Live Speech Driven Head-and-Eye Motion Generators

Binh H. Le, Xiaohan Ma, and Zhigang Deng, *Senior Member, IEEE*

Abstract—This paper describes a fully automated framework to generate realistic head motion, eye gaze, and eyelid motion simultaneously based on live (or recorded) speech input. Its central idea is to learn separate yet inter-related statistical models for each component (head motion, gaze, or eyelid motion) from a pre-recorded facial motion dataset: i) Gaussian Mixture Models and gradient descent optimization algorithm are employed to generate head motion from speech features; ii) Nonlinear Dynamic Canonical Correlation Analysis model is used to synthesize eye gaze from head motion and speech features, and iii) non-negative linear regression is used to model voluntary eye lid motion and log-normal distribution is used to describe involuntary eye blinks. Several user studies are conducted to evaluate the effectiveness of the proposed speech-driven head and eye motion generator using the well-established paired comparison methodology. Our evaluation results clearly show that this approach can significantly outperform the state-of-the-art head and eye motion generation algorithms. In addition, a novel mocap+video hybrid data acquisition technique is introduced to record high-fidelity head movement, eye gaze, and eyelid motion simultaneously.

Index Terms—Facial Animation, Head and Eye Motion Coupling, Head Motion Synthesis, Gaze Synthesis, Blinking Model, and Live Speech Driven

1 Introduction

As one of the prominent communicative signals, non-verbal facial gestures including (but not limited to) the movement of head, eye gaze and eyelid, play a central role to facilitate natural and engaging human-human communication and interaction. Often, humans can easily read and infer the affective and mental states of a person based on his/her facial gesture cues. In concert with the verbal channel, non-verbal facial gestures can be used to visually emphasize certain words and syllables of importance in a conversation [1]. Despite the clear importance of modeling and animating human facial gestures, efficiently generating lifelike and perceptually believable facial gesture such as head and eye motion on-the-fly has still been a largely unresolved research challenge in computer animation, virtual human modeling, and human computer interaction fields for decades.

Generating speech-driven eye and head movement on a talking avatar is challenging, because: 1) The association between speech and facial gestures is essentially a nondeterministic many-to-many mapping, since numerous factors including personality traits, culture background, conversation contexts, and affective state collectively determine the gestures in a conversation. 2) The tie between speech and eye motion has been significantly less observed and documented in the literature. Besides the above difficulties, live speech driven eye and head motion generation is even more challenging due to the following additional reasons: i) "Live speech driven" implies the real-time performance of the algorithm; hence, its runtime performance needs to be highly efficient. ii) Since only the prior and current information enclosed in live speech is available as the runtime input, many widely-used dynamic programming schemes cannot be directly used for the task.

Researchers have conducted various research efforts to model lifelike eye motion and natural head movement on a talking avatar by either using empirical rules [2; 3; 4; 5; 6; 7] or learning statistical models from pre-recorded human motion data [8; 9; 10; 11]. These methods, however, are only suitable for offline synthesis and cannot be driven by live speech. Furthermore, these methods focus on either eye motion or head movement; but, none of them can generate the coordinated head and eye movement in an inter-connected framework. Recently, Levine et al. [12; 13] developed data-driven, live speech driven body language controllers that are capable of dynamically synthesizing realistic body and hand gestures based on live speech input. Although realistic head movement was demonstrated in their method, how to efficiently generate plausible speech-synchronized eye motion has not been addressed.

In this paper, we propose a novel, fully automated framework to generate realistic and synchronized eye and head movement on-the-fly, based on live or pre-recorded speech input. In particular, the synthesized eye motion in our framework includes not only the traditional eye gaze but also the subtle eyelid movement and blink. The central idea of this framework is to learn separate yet inter-related statistical models for each component (head motion, gaze, or eyelid motion) from a pre-recorded facial motion dataset. Specifically, it includes: i) Gaussian Mixture Models and gradient descent optimization algorithm are employed to

• *Binh H. Le, Xiaohan Ma and Zhigang Deng are with the Department of Computer Science, University of Houston, 4800 Calhoun Road, Houston, TX, 77204-3010.*
** This work was accepted by IEEE TVCG on Feb 8th, 2012.*
E-mails: bhle2|xiaohan|zdeng@cs.uh.edu

generate head motion from speech features; ii) Nonlinear Dynamic Canonical Correlation Analysis model is used to synthesize eye gaze from head motion and speech features, and iii) Non-negative linear regression is used to model voluntary eye lid motion and log-normal distribution is used to describe involuntary eye blink.

Through comparative user studies, we show that our approach outperforms various state-of-the-art facial gesture generation algorithms, and the results by our approach are measurably close to the ground-truth. In addition to our contribution in synchronized head and eye motion generation, we also develop a novel mocap+video hybrid data acquisition technique to simultaneously capture high-fidelity head movement, eye gaze, and eyelid motion of a human subject.

To the best of our knowledge, our approach is the first reported system that can dynamically and automatically generate realistic, speech-synchronized eye and head movement solely based on arbitrary live speech input. In particular, we believe that the *live speech driven* and *synchronized head movement, eye gaze, and eyelid movement characteristics* of our approach well separate it from existing similar techniques. Our approach can be widely used for a broad variety of interactive avatars, computer-mediated communication, and virtual world applications.

2 Previous and Related Work

During the past several decades, many researchers have studied various aspects of facial animation including face modeling [14; 15; 16; 17; 18; 19], speech synthesis [20; 21; 22], deformation [23; 24; 25; 26], and expression transferring [27; 28; 29; 30]. Comprehensively reviewing these efforts is beyond the scope of this paper (interested readers are referred to [31]). We only briefly review recent efforts that are most related to this work.

Head motion synthesis: Some of the early work crafted a number of empirical rules to govern the correspondence between input texts and head movement [32; 33; 34]. An increasing amount of interest has been drawn to speech-driven head motion synthesis in recent years [5; 35]. For example, Graf *et al.* [3] studied the conditional probabilities of pitch accents accompanied by primitive head movement (*e.g.*, nodding, nodding with overshoot, and abrupt swinging in one direction). Chuang and Bregler [11] built a database of head motion sequences indexed by pitch features. Then, new head motion can be synthesized by searching for a globally optimal path that maximizes the match of pitch features. Researchers also trained various forms of HMMs for novel head motion synthesis based on a collected audio-head motion dataset [10; 36; 37; 38; 39]. In addition, Gratch and his colleagues have reported statistical models that learn empirical rules from annotated conversation datasets to generate head movement (in particular, nodding) on listening agents [40; 41; 42].

Eye gaze generation. Researchers have proposed many techniques to model eye gaze of avatars in virtual environments or in a conversation [2; 43; 8; 4; 9; 6; 7; 44; 45; 46]. For example, Lee *et al.* [8] statistically analyzed the recorded gaze data and then synthesized saccade movement using the first-order statistics. Thiébaux *et al.* [44] proposed a parameterized gaze controller on top of the open-source SmartBody virtual human system. Ma and Deng [45] proposed an effective technique to synthesize natural eye gaze by modeling the linear coupling between eye and head movement. Oyekoya *et al.* [46] built a rule based system to model the gaze from the saliency attract attention in virtual environments. However, speech content is not taken into consideration in all the above approaches. As pointed out in existing literature [8; 47], humans have different gaze patterns depending on their conversational modes (*e.g.*, talking or listening).

Eyelid and blink motion generation. Compared with the above gaze synthesis, fewer research efforts have been focused on eyelid motion generation in the animation community. The simple yet widely used approach for eyelid motion generation is to linearly interpolate two key eyelid shapes (one for the open eyelid and the other for the close eyelid) [48]. Itti *et al.* [49] incorporated blink into their avatar engine by implementing simple heuristic rules to determine the frequency of blink. Peter and O'Sullivan [50] modeled eyelid motion correlated with gaze shifts by parameterizing the vertical angle of a gaze as an input to determine the magnitude of a blink. Recently, Steptoe *et al.* [51] proposed a parametric model for lid saccades and blink based on the documented insights from ophthalmology and psychology. However, in their method, whether the parameterized eyelid motion can be soundly incorporated with gaze and head movement has not been established and validated yet.

3 System Overview

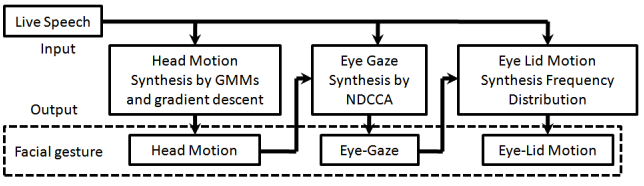

Fig. 1: Schematic overview of the proposed live speech driven head-and-eye motion generator.

Figure 1 shows the schematic overview of our proposed live speech driven head-and-eye motion generator. Basically, it consists of three inter-connected modules: a head motion synthesis module (Section 5), an eye gaze synthesis module (Section 6), and an eyelid motion synthesis module (Section 7). As illustrated in Figure 1, the live speech signal is *the shared input* among the three modules, and thus it functions as the natural motion *synchronizer* in our

approach. Also, the three modules are sequentially interconnected in an elegant way, that is, the output of the head motion synthesis module serves as one of the inputs to the eye gaze synthesis module whose output serves as one of the inputs to the eyelid motion synthesis module.

