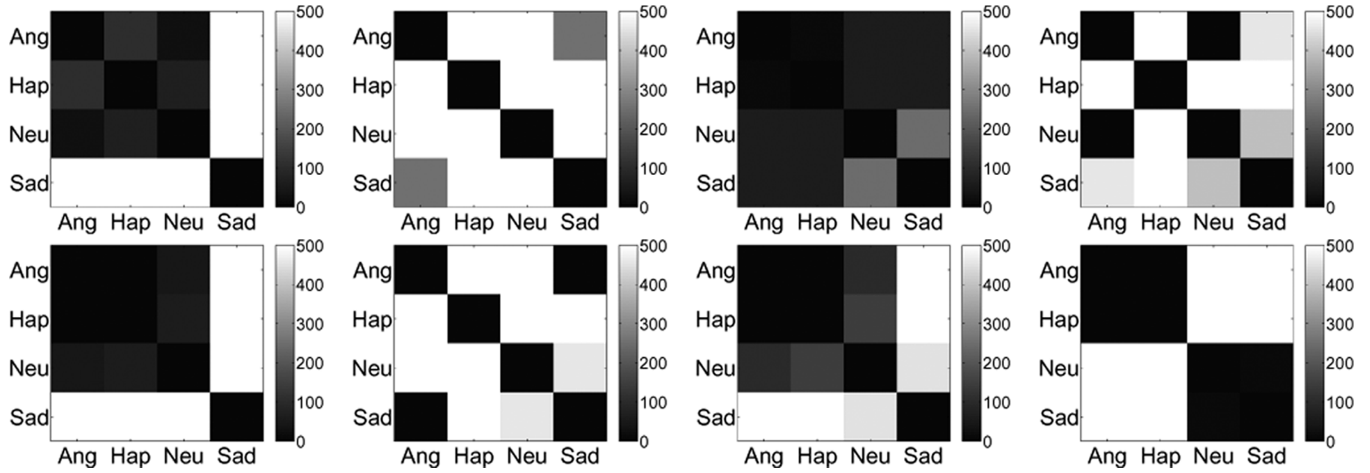


Fig. 4. Kullback–Leibler distance rate (KLDR) of HMMs for eight head-motion clusters. Lighter-shaded regions mean that the HMMs are different, and darker-shaded regions mean that the HMMs are similar. The figure reveals the differences between the emotion-dependent HMMs.

TABLE II
CANONICAL CORRELATION ANALYSIS BETWEEN ORIGINAL AND SYNTHESIZED HEAD MOTION SEQUENCES

	Neutral	Sadness	Happiness	Anger
Mean	0.86	0.88	0.89	0.91
Std	0.12	0.11	0.08	0.08

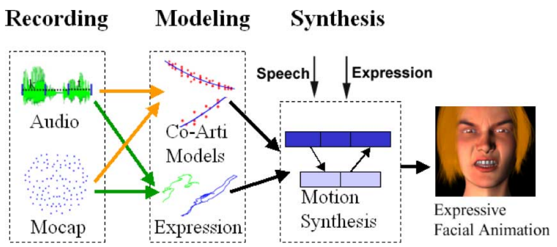


Fig. 5. Overview of the data-driven expressive facial animation synthesis system. The system is composed of three parts: recording, modeling, and synthesis.

C. Configuration of HMMs

The topology of the HMM is defined by the number and the interconnection of the states. In this particular problem, it is not completely clear which HMM topology provides the best description of the dynamics of the head motion. The most common topologies are the left-to-right topology (LR), in which only transitions in forward direction between adjacent states are allowed, and the ergodic (EG) topology, in which the states are fully connected. In our previous work [5], different HMM configurations for head motion synthesis were compared. The best performance was achieved by the LR topology with three states and two mixtures. One possible explanation is that LR topologies have fewer parameters than EG topologies, so they require less data for training. In this paper, the training data is even smaller, since emotion-dependent model are separately trained. Therefore, the HMMs used in the experiments were implemented using an LR topology with two states $(S = 2)$ and two mixtures $(M = 2)$.

Another important parameter that needs to be set is the number of HMMs, which is directly related to the number of clusters K. If K increases, the error quantization of the discrete representation of head poses decreases. However, the discrimination between models will significantly decrease and more training data will be needed. Therefore, there is a tradeoff between the quantization error and the intercluster discrimination. In our previous work, it was shown that realistic head motion sequences were obtained, even when only 16 clusters were used. In this paper, we also used a 16-word-sized codebook $(K = 16)$.

D. Objective Evaluation

Table II shows the average and standard deviation of the first-order canonical correlation between the original and the synthesized head motion sequences. As can be observed, the results show that the emotional sequences generated with the prosodic feature are highly correlated with the original signals $(r > 0.85)$. Notice that the first-order canonical correlation between the prosodic speech features and the original head motion sequence was about $r \approx 0.72$ (see Table I). This result shows that even though the prosodic speech features do not provide complete information to synthesize the head motion, the performance of the proposed system is notably high. This result is confirmed by the subjective evaluations presented in Section VII.

To compare how different the emotional HMMs presented in this paper are, an analytic approximation of the *Kullback–Leibler distance* (KLD) was implemented. The KLD, or relative entropy, provides the average discrimination information between the probability density functions of two random variables. Therefore, it can be used to compare distances between models. Unfortunately, there is no analytic close-form expression for Markov chains or HMMs. Therefore, numerical approximation, such as Monte Carlo simulation, or analytic upper bound for the KLD need to be used [25], [26]. Here, we use the analytic approximation of the *Kullback–Leibler distance rate* (KLDR) presented by Do, which is fast and deterministic. It has been shown that it produces similar results to those obtained through Monte Carlo simulations [26].

Fig. 4 shows the distance between emotional HMMs for eight head-motion clusters. Even though some of the emotional models are close, most of them are significantly different. Fig. 4

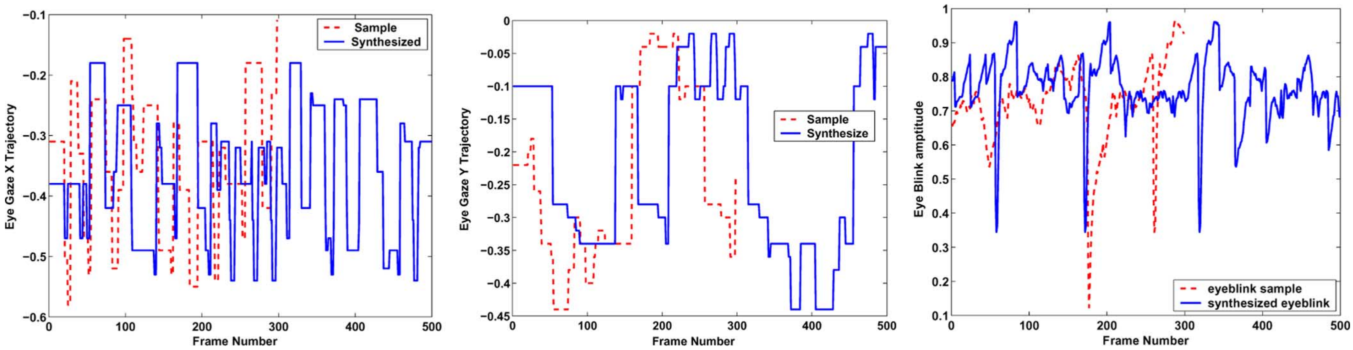


Fig. 6. Synthesized eye-gaze signals. Here, the solid line represents synthesized gaze signals, and dotted line represents captured signal samples.

reveals that happy and angry HMMs are closer than any other emotional category. As discussed in Section IV, the head motion characteristics of happy and angry utterances are similar, so it is not surprising that they share similar HMMs. This result indicates that a single model may be used to synthesize happy and angry head motion sequences. However, in the experiments presented in this paper a separate model was built for each emotion.

The readers are referred to [5] for further details about the head motion synthesis method.

VI. Facial Animation Synthesis

Although this paper is focused on head motion, for realistic animations, every facial component needs to be modeled. In this paper, expressive visual speech and eye motion were synthesized by the techniques presented in [27]–[30]. This section briefly described these approaches, which are very important to creating a realistic talking avatar.

Fig. 5 illustrates the overview of our data-driven facial animation synthesis system. In the recording stage, expressive facial motion and its accompanying acoustic signal are simultaneously recorded and preprocessed. In the modeling step, two approaches are used to learn the expressive facial animation: the neutral speech motion synthesis [27] and the dynamic expression synthesis [28]. The neutral speech motion synthesis approach learns explicit but compact speech coarticulation models by encoding coarticulation transition curves from recorded facial motion capture data, based on a weight-decomposition method that decomposes any motion frame into linear combinations of neighboring viseme frames. Given a new phoneme sequence, this system synthesizes corresponding neutral visual speech motion by concatenating the learned coarticulation models. The dynamic expression synthesis approach constructs a *phoneme-independent expression eigenspace* (PIEES) by a phoneme-based time warping and subtraction that extracts neutral motion signals from captured expressive motion signals. It is assumed that the above subtraction removes "phoneme-dependent" content from expressive speech motion capture data [28]. These phoneme-independent signals are further reduced by *principal component analysis* (PCA) to create an expression eigenspace, referred here PIEES [28]). Then, novel dynamic expression sequences are generated from the constructed PIEES by texture-synthesis approaches originally used for synthesizing similar but different images given a small image sample in graphics field. In the synthesis step, the synthesized neutral speech motions are weight-blended with the synthesized expression signals to generate expressive facial animation.

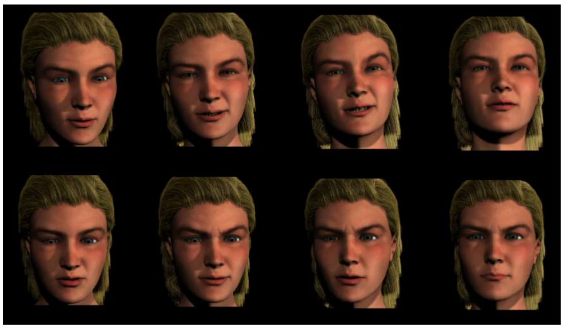


Fig. 7. Synthesized sequence for happy (top) and angry (bottom) sentences.

In addition to expressive visual speech synthesis, we used a texture-synthesis-based approach to synthesize realistic eye motion for talking avatars [29]. Eye gaze is one of the strongest cues in human communication. When a person speaks, he/she looks to our eyes to judge our interest and attentiveness, and we look into his/her eyes to signal our intent to talk. We adopted data-driven texture synthesis approaches [31], originally used in 2-D image synthesis, to the problem of realistic eye motion modeling. Eye gaze and aligned eye blink motion are considered together as an "eye motion texture" sample. The samples are then used to synthesize novel but similar eye motions. In our work, the patch-based sampling algorithm [31] is used, due to its time efficiency. The basic idea is to generate one texture patch (fixed size) at a time, randomly chosen from qualified candidate patches in the input texture sample. Fig. 6 illustrates the synthesized eye motion results.