Another important employed component (but not illustrated in Figure 1) is the *data acquisition and processing* module. Traditionally, a human subject has to wear a cumbersome eye tracker (like the one used in [8]) to record accurate eye gaze movement. In this work, to overcome this limitation, we develop a novel mocap+video hybrid data acquisition system to record accurate head movement, eye gaze, and eyelid movement simultaneously (detailed in Section 4). Hence, the subtle and tight coupling among the three channels of facial gestures are well captured in our recorded dataset.

4 Data Acquisition and Processing

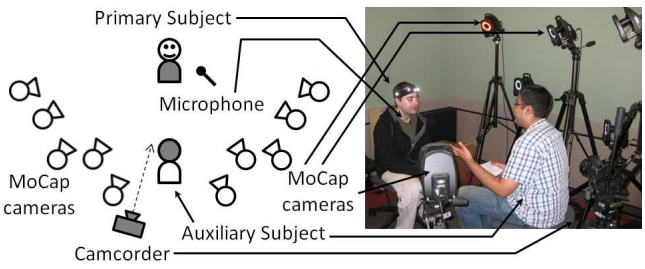

Fig. 2: **Left**: Illustration of the hybrid data acquisition setup. **Right**: A snapshot of the data acquisition process.

System setup. We use two data acquisition setups to record our needed training data. The first setup is a dyadic conversation scenario. As shown in Figure 2, in this setup, two human subjects (one is the primary subject and the other is the auxiliary subject) sit face to face, with a distance of 1.5 meters, to have a natural dyad during the mocap procedure. A VICON optical motion capture system with ten cameras (forming a semi-circle configuration) is used to capture head motion of the primary subject. A high definition camcorder is used to record the eye movement of the primary subject (refer to Section 4.1). And, a wireless microphone is used to record the voice from the primary subject. We only capture the primary subject's motion (*i.e.,* head movement, eye gaze, and eyelid movement). All the cameras and microphone are synchronized and all the recorded data are resampled to 24 fps. Such multi-channel data synchronization is intrinsically supported by the used commercial motion capture system and its software kit. The second setup is a single subject's speaking scenario. It is designed to record a variety of high-fidelity head motion while the captured human subject speaks with different emotions. The subject is asked to speak a custom, phoneme-balanced corpus while keeping his/her head movement as naturally as possible.

In total, we captured 7 minutes data using the first setup (called the "Dyad-Dataset") and 40 minutes data using the second setup (called the "Single-Dataset"). In this work, both the datasets are used for head motion synthesis, and only the Dyad-Dataset is used for eye gaze and eyelid motion synthesis.

Head motion extraction. To extract head motion, we put four markers on the head of the captured subject (the top-right panel in Figure 3). After the motion data is recorded, three Euler angles of the head rotation in each frame are calculated based on the four head markers. We denote the three Euler angles at time t as α_t, β_t, and γ_t , which are the yaw, pitch, and roll angles, respectively.

Speech feature extraction. We use the OpenSmile toolkit [52] to extract pitch and loudness features from the recorded speech. In this process, its sampling rate is set to 24Hz, and its sampling window is set to 0.025 second. Finally, we apply a contour smoothing filter with a window size of 3 frames (0.125 second) to obtain the final pitch and loudness contours. The pitch and loudness features at time t is denoted as p_t and l_t, respectively. Note that which speech features are optimal to drive facial animation has been a widely open research question for decades, as pointed out by Brand [21]. In this work, we choose empirically the combination of pitch and loudness to drive head-and-eye motion.

4.1 Eye Movement Data Acquisition

To accurately track gaze and eyelid motion in this work, we use a Canon EOS t2i with a 50mm f/1.8 lens to record HD video from the face region of the subject. Synchronized with the used optical motion capture system, this camcorder outputs FullHD video at a resolution of 1920×1280 pixels with a frame rate of 29.97 fps. The aperture is set to f/4 to achieve enough depth of field, and the shutter speed is set to 1/60s to remove motion blur. It is put at an approximate distance of 2 meters to the primary mocap subject, as illustrated in Figure 2.

Our eye movement tracking algorithm is based on the assumption that human skull is a rigid object and the center of each eyeball is located on the orbit. Thus, the centers of both the eyeballs can be calculated based on the 3D location and orientation of the head, which can guide our eye tracking algorithm to identify the exact locations of the pupils. Our eye movement tracking process consists of three main steps: camera calibration (build a mapping from the motion capture coordinate system to the camcorder coordinate system), subject calibration (determine the centers of the eyeballs and the straight direction of view), and eye tracking (employ image processing algorithms to extract the image regions containing the pupils and compute gaze and eyelid openness).

Camera calibration. We use the Direct Linear Transformation (DLT) algorithm [53] to build a mapping from the 3D space of the motion capture system to the 2D space of the video camcorder. A L-shape like pattern attached with 16

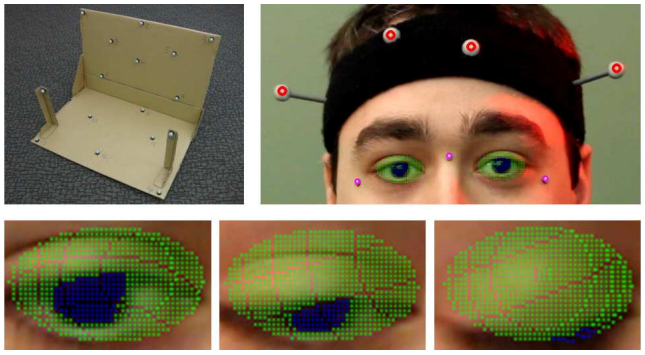

Fig. 3: Eye tracking process. **Top left**: The used camera calibration pattern. **Top right**: The used marker layout in the subject calibration step. **Bottom**: Examples of eye tracking results. Green dots indicate the eye probe regions and blue dots indicate the pupils.

markers (shown in the top-left panel of Figure 3) is captured by both the camera systems. The 3D coordinate of the ith marker in the motion capture coordinate system is denoted as (x_i, y_i, z_i), and its corresponding 2D coordinate (in the video camcorder coordinate system), denoted as (u_i, v_i), is manually annotated in the synchronized video frame. Then, this coordinate system mapping can be modeled by a 3 by 4 matrix $\mathbf{M}$, and the lens distortion is modeled using the second order Brown's distortion model [54]. Finally, by solving a system of linear equations where the unknowns are elements of matrix $\mathbf{M}$ and distortion coefficients, we essentially build a mapping from any 3D point in the motion capture coordinate system to its corresponding 2D point in the camcorder coordinate system.

Subject calibration. In this step, we ask the subject to have a straight look at the camcorder so that we can determine the exact 3D head transformation that corresponds to the straight direction of view. We also put 3 markers on the face to determine the centers of the eyeballs (the top-right panel of Figure 3). The 3 facial markers are carefully configured so that the centers of the eyeballs are at the mid points between two of them. However, due to the unavoidable human errors, the facial markers cannot be placed precisely; thus, we still need to do a correction to achieve the exact locations. We manually rectify the 2D positions of the eyeballs' centers in the first video frame by selecting their correct positions and computing the deviations. The obtained deviations are stored and used for other consecutive frames.

Eye tracking. We put an ellipse-shaped sampling grid on the sphere centered at each eyeball. Since the average eyeball diameter of an adult is about 25mm and the size of the eye is 20×10mm on average [55], we set the radius of the sphere as 25mm and the size of the grid as 20×10mm accordingly. Then, each 3D point in this grid is mapped to its corresponding 2D point in the camcorder video frame, and its color is extracted and thresholded. After that, we assign 1 to the grid points whose pixel colors are close to black and brown (*i.e.*, belonging to the pupils); otherwise, assign 0 to them. Then, an open-close operation is performed in the obtained binary result to remove the hole within the pupil region. Example results are shown in the bottom panel of Figure 3. Finally, in each tracked frame, two eye gaze rotational angles are calculated through a spherical coordinate transformation by assuming the tracked pupil center is located on the surface of the 3D eyeball, and thus the two rotation angles in one frame represent a gaze in this work. The eyelid openness is calculated based on the highest point in the pupil region and further normalized to 0∼1 (0 for "fully closed", 1 for "fully open"). Lastly, we average the tracking results of the two eyes.

5 Speech-Driven Head Motion Synthesis

A proven scheme for live speech driven gesture motion synthesis is to build optimal selection policies that choose suitable motion units from a motion library at runtime [13]. However, it cannot be applied directly to real-time head motion synthesis since a syntactic motion unit needs to have a minimally effective length (*i.e.*, number of frames). In the case of real-time head motion synthesis, this will cause a noticeable delay in synchronization (*i.e.*, out of sync) between the live speech and synthesized head motion. The alternative would be to predict head motion way ahead of the live speech, which is even more technically difficult if not impossible.