Fig. 7 shows frames of the synthesized data for happy and angry sentences. The text of sentences are "We lost them at the last turnoff" and "And so you just abandoned them?", respectively.

VII. Evaluation of Emotional Perception From Animated Sequences

To analyze whether head motion patterns change the emotional perception of the speaker, various combinations of facial animations were created, including deliberate mismatches between the emotional content of the speech and the emotional

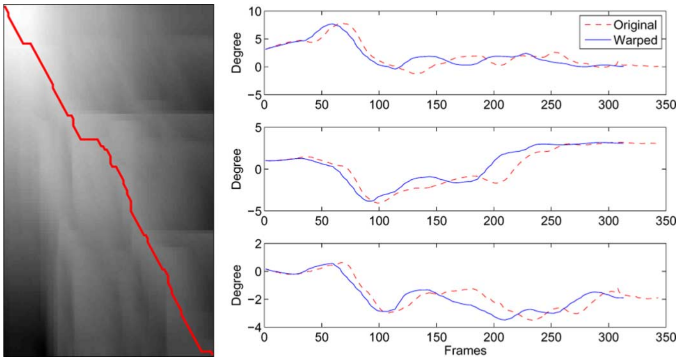


Fig. 8. Dynamic time warping. Optimums path (left panel) and warped head motion signal (right panel).

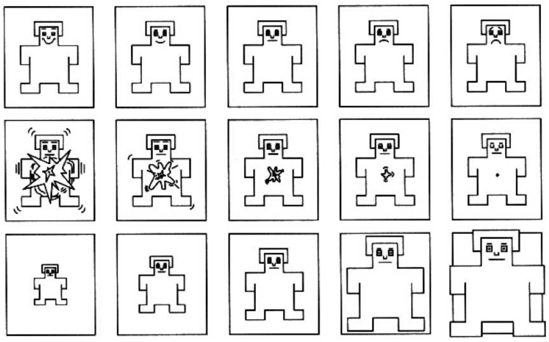


Fig. 9. Self-assessment manikins [35]. The rows illustrate: top, *Valence* [1-positive, 5-negative]; middle, *Activation* [1-excited, 5-calm]; and bottom, *Dominance* [1-weak, 5-strong].

pattern of head motion, for four sentences in our database (one for each emotion). Given that the actress repeated each of these sentences under four emotional states, we generated facial animations with speech associated with one emotion, and recorded head motions associated with a different emotion. Altogether, 16 facial animations were created (four sentences × four emotions). The only consideration was that the timing between the repetitions of these sentences was different, and it was overcome by aligning the sentences using *dynamic time warping* (DTW) [32]. After the acoustic signals were aligned, the optimal synchronization path was applied to the head motion sequences and were used to create the mismatched facial animations (Fig. 8). In the DTW process, some emotional characteristics could be removed, especially for sad sentences, in which the syllable duration is inherently longer than in other emotions. However, most of the dynamic behaviors of emotional head motion sequences are nevertheless preserved. Notice that even though lip and eye motions were also included in the animations, the only parameter that was changed was the head motion.

For assessment, 17 human subjects were asked to rate the emotions conveyed and the naturalness of the synthesized data presented as short animation videos. The animations were presented to the subjects in a random order. The evaluators received instructions to rate their overall impression of the animation and not individual aspects such as head movement or voice quality.

The emotional content was rated using three emotional attributes ("primitives"), namely *valence*, *activation*, and *dominance*, following a concept proposed by Kehrein [33]. *Valence* describes the positive or negative strength of the emotion, *activation* details the excitation level (high versus low), and *dominance* refers to the apparent strength or weakness of the speaker.

Describing emotions by attributes in an emotional space is a powerful alternative to assigning class labels such as *sadness* or *happiness* [34], since the primitives can be easily used to capture emotion dynamics and speaker dependencies. Also, there are different degrees of emotions that cannot be measured if only category labels are just used (e.g., how "happy" or "sad" the stimuli is). Therefore, these emotional attributes are more suitable to evaluate the emotional salience in human perception. Notice that for animation, we propose to use categorical classes, since the specifications of the expressive animations are usually described in terms of category emotions and not emotional attributes.

TABLE III
SUBJECTIVE AGREEMENT EVALUATION, VARIANCE ABOUT THE MEAN

Valence	Activation	Dominance	Naturalness
0.52	0.49	0.48	0.97

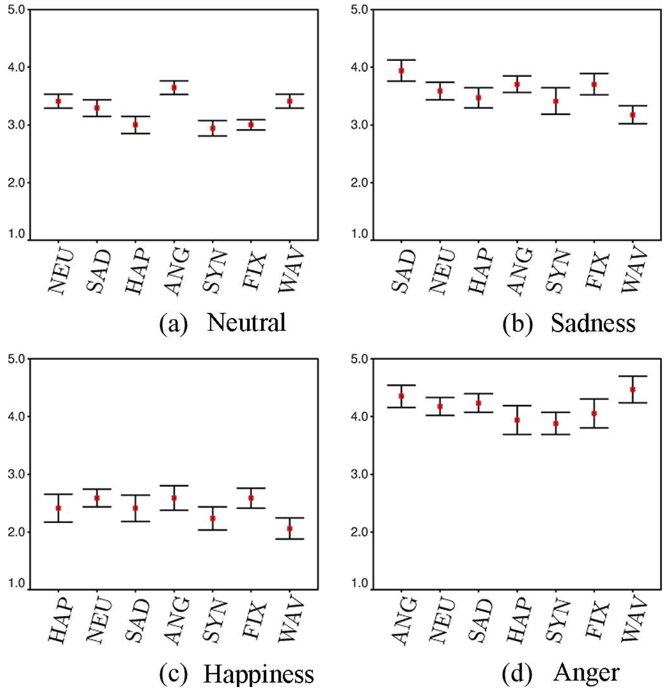
As a tool for emotion evaluation, *self-assessment manikins* (SAMs) have been used [35], [36], as shown in Fig. 9. For each emotion primitive, the evaluators had to select one out of five iconic images ("manikins"). The SAMs system has been previously used successfully for assessment in emotional speech, showing low standard deviation and high interevaluator agreement [36]. Also, using a text-free assessment method bypasses the difficulty that each evaluator has on his/her individual understanding of linguistic emotion labels.

For each SAM row in Fig. 9, the selection was mapped to the range 1 to 5 from left to right. The naturalness of the animation was also rated using a five-point scale. The extremes were called *robot-like* (value 1), and *human-like* (value 5).

In addition to the animations, the evaluators also assessed the underlying speech signal without the video signal. This rating was used as a reference.

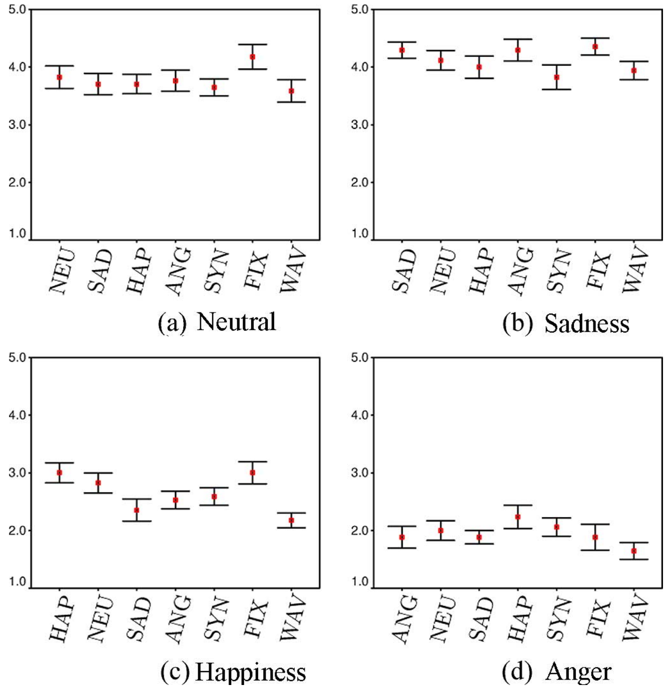
Table III presents the interevaluator average variance in the scores rated by the human subjects, in terms of emotional attributes and naturalness. These measures confirm the high interevaluator agreement of emotional attributes. The results also show that the naturalness of the animation was perceived slightly different between the evaluators, which suggest that the concept of naturalness is more person-dependent than the emotional attribute. However, this variability does not bias our analysis, since we will consider differences between the scores given to the facial animations.

Figs. 10–12 show the results of the subjective evaluations in terms of emotional perception. Each quadrant has the error bars for six different facial animations with head motion synthesized with (from left to right): original sequence (without mismatch), three mismatched sequences (one for each emotion), synthesized sequence, and, fixed head poses. In addition, the result for audio (WAV) was also included. For example, the second error bar in the most upper-left block of Fig. 10 shows the *valence* assessment for the animation with neutral speech and sad head motion sequence. To measure whether the difference in the



(a) Neutral (b) Sadness (c) Happiness (d) Anger

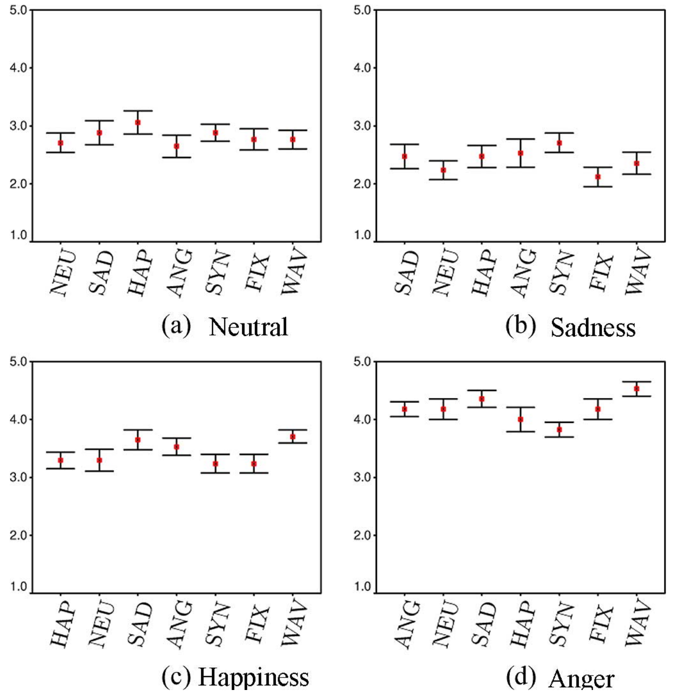
Fig. 10. Subjective evaluation of emotions conveyed in *valence* domain [1-positive, 5-negative]. Each quadrant has the error bars of facial animations with head motion synthesized with (from left to right): original head motion sequence (without mismatch), three mismatched head motion sequences (one for each emotion), synthesized sequence (SYN), and, fixed head poses (FIX). The result of the audio without animation is also shown (WAV).