In this work, we synthesize speech-driven head motion by modeling the distribution between head motion and prosody features. Theoretically, we can learn statistical models (e.g., HMMs or Gaussian Mixtures) to model the distribution and inter-correlation among multiple high dimensional observation datasets such as head motion features and prosody features. However, in reality the training dataset is often limited, so training such a model would suffer from the curse of dimensionality. To tackle this problem, we employ a divide-and-conquer learning strategy to model the correlation between head motion and speech features [56]. Specifically, in this work, we construct a simple Gaussian Mixture model to learn the distribution probabilities between each kinematic feature of head motion (e.g., Euler angles, angular velocities, etc) and prosody features of speech and then combine these models during the runtime synthesis.

Our speech-driven head motion synthesis algorithm consists of two main steps, as shown in Figure 4: (1) At the model training step, Gaussian Mixture Models (GMMs) are trained to capture the cross-channel correlations between the speech features and the kinematic features of the head motion at every frame of the training data. (2) At the runtime synthesis step, an efficient gradient descent search is employed to solve a constraint based optimization problem, which finds the optimal head pose for the current

time frame to maximize the posterior joint distribution probability. It is noteworthy that, our head motion synthesis algorithm does not depend on a motion library at runtime. Instead, it works at frame level, which means the algorithm directly generates the head pose for the current time frame from the live speech.

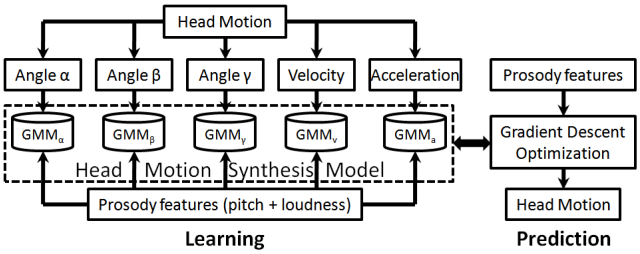

Fig. 4: Schematic view of the speech-driven head motion synthesis module. At the learning stage, the joint probability density functions (PDFs) of kinematic parameters and prosody features are modeled as Gaussian Mixture Models (GMMs). Then, at the runtime prediction step, head motion is dynamically synthesized using the gradient descent optimization method to maximize the likelihood of the joint PDFs, given the inputted prosody features.

5.1 Model Training

Some basic kinematic features of head motion are used in our model. In particular, for every time frame t in the training data, we extract its corresponding five kinematic motion features: three Euler angles α_t, β_t, and γ_t, the angular velocity v_t, and the angular acceleration a_t. The first three parameters (α_t, β_t, and γ_t) are directly extracted from the recorded head motion (refer to Section 4), and the other two parameters are derived from the three Euler angles (Eq. 1).

$$\begin{aligned} v_t &= \left\| \begin{bmatrix} \alpha_t \\ \beta_t \\ \gamma_t \end{bmatrix} - \begin{bmatrix} \alpha_{t-1} \\ \beta_{t-1} \\ \gamma_{t-1} \end{bmatrix} \right\| \\ a_t &= \left\| \begin{bmatrix} \alpha_t \\ \beta_t \\ \gamma_t \end{bmatrix} - 2 \begin{bmatrix} \alpha_{t-1} \\ \beta_{t-1} \\ \gamma_{t-1} \end{bmatrix} + \begin{bmatrix} \alpha_{t-2} \\ \beta_{t-2} \\ \gamma_{t-2} \end{bmatrix} \right\| \end{aligned} \quad (1)$$

Then, for each kinematic parameter $\kappa \in \{\alpha, \beta, \gamma, v, a\}$, we obtain a training set of tuples $\{(\kappa_t, p_t, l_t)\}_{t=1}^{n}$, where p_t and l_t are the pitch and loudness of the speech at time frame t in the training data, and n is the total number of frames in the training data. After that, based on the $\{(\kappa_t, p_t, l_t)\}_{t=1}^{n}$, we train a Gaussian Mixture Model GMM_κ to model the cross-channel probabilistic association between the kinematic parameter κ and the speech features. Since the three features are continuous, the cross-channel probabilistic distributions are modeled using the following function:

$$P(\mathbf{X}) = \sum_{i=1}^{m} c_i \frac{1}{\sqrt{(2\pi)^3 \left|\mathbf{\Sigma}_i\right|}} e^{-\frac{1}{2}(\mathbf{X}-\mu_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{X}-\mu_i)} \quad (2)$$

Here $\mathbf{X} = (\kappa, p, l)^\mathsf{T}$, m is the number of mixtures, c_i, μ_i, and $\mathbf{\Sigma}_i$ are the weight, mean, and covariance matrix of the ith mixture, respectively. We use the widely-used Expectation-Maximization (EM) algorithm to train the above GMMs model. Figure 5 shows the log-likelihoods of the GMMs with different numbers of mixtures. Based on this figure, the number of mixtures for all the GMMs is experimentally set to 10 in this work.

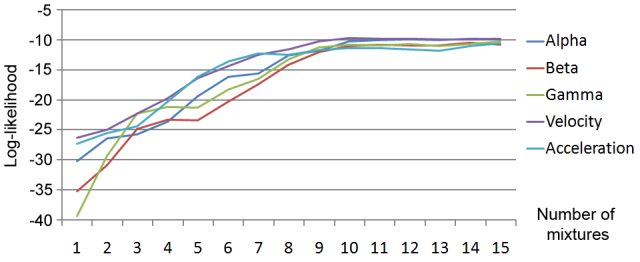

Fig. 5: The log-likelihoods of the trained GMMs with different numbers of mixtures. Note that all the training data are used to determine the number of mixture components.

5.2 Synthesis by Gradient Descent Search

After the above five GMMs models (GMM_κ, $\kappa \in \{\alpha, \beta, \gamma, v, a\}$) are trained, the problem of real-time head motion synthesis can be formulated as finding an optimal head pose for the current time step t, given the previous and current speech features (1 to t) and previous head poses (1 to $t-1$). Given the speech feature (p_t, l_t) at time step t, we can solve for its optimal head motion kinematic features by maximizing the following posterior probability via the Bayesian inference:

$$\begin{aligned} (\alpha_t^*, \beta_t^*, \gamma_t^*) = \arg\max_{\alpha_t, \beta_t, \gamma_t} \prod_{\kappa \in \{\alpha, \beta, \gamma, v, a\}} P(\kappa_t | p_t, l_t) = \\ \arg\max_{\alpha_t, \beta_t, \gamma_t} \prod_{\kappa \in \{\alpha, \beta, \gamma, v, a\}} \frac{P(\kappa_t, p_t, l_t)}{P(p_t, l_t)} \\ \text{Subject to Eq. (1)} \end{aligned} \quad (3)$$

Note that, in the above Eq. 3, $P(p_t, l_t)$ is a constant for the current time step t, and thus, this term can be ignored. However, computing the angular velocity v and angular acceleration a at time step t depends on (i.e., constrained to) the Euler angles at time steps $t-2$, $t-1$, and t (see Eq. 1). We can remove these constraints by plugging Eq. 1 into Eq. 3 and yield the non-constrained optimization as follow:

$$(\alpha_t^*, \beta_t^*, \gamma_t^*) = \arg \max_{\alpha_t, \beta_t, \gamma_t} \prod_{\kappa \in \{\alpha, \beta, \gamma\}} P(\kappa_t, p_t, l_t) \times P\left(\left\| \begin{bmatrix} \alpha_t \\ \beta_t \\ \gamma_t \end{bmatrix} - \begin{bmatrix} \alpha_{t-1}^* \\ \beta_{t-1}^* \\ \gamma_{t-1}^* \end{bmatrix} \right\|, p_t, l_t \right) \times P\left(\left\| \begin{bmatrix} \alpha_t \\ \beta_t \\ \gamma_t \end{bmatrix} - 2 \begin{bmatrix} \alpha_{t-1}^* \\ \beta_{t-1}^* \\ \gamma_{t-1}^* \end{bmatrix} + \begin{bmatrix} \alpha_{t-2}^* \\ \beta_{t-2}^* \\ \gamma_{t-2}^* \end{bmatrix} \right\|, p_t, l_t \right) \quad (4)$$

Then, we adapt the Gradient Descent search technique to find the optimal tuple $(\alpha_t^*, \beta_t^*, \gamma_t^*)$ by maximizing the above Eq. 4, assuming $(\alpha_{t-1}^*, \beta_{t-1}^*, \gamma_{t-1}^*)$ and $(\alpha_{t-2}^*, \beta_{t-2}^*, \gamma_{t-2}^*)$ are given (i.e., synthesized in previous time steps). Note that, the $P(.)$ function in Eq. 4 is not always continuous and differentiable with respect to $[\alpha_t \ \beta_t \ \gamma_t]^T$; however, it is only non-differentiable at one particular point, and even this happens, in practice, the gradient descent search would immediately bring the solution to a differentiable position at the next step.