(a) Neutral (b) Sadness (c) Happiness (d) Anger

Fig. 11. Subjective evaluation of emotions conveyed in *activation* domain [1-excited, 5-calm]. Each quadrant has the error bars of facial animations with head motion synthesized with (from left to right): original head motion sequence (without mismatch), three mismatched head motion sequences (one for each emotion), synthesized sequence (SYN), and, fixed head poses (FIX). The result of the audio without animation is also shown (WAV).

means of two of these groups are significant, the two-tailed Student's t-test was used.



(a) Neutral (b) Sadness (c) Happiness (d) Anger

Fig. 12. Subjective evaluation of emotions conveyed in *activation* domain [1-weak, 5-strong]. Each quadrant has the error bars of facial animations with head motion synthesized with (from left to right): original head motion sequence (without mismatch), three mismatched head motion sequences (one for each emotion), synthesized sequence (SYN), and, fixed head poses (FIX). The result of the audio without animation is also shown (WAV).

In general, the figures show that the emotional perception changes in presence of different emotional head motion patterns.

In the *valence* domain (Fig. 10), the results show that when the talking avatar with angry speech is animated with happy head motion, the attitude of the character is perceived more positive. The t-test result indicates that the difference in the scores between the mismatched and the original animations is statistical significant ($t = 2.384$, $df = 16$, $p = 0.03$). The same result is also held when sad and neutral speeches are synthesized with happy head motion sequences. For these pairs, the t-test results are ($t = 2.704$, $df = 16$, $p = 0.016$), and ($t = 2.384$, $df = 16$, $p = 0.03$), respectively. These results suggest that the temporal pattern in happy head motion makes the animation to have a more positive attitude.

Fig. 10 also shows that when neutral or happy speech is synthesized with angry head motion sequences, the attitude of the character is perceived slightly more negative. However, the t-test reveals that these differences are not completely significant.

In the *activation* domain (Fig. 11), the results show that the animation with happy speech and angry head motion sequence is perceived with a higher level of excitation. The t-test result indicates that the differences in the scores are significant ($t = 2.426, df = 16, p = 0.027$). On the other hand, when the talking avatar with angry speech is synthesized with happy head motion, the animation is perceived slightly more calmed, as observed in Fig. 11. Notice that in the acoustic domain, anger is usually perceived more excited than happiness, as reported in [37], [38], and as shown in the evaluations presented here (see the last bars of (c) and (d) in Fig. 11). Our results suggest that the same trend is observed in the head motion domain: angry head motion sequences are perceived more excited than happy head motion sequences.

When animation with happy speech is synthesized with sad head motion, the talking avatar is perceived more excited ($t = 2.184$, $df = 16$, $p = 0.044$). It is not clear whether this result, which is less intuitive than the previous results, may be a true effect generated by the combination of modalities, which together produce a different percept (similar to the McGurk effect [39]), or may be an artifact introduced in the warping process.

In the *dominance* domain, Fig. 12 shows that the mismatched head motion sequences do not modify in significant ways how dominant the talking avatar is perceived as compared to the animations with the original head motion sequence. For example, the animation with neutral speech and with happy head motion is perceived slightly stronger. A similar result is observed when animation with happy speech is synthesized with an angry head motion sequence. However, the t-test reveals that the differences in the means of the scores of the animations with mismatched and original head motion sequences are not statistical significant: ($t = -1.461$, $df = 16$, $p = 0.163$), and ($t = -1.289$, $df = 16$, $p = 0.216$), respectively. These results suggest that head motion has a lower influence in the *dominance* domain than in the *valence* and *activation* domains. A possible explanation of this result is that human listeners may be more cognizant of other facial gestures such as eyebrow and forehead motion to infer how dominant the speaker is. Also, the intonation and the energy of the speech may play a more important role than head motion gesture for dominance perception.

Notice that the emotional perception of the animations synthesized without head motion usually differs from the emotion perceived from the animations with the original sequences. This is especially clear in the *valence* domain, as can be observed in Fig. 10. The differences in the mean of the scores in Fig. 10(a) and (b) between the fixed head motion and the original animations are statistical significant, as shown by the t-test: Fig. 10(a) ($t = 2.746$, $df = 16$, $p = 0.014$) and Fig. 10(b) ($t = 2.219$, $df = 16$, $p = 0.041$). For Fig. 10(c) and (d) the differences in the means observed in the figure are not totally significant: Fig. 10(c) ($t = -1.144$, $df = 16$, $p = 0.269$), Fig. 10(d) ($t = 2.063$, $df = 16$, $p = 0.056$). This result suggests that head motion has a strong influence on the perception of how positive or negative the affective state of the avatar is.

Figs. 10–12 also suggest that the emotional perception of the acoustic signal changes when facial animation is added, emphasizing the multimodal nature of human emotional expression. This is particularly noticeable in sad sentences, in which the t-test between the means of the scores of the original animation and the acoustic signal gives ($t = 4.190$, $df = 16$, $p = 0.01$) in the *valence* domain, and ($t = 2.400$, $df = 16$, $p = 0.029$) in the *activation* domain. Notice that in this analysis, the emotional perception of the acoustic signal is directly compared to the emotional perception of the animation. Therefore, the differences in the results are due to not only the head motion, but also the other facial gestures included in the animations (see Section VI). These results suggest that facial gestures (including head motion) are extremely important to convey the desired emotion.