Since natural head motion is continuous, the head pose at t is expected to be reasonably close to the head pose at $t-1$. We randomly generate an initial (starting) solution for the Gradient Descent search from the following normal distribution (Eq. 5), where the standard deviation parameter σ is computed from the feature distribution in the training data.

$$(\alpha_t, \beta_t, \gamma_t) = (\alpha_{t-1}^*, \beta_{t-1}^*, \gamma_{t-1}^*) + \mathbf{w} \qquad \mathbf{w} \approx \mathcal{N}(0, \sigma^2) \quad (5)$$

Note that the above gradient descent search is performed in a low, three dimensional space and the generated initial solution (guess) is reasonably close to the optimal one; thus, the optimization process can converge to the local maximum very quickly. In our experiments, we found that the optimization process typically took on average 17 iterations to reach the local maximum.

6 Eye Gaze Synthesis

It has been well documented that eye gaze has a measurable correlation with head motion and speech [2; 8]. Therefore, in order to capture the intrinsic gaze dynamics and the subtle coupling among gaze, head motion, and speech features, we design a Nonlinear Dynamic Canonical Correlation Analysis (NDCCA) scheme for the gaze synthesis task in our model.

Specifically, we first nonlinearly transform the recorded gaze, speech features, and head motion (*i.e.,* three Euler angles) to a high-dimensional feature space. Then, we perform linear Dynamic Canonical Correlation Analysis (DCCA) to model the intrinsic dynamic coupling between the different signals, inspired by the work of [45]. Since the first transformation is nonlinear, any linear operation in the transformed feature space corresponds to a nonlinear operation in the original space of gaze and head motion. Thus, our gaze synthesis model is essentially nonlinear. Note that although our gaze synthesis method is inspired by the work of [45], their approach can only model the linear coupling between gaze and head movement and the speech channel is completely ignored, which is insufficient to model the sophisticated cross-channel association among the gaze, head motion, and speech.

In follow-up sections, we first briefly review the background of nonlinear canonical correlation analysis (NCCA) (Section 6.1), and then describe our NDCCA-based gaze synthesis algorithm (Section 6.2).

6.1 Nonlinear Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is traditionally used to model the underlying correlation between two datasets [57]. Assuming $h \in \mathbb{R}^p$ and $e \in \mathbb{R}^q$ are two multi-dimensional observations, in traditional CCA, we try to find the basis pairs $u \in \mathbb{R}^p$ and $v \in \mathbb{R}^q$ so that the correlation between the projections $y_h = u^\mathsf{T} h$ and $y_e = v^\mathsf{T} e$ is maximized. Using the Lagrange multipliers, we can reformulate this problem to a constrained optimization [58].

$$\max_{u,v} E\{(y_h y_e) + \frac{1}{2}\lambda_h(1 - y_h^2) + \frac{1}{2}\lambda_e(1 - y_e^2)\} \qquad \text{s.t. } y_h = u^\mathsf{T} h \text{ and } y_e = v^\mathsf{T} e \quad (6)$$

Where $E\{.\}$ denotes the expected value operator.

In order to bound the bases, the term $\frac{1}{2}\lambda_h(1 - y_h^2) + \frac{1}{2}\lambda_e(1 - y_e^2)$ is added to limit the variances of y_h and y_e, i.e. $E(y_h^2 = 1)$ and $E(y_e^2 = 1)$. Here, the two Lagrange multipliers λ_h and λ_e are used to control the strength of those soft constraints on the variances of y_h and y_e. Then, the constrained optimization problem (Eq. 6) can be solved by a linear neural network as proposed by Lai and Fyfe [58]. This linear neural network is trained by the following joint learning rules:

$$\Delta u = \eta h(y_e - \lambda_h y_h), \Delta\lambda_h = \eta_0(1 - y_h^2)$$
$$\Delta v = \eta e(y_h - \lambda_e y_e), \Delta\lambda_e = \eta_0(1 - y_e^2)$$

Where η and η_0 are the learning rates.

Using the above training process, we can find a set of $r = \min(p, q)$ uncorrelated base pairs where the i-th pair can be denoted as (u^i, v^i). For convenience, we can put those base vectors together to form matrices $U \in \mathbb{R}^{p \times r}$ and $V \in \mathbb{R}^{q \times r}$, where u^i is the i-th column of U and v^i is the i-th column of V. Using the set of base pairs, we can project the observations h and e into the reduced coordinates as follow:

$$\hat{h} = U^{\mathsf{T}} h \qquad \hat{e} = V^{\mathsf{T}} e \tag{7}$$

Since our final goal is synthesis, we also need to build a back mapping from the projection $\hat{e}$ to the original data e. This back mapping can be modeled as a linear regression as suggested in [26], in which the back mapping matrix $V^* \in \mathbb{R}^{r \times q}$ is solved by linear least squares as follow:

$$V^* = \arg\min_{V} \sum_{t=1}^{n} ||V^{\mathsf{T}} \hat{e}_t - e_t||^2 \quad , V \in \mathbb{R}^{r \times q} \tag{8}$$

Where n is the total number of training examples.

However, the obtained linear correlation may not be able to capture the complex nonlinear association between the two datasets. As such, we use the kernel mapping method [26] to construct a nonlinear CCA in order to establish a nonlinear coupling between the two datasets. Thus, Eq. 7 can be reformulated as follows (Eq. 9):

$$\hat{h} = U^{\mathsf{T}} k(h) \tag{9}$$

Where k is a non-linear kernel function. Note that to avoid solving the backward mapping at the runtime, we only apply the non-linear mapping function to h. The kernel function in this work is experimentally set to the following:

$$k : \mathbb{R}^p \rightarrow \mathbb{R}^p \qquad h \mapsto \phi \text{ s.t. } \phi_i = h_i^2 + h_i, \forall i = 1..p \tag{10}$$

After the original dataset is transformed to the nonlinear space, the objective function (Eq. 6) can be used again to solve the desired U and V. Then, we project the dataset into the reduced coordinates $(\hat{h}, \hat{e})$ and perform the following linear regression (Eq. 11) to solve $C_e \in \mathbb{R}^{q \times p}$, the numerical coupling matrix between $\hat{h}$ and $\hat{e}$.

$$C_e = \arg\min_{C} \sum_{t=1}^{n} ||C\hat{h}_t - \hat{e}_t||^2 \quad , C \in \mathbb{R}^{q \times p} \tag{11}$$

Where $(\hat{h}_t, \hat{e}_t)$ denotes the t-th example in the training dataset of n examples.

6.2 NDCCA-based Eye Gaze Synthesis

Now we describe how to use the above NCCA to further build a dynamic coupling model for eye gaze synthesis. As mentioned in Section 4, a head motion frame is represented as three Euler angles, and an eye gaze is represented as two rotation angles. We also use the speech loudness feature as one of the inputs to our gaze synthesis model, since the correlation between eye gaze patterns and speech loudness features has been previously reported [47].

Specifically, n denotes the total number of frames in our training data set. Both the head motion and speech loudness features in the training data are combined frame by frame to a head-speech dataset $\{h_t : t = 1..n\}$ (where $h_t \in \mathbb{R}^4$ is the head-speech frame at time step t), and the extracted eye gaze dataset is denoted as $\{e_t : t = 1..n\}$ (where $e_t \in \mathbb{R}^2$ is the eye gaze at time step t). Then, the NCCA-based coupling model (Section 6.1) is trained to build an eye gaze synthesis model that connects the eye gaze patterns with the head-speech characteristics. As such, the projection matrix U, the back projection V^*, and the coupling matrix C_e can be determined (refer to Section 6.1).

Expanding NCCA to NDCCA. The above trained NCCA model captures the cross-channel association among speech, gaze, and head movement, but the intrinsic gaze dynamics are not modeled. In this work, we choose to further add the gaze dynamic model proposed in the work of [45] to tackle this limitation. In particular, the gaze dynamics are modeled by a linear regression matrix $A^* \in \mathbb{R}^{2\times2}$. This gaze dynamic model can be solved by the linear least squares in Eq. 12, which minimizes the sum of squared errors between the gaze of two consecutive time frames t and $t-1$ projected into the reduced coordinate system.

$$A^* = \arg\min_{A} \sum_{t=1}^{n-1} ||\hat{e}_t - A\hat{e}_{t-1}||^2 \quad , A \in \mathbb{R}^{2\times2} \tag{12}$$

Where A^* is the dynamic transition matrix to be estimated. Given $\{\hat{e}_t : t = 1..n\}$, A^* can be efficiently solved [45]. To the end, we construct a NDCCA model for our eye gaze synthesis.

Gaze synthesis. Given the live speech input and the synthesized head motion at the time frame t from the head motion synthesis module (the combined input is denoted as h_t), we can synthesize the accompanying eye gaze using the above trained NDCCA model, described as follows.

The synthesis procedure is straightforward after the training process, where we determine the projection matrix U (Eq. 6), the back projection matrix V^* (Eq. 8), the coupling matrix C_e (Eq. 11), and the dynamic transition matrix A^* (Eq. 12). Basically, we first compute $\hat{e}_t$ using Eq. 13.

$$\hat{e}_t = C_e U^{\mathsf{T}} k(h_t) \tag{13}$$

Then, the dynamic transition matrix A^* is used to further smooth $\hat{e}_t$ (Eq. 14).

$$\hat{e}_t = \lambda_1 \hat{e}_t + \lambda_2 (A^* \hat{e}_{t-1}) \tag{14}$$

In the above Eq. 14, we empirically set two weight coefficients λ_1 and λ_2 to 0.7 and 0.3, respectively. Finally, the outputted eye gaze can be reconstructed using the following Eq. 15.

$$e_t = {V^*}^{\mathsf{T}} \hat{e}_t \tag{15}$$

7 Eyelid Motion Synthesis

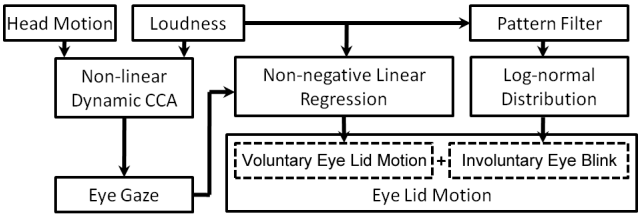

Fig. 6: Pipeline of the eyelid motion synthesis module.

There are two major categories of eyelid motion: *involuntary blink* and *voluntary eyelid motion* [51]. Involuntary eye blink is for the protection of the eyes and it is usually fully close, while voluntary eyelid motion is entwined in human psychology and gesture, *e.g.*, gaze-evoked blink [59; 50]. Inspired by this eyelid motion classification, we decompose eyelid motion into two components: involuntary eye blink and voluntary eyelid motion, and model them separately. Basically, the voluntary eyelid motion is modeled through a non-negative linear regression model, and the involuntary eye blink is explicitly described as two Log-normal distributions, constructed via data fitting. Figure 6 shows the pipeline of the eyelid motion synthesis.

7.1 Involuntary Blink Model

As reported in the literature [60], eye blink rate patterns of normal subjects follow a certain frequency distribution. Also, the blink pattern of a person generally has a strong correlation with his/her conversational role, in particular, speakers exhibit higher-frequency blink than listeners [1].

In this work, we introduce a simple yet effective eye blink generation algorithm by constructing a probabilistic distribution model as follows. First, we analyze the recorded eye blink data and construct two blink rate distribution models (one for the talking mode and the other for the listening mode). Then, in the blink synthesis process, we first determine the mode (talking or listening) based on the live speech input and then compute eye blink using its corresponding blink rate distribution model.

We experimentally found that two different Log-normal distributions (Eq. 16) can be used to soundly fit the recorded blink data in the talking and listening modes, respectively. Our empirical finding is also consistent with the previous reports by other researchers that blink distributions for both the talking and listening modes are close to the normal distribution [60].

$$\ln L \approx \mathcal{N}(\mu, \sigma^2) \tag{16}$$

In Eq. 16, L represents the blink pattern. The two parameters, μ and σ, in the above Eq. 16 are determined through data fitting: μ=21.1 and σ=3.6 for the talking mode, and μ=15.4 and σ=3.9 for the listening mode.

7.2 Gaze-Evoked Eyelid Motion Model

In the psychology literature [59], the relationship between eyelid motion and eye gaze is typically referred to as *gaze-evoked eyelid motion*. The gaze-evoked eyelid motion usually starts simultaneously with the eye gaze, and its magnitude often increases in accordance with the amplitude of the eye gaze [50]. We propose a non-negative linear regression coupling model to build the connection between eyelid motion and gaze. We choose the non-negative linear regression model since it can reduce overfitting in the learning process.

Assuming the extracted eye gaze dataset is $\{e_t : t = 1..n\}$ ($e_t \in \mathbb{R}^2$ is the the eye gaze in time frame t0) and the gaze-evoked eyelid motion dataset is $\{b_t : t = 1..n\}$; here $b_t \in \mathbb{R}$ is the acquired eyelid motion minus fully closed eye blink in time frame t. Specifically, b_t is a normalized value between 0 (closed) and 1 (open), as described in the data acquisition Section 4. We use the non-negative least-square technique [57] to learn the linear regression matrix $C_b \in \mathbb{R}^{1\times 2}$. This learning is done by minimizing the following Eq. 17.

$$C_b = \arg\min_{C} \sum_{t=1}^{n} ||C.e_t - b_t||^2 \tag{17}$$
$$\text{Subject to: } C \in \mathbb{R}^{1\times 2} \text{ and } C \geq \vec{0}$$

7.3 Eyelid Motion Synthesis

Given a synthesized eye gaze e_t at the time frame t, we can use the above constructed non-negative linear coupling model to predict the voluntary (*i.e.*, gaze-evoked) eyelid motion b_t, by computing the following linear equation (Eq. 18).

$$b_t = C_b e_t \tag{18}$$

Also, we can use the above constructed Log-normal probability distribution models to generate the involuntary blink l_t by sampling blink rates according to the speech loudness feature (i.e., determine the talking or listening mode). The sampling equation for the Log-normal probability distribution is as follows (Eq. 19):

$$P(l_t \equiv blink) = e^{\mu+\sigma\mathcal{N}(0,1)} \quad (19)$$

The final synthesized eyelid motion is the sum of both the voluntary eyelid motion b_t and the involuntary blink l_t.

8 Results and Evaluations

We tested our head-and-eye motion generator by inputting various audio clips including live speech and pre-recorded speech. We found that most of the synthesized results are realistic and perceptually believable (example snapshots are shown in Figure 7).

Fig. 7: Example snapshots of synthetic avatar head and eye movement based on live (or pre-recorded) speech input by our approach.

Runtime performance. Our live speech driven head-and-eye motion generator is highly efficient in terms of runtime performance. We measured the average running time of our system (implemented in C++ without GPU-acceleration) on an off-the-shelf desktop computer (configuration: Intel Quad Core 2.8GHz, 4GB RAM, Windows 7 OS). The used 3D avatar model has 30K triangles. The average running time for each frame synthesis in our system is about 21 ms, which is competent to ensure the whole system to run at 24 fps (the minimum real-time frame rate). We also found that within the measured 21 ms (one cycle), our motion synthesis module took significantly less computing time than other supporting modules (*e.g.*, mesh deformation and rendering).

8.1 Comparative User Studies

In order to evaluate the effectiveness of our approach, we conducted three comparative user studies: i) a head motion comparative study, ii) an eye motion comparative study, and iii) a facial gesture comparative study. We employed the well-established paired comparison methodology [61] for the three user studies. Basically, we used various methods (include ours) to synthesize animation clips and then asked participants to compare them in a paired way. Although our proposed method can perform online, we still used pre-recorded speech clips for the user studies due to their repeatability (so the same pre-recorded speech clips can be inputted to various synthesis methods for result comparison). And, instead of asking the participants to explicitly rate the naturalness of individual facial animation clips, the participants were only asked to select the more natural one between the two clips (forming a comparison pair), which makes the decision much easier and thus increases the accuracy and robustness of the subjective evaluation outcomes.

In all the user studies, the same 3D avatar model and the same resolution of 320*320 pixels were used in all the clips. A total of 20 student volunteers participate in all the three comparative studies. Each comparison pair was displayed on a 1280x1024 LCD monitor with a 0.6m distance away from the participants. Between the two animation clips in each pair, the participants were asked to select the one which appeared to be more natural and perceptually realistic for them. They were allowed to leave undecided choices if they cannot decide which clip is perceptually better (that is, perceptually indistinguishable for them). To counter balance the order of the visual stimuli, the comparison pairs were displayed in a random order for each participant. The participants can choose to play the two animation clips in a pair one after another (do not need to play them simultaneously) and they can view the clips for unlimited times before made their decisions. Figure 8 shows a snapshot of a side-by-side comparison pair used in the evaluation.

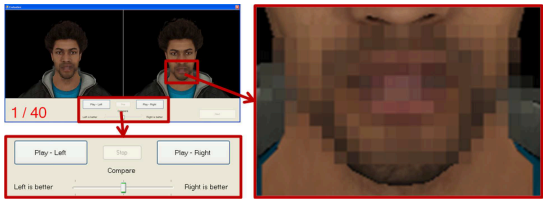

Fig. 8: A snapshot of the user study interface. The command buttons and the mosaiced avatar mouth region are magnified.

8.1.1 Head Motion Comparative Study

In this study, we first manually segmented the first minute of the test speech dataset published by Levine *et al.* [12] to 10 short audio clips. The duration of each audio clip was 5-10 seconds. Then, based on the 10 test audio clips, we generated a total of 50 head animation clips (with audio) using five different approaches: 10 by the prosody-driven body language synthesis algorithm by Levine et al. '09 [12], 10 by the Mood-Swings approach by Chuang et al. '05 [11], 10 by the HMMs-based approach by Busso et al. '05 [10], 10 by using the captured (ground-truth) head motion, and 10 by our approach (this work). The synthesized animations by [12] were directly extracted from their published results. In our implementation of the Mood-Swings framework [11], we first manually segmented the training data to 643 sentences and then further segmented the sentences to 8135 pitch segments. The path search in the Mood-Swings framework was implemented by retaining the top 10 pitch matches only. In our implementation of the HMMs-based approach [10], we quantized the head motion Euler angles to 16 clusters and used them to train the same

number of HMMs. Based on the 50 head animation clips, we produced 40 side-by-side comparison pairs: 10 pairs for each comparison case, that is, 10 pairs for the comparison between our approach and each of the other four approaches (i.e., [11; 10; 12] and the ground-truth).