TABLE IV
NATURALNESS ASSESSMENT OF RIGID HEAD MOTION SEQUENCES
[1-*robot-like*, 5-*human-like*]

Head Motion	Neutral		Sadness		Happiness		Anger	
Data	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Original	3.76	0.90	3.76	0.83	3.71	0.99	3.00	1.00
Synthesized	4.00	0.79	3.12	1.17	3.82	1.13	3.71	1.05
Fixed Head	3.00	1.06	2.76	1.25	3.35	0.93	3.29	1.45

Table IV shows how the listeners assessed the naturalness of the facial animation with head motion sequences generated with the original and with the synthesized data. It also shows the results for animations without head motion. These results show that head motion significantly improves the naturalness of the animation. Furthermore, with the exception of sadness, the synthesized sequences were perceived even more natural than the real head motion sequences, which indicates that the head motion synthesis approach presented here was able to generate realistic head motion sequences.

VIII. CONCLUSION

Rigid head motion is an important component in human–human communication that needs to be appropriately added into computer facial animations. The subjective evaluations presented in this paper show that including head motion into talking avatars significantly improves the naturalness of the animations.

The statistical measures obtained from the audiovisual database reveal that the dynamics of head motion sequences are different under different emotional states. Furthermore, the subjective evaluations also show that head motion changes the emotional perception of the animation, especially in the valence and activation domain. The implications of these results are significant. Head motion can be appropriately included in the facial animation to emphasize its emotional content.

In this paper, an extension of our previous head motion synthesis approach was implemented to handle expressive animations. Emotion-dependent HMMs were designed to generate the most likely head motion sequences driven by speech prosody. The objective evaluations show that the synthesized and the original head motion sequences were highly correlated, suggesting that the dynamics of head motion were successfully modeled by the use of prosodic features. Also, the subjective evaluations show that on average, the animations with synthesized head motion were perceived as realistic when compared with the animation with the original head motion sequence.

The results of this paper indicate that head motion provides important emotional information that can be used to discriminate between emotions. It is interesting to notice that in the current multimodal emotion recognition systems, head motion is usually removed in the preprocessing step. Although head motion is speaker-dependent, as is any gesture, it could be used to determine emotional versus nonemotional affective states in human–machine interaction systems.

We are currently working to modify the system to generate head motion sequences that not only look natural, but also preserve the emotional perception of the input signal. Even though the proposed approach generates realistic head motion sequences, the results of the subjective evaluations show that in some cases the emotional content in the animations were perceived slightly different from the original sequences. Further research is needed to shed light into the underlying reasons. It may be that different combinations of modalities create different emotion percepts, similar to the famous McGurk effect [39]. Or, it may be that the modeling and techniques used are not perfect enough, creating artifacts. For instance, it may be that the emotional HMMs preserve the phase but not the amplitude of the original head motion sequence. If this is the case, the amplitude of the head motion could be externally modified to match the statistics of the desired emotion category.

One limitation of this work is that head motion sequences considered here did not include the three DOF of head translation. Since human neck translates the head, especially back and forward, our future work will investigate how to jointly model the six DOF of the head.

In this paper, head motion sequences from a single actress were studied, which is generally enough for synthesis purposes. An open area that requires further work is the analysis of interperson variabilities and dependencies in head motion patterns. We are planning to collect more data from different subjects to address the challenging questions triggered by this topic.

We are also studying the relation between the speech and other facial gestures, such as eyebrow motion. If these gestures are appropriately included, we believe that the overall facial animation will be perceived to be more realistic and compelling.

Acknowledgment

The authors would like to thank J.P. Lewis and M. Bulut for helping data capture, H. Itokazu, B. St. Clair, S. Drost, and P. Fox for face model preparation.

References

[1] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Bateson, “Visual prosody and speech intelligibility: Head movement improves auditory speech perception,” *Psychol. Sci.*, vol. 15, no. 2, pp. 133–137, Feb. 2004.

[2] H. Hill and A. Johnston, “Categorizing sex and identity from the biological motion of faces,” *Current Biol.*, vol. 11, no. 11, pp. 880–885, Jun. 2001.

[3] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, “Visual prosody: Facial movements accompanying speech,” in *Proc. IEEE Int. Conf. Autom. Faces Gesture Recognition*, Washington, DC, May 2002, pp. 396–401.

[4] T. Kuratate, K. G. Munhall, P. E. Rubin, E. V. Bateson, and H. Yehia, “Audio-visual synthesis of talking faces from speech production correlates,” in *Proc. 6th Eur. Conf. Speech Commun. Technol. Eurospeech*, Budapest, Hungary, Sep. 1999, pp. 1279–1282.

[5] C. Busso, Z. Deng, U. Neumann, and S. Narayanan, “Natural head motion synthesis driven by acoustic prosodic features,” *Comput. Animation Virtual Worlds*, vol. 16, no. 3–4, pp. 283–290, Jul. 2005.

[6] M. Costa, T. Chen, and F. Lavagetto, “Visual prosody analysis for realistic motion synthesis of 3-D head models,” in *Proc. Int. Conf. Augmented, Virtual Environments Three Dimensional Imaging (ICAV3-D)*, Ornos, Mykonos, Greece, May–Jun. 2001, pp. 343–346.

[7] R. W. Picard, “Affective computing,” MIT Media Lab., Perceptual Comput. Section, Mass. Inst. Technol., Cambridge, MA, Tech. Rep. 321, Nov. 1995.