#	u	χ^2	p value	**Motion Capture**	**Levine et al. '09**	**Busso et al. '05**	**Chuang et al. '05**
1	0.382	49.697	<0.001	9/10	8/10	11/7	13/7
2	0.305	40.799	<0.001	5/2	11/7	12/7	11/8
3	0.347	45.599	<0.001	5/11	9/7	14/6	11/8
4	0.375	48.799	<0.001	8/10	9/7	12/7	12/6
5	0.536	67.197	<0.001	10/9	8/10	14/6	11/9
6	0.543	64.199	<0.001	7/10	8/10	15/4	12/7
7	0.431	55.197	<0.001	8/9	11/9	12/6	11/8
8	0.619	76.000	<0.001	8/9	10/7	12/7	14/5
9	0.880	106.400	<0.001	10/8	12/7	12/3	14/5
10	0.108	26.419	<0.01	4/11	8/4	8/10	8/11

TABLE 1: Consistency and agreement test statistics for the head motion comparative evaluations. The number pair (e.g., X/Y) shown in each cell of the right part of the table denotes that the total number of the participants who voted for our approach is X and the total number of the participants who voted for the other comparative approach is Y. The four comparative approaches are: transferring original motion capture data, Levine et al. '09 [12], Busso et al. '05 [10], and Chuang et al. '05 [11].

To eliminate any potential perceptual influence of eye motion, we did not add eye motion to any of the animation clips (*i.e.,* the eyes stayed still, looking straight ahead). Also, for each test audio clip, we generated its lip-sync motion using an ad hoc approach, and the same lip-sync approach was used by all the above five different head motion generation approaches. Moreover, we intentionally applied a strong mosaic filter to the mouth region in all the clips in order to reduce potential perceptual distractions from the non-perfect, synthesized lip-sync quality (refer to Figure 8). It is noteworthy that in this comparison, we chose to compare our approach with the Levine et al. 09's work [12], not their latest work [13] due to the following reasons. First, the work of [13] primarily focuses on hand gesture synthesis and its algorithm even does not consider any head motion features, while the work of [12] explicitly includes head motion kinematic features. Second, the dataset used in [13] is not publicly accessible at this point, while the dataset of [12] is publicly accessible, which makes a fair comparison possible.

Agreement test of the subjective evaluation outcomes: We carried out a statistical test, *agreement test* [61], to see whether the participants rated all the pairs in a consistent way in our comparative user studies. Basically, the agreement test is used to indicate the overall agreement among all the votes in paired comparisons, which has been widely used in paired comparison studies as an appropriate measure [61; 45]. Specifically, in this work, we choose to calculate the Coefficient of Agreement (COA) [62] to demonstrate that there is certain consistency among users votes in our user study; otherwise, if there's no such consistency, we can interpret that the obtained user study results are somehow contaminated (e.g., voting randomness) and thus invalid. In this case, we cannot perform further analysis on such study results. Also, we further use the Chi-Square test statistics (χ^2) to compute the statistical significance of the COA [63].

Here we take #1 pair in the head motion comparative study as an example: As shown in the first row of Table 1, among the total 20 participants, 9 participants rated our approach better than the original motion capture data, 8 rated ours better than [12], 11 rated ours better than [10], and 13 rated ours better than [11]. Following the test method proposed by Kendall and Babington-Smith [62], the COA of sample #1 for "head motion realism comparison" is $u = 0.382$, its χ^2 is 49.697, and its corresponding p-value is 0.001 (1% χ^2 is 29.59) given DOF = 10 (C_2^5). Thus, our Chi-Square test statistics results indicate that for the used head animation #1 pair there was a statistically strong agreement among the participants. The COA test results of all the head animation pairs are shown in Table 1. It is noteworthy that in Table 1 (follow-up Tables 2 and 3 have the same problem), the numbers X and Y in one entry may not always sum to 20 (the total number of the participants in our user studies) because the participants in our user studies were allowed to leave undecided choices if they cannot decide which clip is perceptually better (that is, perceptually indistinguishable for them).

8.1.2 Eye Motion Comparative Study

We randomly selected 10 test audio clips from the captured motion dataset (described in Section 4) and the 10 test clips were not used in the model training stage. Based on the 10 test audio clips and their corresponding ground-truth (recorded) head motion, we generated a total of 40 head+eye animation clips (with audio) using the following four different approaches: 10 by the linear head-eye coupling algorithm by Ma and Deng '09 [45], 10 by the "eyes alive" model by Lee et al. '02 [8], 10 by using the original captured eye motion, and 10 by our approach. In our implementation of the "eyes alive" model [8], we used the empirical rules and parameters presented in their work to generate the eye gaze. In our implementation of the linear head-eye coupling algorithm [45], we used the same dataset as in our approach to train its statistical model. We produced 30 side-by-side comparison pairs: 10 pairs for each comparison case, that is, 10 pairs for the comparison between our eye motion synthesis algorithm and each of the other three eye motion generation approaches. It is noteworthy that since our approach is only driven by intuitive inputs such as speech and head motion, in this comparative study, we did not compare our approach with the attention-driven gaze and blink model [50] due to the difficulty of a fair comparison.

For each test audio clip, we essentially had 4 animation clips (one for each eye motion generation approaches, see

above). The 4 animation clips used the same pre-recorded (ground-truth) head motion and the same synthesized lip-sync motion. Like in the above head motion comparative study, we applied the same mosaic filter to the mouth region in all the clips due to the same reason. Note that two chosen approaches [8; 45] cannot generate eyelid motion or blink; as such, we kept their original forms and did not add eyelid motion to their corresponding 20 (=10x2) clips. We carried out the same COA statistical test as in the above head motion comparative study. The COA test results of all the pairs in the eye motion comparative study are shown in Table 2.

#	u	χ^2	p value	**Motion Capture**	**Lee et al. '02**	**Ma and Deng '09**
1	0.177	65.599	<0.001	10/4	18/2	11/6
2	0.211	75.199	<0.001	10/7	17/3	16/3
3	0.101	44.000	<0.001	2/17	12/7	13/5
4	0.119	49.000	<0.001	4/1	8/0	17/2
5	0.158	60.197	<0.001	8/1	16/2	13/5
6	0.148	57.199	<0.001	6/14	15/1	14/5
7	0.216	76.599	<0.001	9/10	17/2	17/2
8	0.105	45.200	<0.001	3/16	8/1	16/3
9	0.065	33.600	<0.001	5/2	11/7	8/5
10	0.080	38.000	<0.001	7/2	8/10	12/4

TABLE 2: Consistency and agreement test statistics for the eye motion comparative evaluations. The number pair (e.g., X/Y) shown in each cell of the right part of the table denotes that the total number of the participants who voted for our approach is X and the total number of the participants who voted for the other comparative approach is Y. The three comparative approaches are: transferring original motion capture data, Lee et al. '02 [8], and Ma and Deng '09 [45].

8.1.3 Facial Gesture Comparative Study

Based on the same 10 test audio clips (i.e., those used in the above eye motion comparative study), we generated a total of 20 facial gesture (head, gaze and eyelid) clips with audio using two different approaches: 10 by our facial gesture generator and the other 10 by using the original captured head and eye motion. Similarly, we applied the same mosaic filter to the mouth region in all the clips. Based on the resultant 20 animation clips, we produced 10 side-by-side facial gesture comparison pairs. The COA test results of all the pairs in the facial gesture comparative study are shown in Table 3.

8.2 Analysis of User Study Results

Through the above COA tests, we validated that the rating results of our paired comparisons were statistically valid. Thus, based on the obtained user ratings, we summed up the user selections in each comparative user study. The results are summarized and illustrated in Figure 9. To quantify the statistical significance of the results, we also performed a two-tailed independent one-sample t-test and reported the p-value for each row in Figure 9.

#	u	χ^2	p value	**Motion Capture**
1	0.378	8.200	<0.05	9/9
2	0.221	5.200	<0.05	7/8
3	0.221	5.200	<0.05	7/10
4	0.221	5.200	<0.05	7/11
5	0.221	5.200	<0.05	7/5
6	0.221	5.200	<0.05	7/11
7	0.378	8.200	<0.05	9/9
8	0.378	8.200	<0.05	9/10
9	0.473	10.000	<0.01	10/9
10	0.157	4.000	<0.05	6/11

TABLE 3: Consistency and agreement test statistics for the facial gesture comparative evaluations. The number pair (e.g., X/Y) shown in each cell of the right part of the table denotes that the total number of the participants who voted for our approach is X and the total number of the participants who voted for the original motion capture data is Y.