[8] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Bechet, B. Douville, S. Prevost, and M. Stone, “Animated conversation: Rule-based generation of facial expression gesture and spoken intonation for multiple conversational agents,” in *Comput. Graphics (Proc. ACM SIGGRAPH’94)*, Orlando, FL, 1994, pp. 413–420.

[9] S. Kettebekov, M. Yeasin, and R. Sharma, “Prosody based audiovisual coanalysis for coverbal gesture recognition,” *IEEE Trans. Multimedia*, vol. 7, no. 2, pp. 234–242, Apr. 2005.

[10] M. Brand, “Voice puppetry,” in *Proceedings of the 26th Annu. Conf. Comput. Graphics Interactive Tech. (SIGGRAPH)*, New York, 1999, pp. 21–28.

[11] K. Kakihara, S. Nakamura, and K. Shikano, “Speech-to-face movement synthesis based on HMMS,” in *IEEE Int. Conf. Multimedia Expo (ICME)*, New York, Apr. 2000, vol. 1, pp. 427–430.

[12] B. Hartmann, M. Mancini, and C. Pelachaud, “Formational parameters and adaptive prototype instantiation for MPEG-4 compliant gesture synthesis,” in *Proc. Comput. Animation*, Geneva, Switzerland, Jun. 2002, pp. 111–119.

[13] S. Kopp and I. Wachsmuth, “Model-based animation of co-verbal gesture,” in *Proc. Comput. Animation*, Geneva, Switzerland, Jun. 2002, pp. 252–257.

[14] H. Yehia, T. Kuratate, and E. V. Bateson, “Facial animation and head motion driven by speech acoustics,” in *Proc. 5th Seminar Speech Prod.: Models Data*, Kloster Seeon, Bavaria, Germany, May 2000, pp. 265–268.

[15] M. B. Stegmann and D. D. Gomez, “A brief introduction to statistical shape analysis,” in *Informatics and Mathematical Modelling, Technical Univ. Denmark*, Mar. 2002 [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?403

[16] Z. Deng, C. Busso, S. Narayanan, and U. Neumann, “Audio-based head motion synthesis for avatar-based telepresence systems,” in *ACM SIGMM 2004 Workshop on Effective Telepresence (ETP 2004)*. New York: ACM Press, 2004, pp. 24–30.

[17] P. Boersma and D. Weeninck, “Praat, a system for doing phonetics by computer,” Inst. Phonetic Sci., Univ. Amsterdam, Amsterdam, The Netherlands, Tech. Rep. 132, 1996 [Online]. Available: http://www.praat.org.

[18] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, “An articulatory study of emotional speech production,” in *Proc. 9th Eur. Conf. Speech Commun. Technol. (Interspeech2005—Eurospeech)*, Lisbon, Portugal, Sep. 2005, pp. 497–500.

[19] Y. Linde, A. Buzo, and R. Gray, “An algorithm for vector quantizer design,” *IEEE Trans. Commun.*, vol. COMM-28, no. 1, pp. 84–95, Jan. 1980.

[20] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland*, The HTK Book*. Cambridge, U.K.: Entropic Cambridge Research Laboratory, 2002.

[21] L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[22] D. Eberly*, 3-D Game Engine Design: A Practical Approach to Real-Time Computer Graphics*. San Francisco, CA: Morgan Kaufmann, 2000.

[23] K. Shoemake, “Animating rotation with quaternion curves,” in *Computer Graphics (Proc. SIGGRAPH85)*, Jul. 1985, vol. 19, no. 3, pp. 245–254.

[24] Maya(r) software. Alias Systems Division, Silicon Graphics, Ltd., 2005 [Online]. Available: http://www.alias.com

[25] J. Silva and S. Narayanan, “Average divergence distance as a statistical discrimination measure for hidden Markov models,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 3, pp. 890–906, May 2006.

[26] M. Do, “Fast approximation of Kullback–Leibler distance for dependence trees and hidden Markov models,” *IEEE Signal Process. Lett.*, vol. 10, no. 4, pp. 115–118, Apr. 2003.

[27] Z. Deng, J. Lewis, and U. Neumann, “Synthesizing speech animation by learning compact speech co-articulation models,” in *Proc. Computer Graphics Int. (CGI 2005)*, Stony Brook, NY, Jun. 2005, pp. 19–25.

[28] Z. Deng, M. Bulut, U. Neumann, and S. Narayanan, “Automatic dynamic expression synthesis for speech animation,” in *Proc. IEEE 17th Int. Conf. Comput. Animation Social Agents (CASA 2004)*, Geneva, Switzerland, Jul. 2004, pp. 267–274.

[29] Z. Deng, J. Lewis, and U. Neumann, “Automated eye motion using texture synthesis,” *IEEE Comput. Graphics Applicat.*, vol. 25, no. 2, pp. 24–30, Mar./Apr. 2005.

[30] Z. Deng, U. Neumann, J. Lewis, T. Kim, M. Bulut, and S. Narayanan, "Expressive facial animation synthesis by learning speech coarticultion and expression spaces," *IEEE Trans. Visualization Comput. Graphics*, vol. 12, no. 6, pp. 1523–1534, Nov./Dec. 2006.

[31] L. Liang, C. Liu, Y. Xu, B. Guo, and H. Shum, "Real-time texture synthesis by patch-based sampling," *ACM Trans. Graphics*, vol. 20, no. 3, pp. 127–150, Jul. 2001.