In the head motion comparative study, there were 96 ratings from the participants who voted our approach over [12], while 78 ratings from them who voted [12] over ours. As clearly illustrated in Figure 9, our approach gained more preferred ratings from the participants than the other three algorithms (excluding the original captured motion). Meanwhile, our approach had fewer preferred ratings (11 fewer votes) than the original captured head motion, but the difference is statistically insignificant (p-value is 0.347).

In the eye motion comparative study, the eye animation results by our approach gained significantly more votes than the other two algorithms [8; 45]. This is not surprising, since the two algorithms [8; 45] did not model eyelid motion, given the fact that eyelid motion is one of the vital components of realistic avatars [51]. In our experiments, our proposed eye motion method outperformed the other two methods [8; 45]; however, it is not completely clear which factor(s) (i.e., gaze, eyelid, or combined) contribute to the outcome difference, since our proposed method models two factors (gaze and eyelid) simultaneously while the other two only consider one factor (gaze). In addition, our approach gained slightly fewer votes (6 fewer votes) than the original captured eye motion, and a significant portion of votes were "undecided choices" (in other words, difficult to make a pick for some of the participants), which is additional evidence to support that the eye motion results by our approach are reasonably close to the captured (ground-truth) eye motion.

In the overall facial gesture (head, gaze and eyelid) comparative study, our approach gained fewer votes (15 fewer votes) than the original captured motion (including eye and head motion). Although the voting difference is still statistically insignificant (p-value is 0.187), which indicates that the synthetic facial gestures generated by our approach

were reasonably close to the captured motion; however, they cannot completely fool the participants (*i.e.*, cannot tell the difference between the synthesized motion and the ground-truth ones), and there are some room to further improve our algorithms.

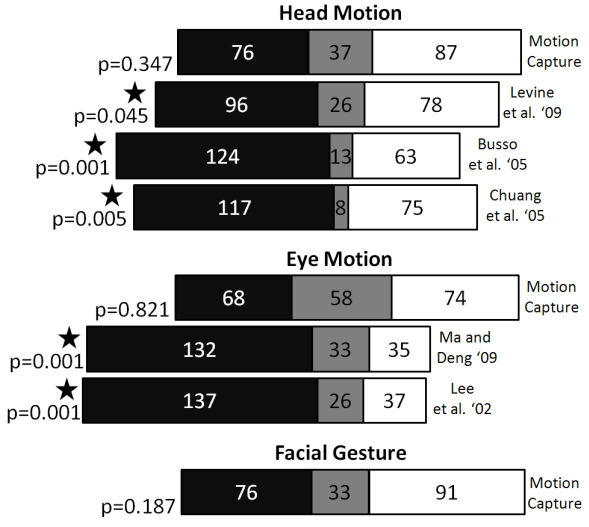

Fig. 9: The subjective evaluation results of our three comparative user studies. Black boxes at the left side indicate the total number of times when the participants voted the results by our algorithm over those by the other approach (in a pair). White boxes at the right side indicate the total number of times when the participants voted the results by the other approach over those by our algorithm. Gray boxes in the middle indicate "undecided choices" (*i.e.*, perceptually equivalent). The symbol $\star$ indicates the computed statistical significance according to a two-tailed independent one-sample t-test with p-value < 0.05.

9 Discussion and Conclusions

In this paper, we propose a novel, fully automated framework to generate realistic head motion, eye gaze, and eyelid motion simultaneously based on live or recorded speech input. Its core idea is to learn separate yet inter-related statistical models for each component (head motion, gaze, or eyelid motion) from a pre-recorded facial motion dataset. In this way, the synthesized facial gestures are intrinsically synchronized, that is, the speech signal serves as the shared control signal across different algorithm modules. In addition to our contribution in automated head-and-eye motion generation, we also design a novel mocap+video hybrid data acquisition system to simultaneously record and automatically extract accurate head movement, gaze, and eyelid motion. We believe that, not limited to this work, such a high-fidelity, multi-channel data acquisition approach can be potentially used in numerous virtual human modeling and animation, affective computing, virtual reality, and HCI applications.

Since our framework can synthesize realistic head-and-eye motion from live speech, it can have numerous potential applications in entertainment, virtual worlds, HCI, etc. By dynamically synthesizing facial gestures at frame level, our approach does not need to use a pre-constructed motion library (by contrast, a motion library is needed in [13]); hence, our efficient runtime motion synthesis module has a very small size, which makes it possible to deploy our approach to modern smartphones that have an ever-increasingly powerful computing capability these days.

In the future, we plan to explore the possibility of extending the current framework for the synthesis of other facial motion (*e.g.*, lip-sync) and hand gesture. As the first complete computational model for the coordinated generation of head motion, eye gaze, and eyelid motion, several limitations still exist in the current work.

- First, our current approach does not consider the semantics enclosed in the input speech; thus, certain meaningful non-verbal facial gestures may not be accurately generated. For example, in many cultures, when people agree to something, they tend to nod their heads. However, extracting and incorporating such semantic information from speech (*e.g.*, certain recognized keywords), via advanced speech recognition and language processing techniques, to improve our current approach may be a feasible solution, since the feasibility of such a paradigm has been demonstrated in the *offline* synthesis case of the speech-driven body language controller [13]. However, this direction would face many challenges since extracting contextual information from live speech is a non-trivial task, and it may need to introduce a certain amount of delay to the process.
- Second, different from various global (with respect to the whole utterance) optimization schemes extensively used in previous efforts, our current approach only considers and utilizes prior and current information available in the live speech to generate realistic facial gestures in real-time, and thus it does not take advantage of forthcoming information in the speech. However, such a non-global optimization strategy may not be the optimal solution to offline facial motion synthesis driven by pre-recorded speech. Also, velocity and acceleration only add the temporal information in a short period of time and thus the suprasegmental relationship between speech and gestures may not be fully captured in the current model. However, capturing suprasegmental relationship between speech and head motion is very challenging and it is definitely a good direction for future exploration.
- Third, the speaker-independent factor is not investigated in our current work, although we show that our approach can generate plausible head-and-eye motion even if the input speeches are from different human subjects (i.e., not the training/captured subject), such as the audio clips of different film actors that are chosen to drive animations in the enclosed demo video. However, our current work does not have a

proven methodology basis to conclude its true speaker-independence. In the future, we would like to investigate a systematic solution to this issue such as properly normalizing speech features.

ACKNOWLEDGMENTS

This work is supported in part by NSF IIS-0914965, Texas NHARP 003652-0058-2007, and research gifts from Google and Nokia. We also would like to thank the RocketBox Libraries for providing the used high-quality 3D avatar, and thank Nikhil Navkar, Richard Gilliam, and Joy Nash for motion capture helps. 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] I. Albrecht, J. Haber, and H.-P. Seidel, "Automatic generation of non-verbal facial expressions from speech," in *Proc. of CGI'02*, 2002, pp. 283–293.

[2] S. Chopra-Khullar and N. I. Badler, "Where to look? automating attending behaviors of virtual human characters," in *AGENTS '99: Proceedings of the third annual conference on Autonomous Agents*, 1999, pp. 16–23.

[3] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: Facial movements accompanying speech," in *Proc. of IEEE Conf. on Automatic Face and Gesture Recognition (FGR)*, 2002, p. 396.

[4] V. Vinayagamoorthy, M. Garau, A. Steed, and M. Slater, "An eye gaze model for dyadic interaction in an immersive virtual environment: Practice and experience," *Computer Graphics Forum*, vol. 23, no. 1, pp. 1–11, 2004.

[5] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," *Psychological Science*, vol. 15, no. 2, pp. 133–137, 2004.

[6] E. Gu and N. I. Badler, "Visual attention and eye gaze during multiparty conversations with distractions," in *Proceeding of International Conference on Intelligent Virtual Agents 2006*, 2006, pp. 193–204.

[7] S. Masuko and J. Hoshino, "Head-eye animation corresponding to a conversation for CG characters," *Computer Graphics Forum*, vol. 26, no. 3, pp. 303–312, 2007.

[8] S. P. Lee, J. B. Badler, and N. I. Badler, "Eyes alive," in *SIGGRAPH '02*, 2002, pp. 637–644.

[9] Z. Deng, J. P. Lewis, and U. Neumann, "Automated eye motion using texture synthesis," *IEEE Comput. Graph. Appl.*, vol. 25, no. 2, pp. 24–30, 2005.

[10] C. Busso, Z. Deng, U. Neumann, and S. Narayanan, "Natural head motion synthesis driven by acoustic prosodic features: Virtual humans and social agents," *Comput. Animat. Virtual Worlds*, vol. 16, no. 3-4, pp. 283–290, 2005.

[11] E. Chuang and C. Bregler, "Mood swings: expressive speech animation," *ACM Trans. Graph.*, vol. 24, no. 2, pp. 331–347, 2005.

[12] S. Levine, C. Theobalt, and V. Koltun, "Real-time prosody-driven synthesis of body language," *ACM Trans. Graph.*, vol. 28, pp. 172:1–172:10, December 2009.

[13] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun, "Gesture controllers," *ACM Trans. Graph.*, vol. 29, pp. 124:1–124:11, July 2010.

[14] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, "Synthesizing realistic facial expressions from photographs," *Proc. of ACM SIGGRAPH'98*, vol. 32, pp. 75–84, 1998.

[15] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. of ACM SIGGRAPH'99*, 1999, pp. 187–194.

[16] Y. Lee, D. Terzopoulos, and K. Waters, "Realistic modeling for facial animation," in *Proc. of ACM SIGGRAPH'95*, 1995, pp. 55–62.

[17] 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. Graph.*, vol. 25, no. 3, pp. 1013–1024, 2006.

[18] 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. Graph.*, vol. 27, no. 5, pp. 1–10, 2008.

[19] T. Weise, H. Li, L. Van Gool, and M. Pauly, "Face/off: live facial puppetry," in *SCA'09*. New York, NY, USA: ACM, 2009, pp. 7–16.

[20] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: driving visual speech with audio," in *SIGGRAPH '97*, 1997, pp. 353–360.

[21] M. Brand, "Voice puppetry," in *SIGGRAPH '99*, 1999, pp. 21–28.

[22] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *SIGGRAPH '02*, 2002, pp. 388–398.

[23] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz, "Spacetime faces: High-resolution capture for modeling and animation," *ACM Trans. on Graph.*, vol. 23, no. 3, pp. 548–558, 2004.

[24] E. Sifakis, I. Neverov, and R. Fedkiw, "Automatic determination of facial muscle activations from sparse motion capture marker data," *ACM Trans. on Graph.*, vol. 24, no. 3, pp. 417–425, 2005.

[25] 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. Graph.*, vol. 26, no. 3, p. 33, 2007.

[26] W.-W. Feng, B.-U. Kim, and Y. Yu, "Real-time data driven deformation using kernel canonical correlation analysis," in *ACM SIGGRAPH 2008 papers*, ser. SIGGRAPH '08, 2008, pp. 91:1–91:9.

[27] L. Williams, "Performance-driven facial animation," in *Proc. of ACM SIGGRAPH'90*, 1990, pp. 235–242.

[28] J.-Y. Noh and U. Neumann, "Expression cloning," in *Proc. of ACM SIGGRAPH'01*, 2001, pp. 277–288.

[29] R. W. Sumner and J. Popović, "Deformation transfer for triangle meshes," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 399–405, 2004.

[30] H. Li, T. Weise, and M. Pauly, "Example-based facial rigging," in *Proc. of ACM SIGGRAPH 2010*, 2010, pp. 32:1–32:6.

[31] Z. Deng and J. Y. Noh, "Computer facial animation: A survey," in *Data-Driven 3D Facial Animation*. Spring-Verlag Press, 2007.

[32] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, "Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents," in *Proc. of SIGGRAPH'94*. New York, NY, USA: ACM, 1994, pp. 413–420.

[33] C. Pelachaud, N. I. Badler, and M. Steedman, "Generating facial expressions for speech," *Cognitive Science*, vol. 20, pp. 1–46, 1994.

[34] D. DeCarlo, C. Revilla, M. Stone, and J. Venditti, "Making discourse visible: Coding and animating conversational facial displays," in *Proc. of IEEE Computer Animation'02*, 2002.

[35] X. Ma, B. H. Le, and Z. Deng, "Perceptual analysis of talking avatar head movements: a quantitative perspective," in *CHI'11: Proceedings of the 2011 annual conference on Human factors in computing systems*. New York, NY, USA: ACM, 2011, pp. 2699–2702.

[36] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 15, no. 3, pp. 1075–1086, 2007.

[37] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, pp. 1330–1345, August 2008.

[38] J. Lee and S. Marsella, "Learning a model of speaker head nods using gesture corpora," in *Proc. of AAMAS*, 2009, pp. 289–296.

[39] G. O. Hofer, "Speech-driven animation using multi-modal hidden markov models, PhD thesis," Ph.D. dissertation, University of Edinburgh, 2010.

[40] R. M. Maatman, J. Gratch, and S. Marsella, "Natural behavior of a listening agent," in *IVA: Proc. of International Conference on Intelligent Virtual Agents*, 2005, pp. 25–36.

[41] J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. J. van der Werf, and L.-P. Morency, "Virtual rapport," in *Proc. of International Conference on Intelligent Virtual Agent (IVA) 2006*, 2006, pp. 14–27.

[42] S. Masuko and J. Hoshino, "Generating head eye movement for virtual actor," *Syst. Comput. Japan*, vol. 37, pp. 33–44, November 2006.

[43] R. A. Colburn, M. F. Cohen, and S. M. Drucker, "The role of eye gaze in avatar mediated conversational interfaces," *Microsoft Research Technical Report, MSR-TR-2000-81*, 2000.

[44] M. Thiébaux, B. Lance, and S. Marsella, "Real-time expressive gaze animation for virtual humans," in *Proc. of AAMAS*, 2009, pp. 321–328.

[45] X. Ma and Z. Deng, "Natural eye motion synthesis by modeling gaze-head coupling," in *Proc. of IEEE VR '09*, 2009, pp. 143–150.

[46] O. Oyekoya, W. Steptoe, and A. Steed, "A saliency-based method of simulating visual attention in virtual scenes," in *Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology*. ACM, 2009, pp. 199–206.

[47] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt, "Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes," in *Proc. of CHI '01*, 2001, pp. 301–308.

[48] D. Bitouk and S. K. Nayar, "Creating a speech enabled avatar from a single photograph," in *Proc. of IEEE VR'08*, March 2008, pp. 107–110.

[49] L. Itti and N. Dhavale, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Proc. SPIE*, 2003, pp. 64–78.

[50] C. Peters and C. O'Sullivan, "Attention-driven eye gaze and blinking for virtual humans," in *ACM SIGGRAPH 2003 Sketches & Applications*, 2003, pp. 1–1.

[51] W. Steptoe, O. Oyekoya, and A. Steed, "Eyelid kinematics for virtual characters," *Computer Animation and Virtual Worlds*, vol. 21, no. 1, pp. 161–171, 2010.

[52] F. Eyben, M. Wollmer, and B. Schuller, "OpenEAR - introducing the munich open-source emotion and affect recognition toolkit," in *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–6.

[53] Y. Abdel-Aziz and H. Karara, "Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry." in *Proceedings of the Symposium on Close-Range Photogrammetry*, 1971, pp. 1–18.

[54] D. Brown, "Decentering distortion of lenses," *Photogrammetric Engineering*, vol. 7, pp. 444–462, 1966.

[55] P. Riordan-Eva, D. Vaughan, and T. Asbury, *General Ophthalmology*. Stanford University Press, 2004.

[56] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.

[57] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. Academic Press, 1979.

[58] P. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," in *Proc. of IEEE-INNS-ENNS Int'l Conf. on Neural Networks (IJCNN'00)*, 2000.

[59] C. Evinger, K. Manning, J. Pellegrini, M. Basso, A. Powers, and P. Sibony, "Not looking while leaping: the linkage of blinking and saccadic gaze shifts," *Experimental Brain Research*, vol. 100, no. 1, pp. 337–344, 1994.

[60] A. R. Bentivoglio, S. B. Bressman, E. Cassetta, D. Carretta, P. Tonali, and A. Albanese, "Analysis of blink rate patterns in normal subjects," *Movement Disorder*, vol. 12, no. 6, pp. 1028–1034, 1997.

[61] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, "Evaluation of tone mapping operators using a high dynamic range display," in *SIGGRAPH '05: ACM SIGGRAPH 2005 Papers*, New York, NY, USA, 2005, pp. 640–648.

[62] M. G. Kendall and B. Babington-Smith, "On the method of paired comparisons," *Biometrika*, vol. 31, pp. 324–345, 1940.

[63] S. Siegel, *Nonparametric statistics for the behavioral sciences*. McGraw-Hill Book Company, Inc., 1956.

Binh H. Le is currently a PhD student at 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, and virtual human modeling and animation. He received his B.S. in Computer Science from the Vietnam National University, Vietnam, in 2008. He held a Vietnam Educational Foundation Fellowship from 2008 to 2010.

Xiaohan Ma is currently a PhD student at 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, virtual human modeling and animation, HCI, and GPU computing. He received his both M.S in Computer Science and B.S. in Computer Science from Zhejiang University (China) in 2005 and 2007, respectively.

Zhigang Deng is currently 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 completed his Ph.D. in Computer Science at the Department of Computer Science at the University of Southern California (Los Angeles, CA) in 2006. Prior that, he also received B.S. degree in Mathematics from Xiamen University (China), and M.S. in Computer Science from Peking University (China). Over the years, he has worked at the Founder Research and Development Center (China) and AT&T Shannon Research Lab. He is a senior member of IEEE and a member of ACM and ACM SIGGRAPH.