[32] J. Deller, J. Hansen, and J. Proakis, *Discrete-Time Processing of Speech Signals*. Piscataway, NJ: IEEE Press, 2000.

[33] R. Kehrein, "The prosody of authentic emotions," in *Proc. Speech Prosody*, Aix-en-Provence, France, Apr. 2002, pp. 423–426.

[34] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Commun.*, vol. 40, no. 1–2, pp. 5–32, Apr. 2003.

[35] L. Fischer, D. Brauns, and F. Belschak, *Zur Messung von Emotionen in der angewandten Forschung*. Lengerich, Germany: Pabst, 2002.

[36] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'05)*, San Juan, PR, Dec. 2005, pp. 381–385.

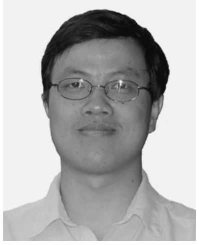
[37] R. Cowie, E. Douglas-Cowie, B. Apolloni, J. Taylor, A. Romano, and W. Fellenz, "What a neural net needs to know about emotion words," in *Proc. Circuits, Syst., Commun., Comput. (CSCC)*, Athens, Greece, Jul. 1999, pp. 5311–5316.

[38] M. Schröder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. Gielen, "Acoustic correlates of emotion dimensions in view of speech synthesis," in *Proc. Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Aalborg, Denmark, Sep. 2001, vol. 1, pp. 87–90.

[39] H. McGurk and J. W. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, Dec. 1976.

Carlos Busso (S'01) received the B.S. and M.S. degrees (with high honors) in electrical engineering from University of Chile, Santiago, Chile, in 2000 and 2003, respectively. He is currently pursuing the Ph.D. degree in electrical engineering at the University of Southern California (USC), Los Angeles.

Since 2003, he has been a student member at the Speech Analysis and Interpretation Laboratory (SAIL), USC. His research interests are in digital signal processing, speech and video signal processing, and multimodal interfaces. His current research includes modeling and understanding human communication and interaction, with application in recognition and synthesis.



Zhigang Deng (M'06) received the B.S. degree in mathematics from Xiamen University, Xiamen, China, in 1997, the M.S. degree in computer science from Peking University, Beijing, China, in 2000, and the Ph.D. degree in computer science from the University of Southern California, Los Angeles, in 2006.

He is an Assistant Professor in the Department of Computer Science, University of Houston, Houston, TX. His research interests include computer graphics, computer animation, human–computer interaction, and visualization.

Dr. Deng is a member of the ACM, ACM SIGGRAPH, and the IEEE Computer Society.

Michael Grimm (S'03) received the M.S. degree in electrical engineering (Dipl.-Ing.) from the University of Karlsruhe (TH), Karlsruhe, Germany, in 2003. He is currently pursuing the Ph.D. degree in signal processing at the Institute for Communications Engineering (INT), TH.

He was a Visiting Scientist with the Speech Analysis and Interpretation Lab (SAIL) of the University of Southern California (USC), Los Angeles CA, in 2005. His research interests include digital speech processing, pattern recognition, and natural language understanding. His research activities focus on audio-visual scene analysis and user modeling in the context of man-robot interaction.

Ulrich Neumann received the M.S. degree in electrical engineering from the State University of New York, Buffalo, in 1980 and the Ph.D. degree in computer science from the University of North Carolina, Chapel Hill, in 1993, where his focus was on parallel algorithms for interactive volume-visualization.

He is an Associate Professor of Computer Science, with a joint appointment in Electrical Engineering, University of Southern California (USC), Los Angeles. His current research relates to immersive environments and virtual humans. He held the Charles Lee Powell Chair of Computer Science and Electrical Engineering and was the Director of the Integrated Media Systems Center (IMSC), an NSF Engineering Research Center (ERC) from 2000 to 2004. He directs the Computer Graphics and Immersive Technologies (CGIT) Laboratory at USC. In his commercial career, he designed multiprocessor graphics and DSP systems, cofounded a video game corporation, and independently developed and licensed electronic products.

Dr. Neumann won an NSF CAREER Award in 1995 and the Jr. Faculty Research Award at USC in 1999.

Shrikanth Narayanan (S'88–M'95–SM'02) received the Ph.D. degree from the University of California, Los Angeles, in 1995.

He was with AT&T Research (originally AT&T Bell Labs), first as a Senior Member, and later as a Principal Member of the Technical Staff from 1995 to 2000. Currently, he is a Professor of Electrical Engineering, with joint appointments in Computer Science, Linguistics, and Psychology at the University of Southern California (USC), Los Angeles. He is a member of the Signal and Image Processing Institute and a Research Area Director of the Integrated Media Systems Center, an NSF Engineering Research Center, at USC. He has published over 190 papers and has ten granted/pending U.S. patents. His research interests are in signals and systems modeling with applications to speech, language, multimodal, and biomedical problems.

Prof. Narayanan was an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (2000–2004) and is currently an Associate Editor of the *IEEE Signal Processing Magazine*. He serves on the Speech Processing and Multimedia Signal Processing technical committees of the IEEE Signal Processing Society and the Speech Communication committee of the Acoustical Society of America. He is a Fellow of the Acoustical Society of America and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is a recipient of an NSF CAREER Award, USC Engineering Junior Research Award, USC Electrical Engineering Northrop Grumman Research Award, a Provost Fellowship from the USC Center for Interdisciplinary research, a Mellon Award for Excellence in Mentoring, and a corecipient of a 2005 Best Paper Award from the IEEE Signal Processing Society.