Speech-Driven Facial Animation With Realistic Dynamics

R. Gutierrez-Osuna, *Member, IEEE*, P. K. Kakumanu, *Student Member, IEEE*, A. Esposito, O. N. Garcia, *Fellow, IEEE*, A. Bojorquez, J. L. Castillo, and I. Rudomin

***Abstract*—This paper presents an integral system capable of generating animations with realistic dynamics, including the individualized nuances, of three-dimensional (3-D) human faces driven by speech acoustics. The system is capable of capturing short phenomena in the orofacial dynamics of a given speaker by tracking the 3-D location of various MPEG-4 facial points through stereovision. A perceptual transformation of the speech spectral envelope and prosodic cues are combined into an acoustic feature vector to predict 3-D orofacial dynamics by means of a nearest-neighbor algorithm. The Karhunen–Loéve transformation is used to identify the principal components of orofacial motion, decoupling perceptually natural components from experimental noise. We also present a highly optimized MPEG-4 compliant player capable of generating audio-synchronized animations at 60 frames/s. The player is based on a pseudo-muscle model augmented with a nonpenetrable ellipsoidal structure to approximate the skull and the jaw. This structure adds a sense of volume that provides more realistic dynamics than existing simplified pseudo-muscle-based approaches, yet it is simple enough to work at the desired frame rate. Experimental results on an audiovisual database of compact TIMIT sentences are presented to illustrate the performance of the complete system.**

***Index Terms*—face image analysis and synthesis, lip synchronization, 3-D audio/video processing.**

I. Introduction

LIP READING plays a significant role in spoken language communication. It is essential for the hearing-impaired, and also used by normal listeners as an aid to improve the intelligibility of speech signals in noisy environments [1]. Lip movement is also useful for understanding facial expressions and developing tools for human–human and human–machine communication [2]–[4]. Facial animation can provide practical and subjective benefits in human–computer interaction, such as cues in understanding naturally spoken language in noisy environments or ungrammatical utterances, and provide positive feelings during waiting times [1]–[5]. Therefore, computer simulation of human faces capable of accurately reflecting lip movement and emotional states has been a flourishing research area for a few decades, resulting in a vast number of facial models and animation systems [6], [8]. Interest in this area has been reinforced with the addition of facial animation features in the MPEG-4 standard [9].

Despite all these efforts, current facial models and animation techniques are still inaccurate approximations of natural faces, particularly due to the fact that the dynamics of human facial expressions are not yet well understood, and often are not even captured or reproduced in synchrony with the corresponding speech. As a result, the development of appropriate paradigms and tools to animate synthetic faces remains a challenging task. As a step toward accomplishing these goals, this article proposes a methodology for learning speech-based orofacial dynamics from video. The method operates at a subphonemic level, mapping speech acoustic features (e.g., spectral power, F0) onto orofacial configurations, thus bypassing problems with phonetic recognition and the many-to-one relationship between phonemes and visemes. Using a tapped delay line to account for coarticulation, we propose a nearest-neighbor mapping that is competitive with the computationally more intensive training phase of time-delay neural networks (TDNNs) or Hidden Markov Models (HMMs). An animation player is also developed to take advantage of the fine-grained temporal resolution. The player is based on a conventional pseudo-muscle model, adapted for direct use with MPEG-4 Facial Points (FPs), and includes an underlying solid structure that cannot be penetrated. This structure is particularly important for the realistic synthesis of facial dynamics outside the FPs, and provides the model with a sense of volume that is absent in other low cost approaches, as described next.

II. Related Work

Both text and speech have been used as control inputs for animating human faces, also known as "talking heads". In text-driven facial animation, the process generally involves determining a mapping from text (orthographic or phonetic) onto visemes by means of vector quantization [10], [11] or a rule-based system [12], [13]. Facial animation driven by speech can be approached in a similar fashion by deriving the phoneme sequence directly from the speech signal, as is done in speech recognition [14]–[16]. However, mapping from acoustic frames to phonetic units and then to animation frames involves introducing uncertainty that often slights prosodic variables (e.g.,

Manuscript received August 15, 2002; revised July 20, 2003. This work was supported by from NSF/CISE Award 9906340, NSF/CAREER Award 9984426, WSU/OBR Award 00-509-10, CONACyT Grant C30033, and NSF/KDI Award 9980054. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chalapathy Neti.

R. Gutierrez-Osuna is with the Deptartment of Computer Science, Texas A&M University, College Station, TX 77843 USA (e-mail: rgutier@cs.tamu.edu).

P. K. Kakumanu is with the Department of Computer Science and Engineering, Wright State University, Dayton, OH 45435-0001 USA (e-mail: kpraveen@cs. wright.edu).

A. Esposito is with the Department of Psychology at the Second University of Naples, Naples, Italy (e-mail: iiass.anna@tin.it).

O. N. Garcia is with the College of Engineering, University of North Texas, Denton, TX 76203 USA (e-mail: ogarcia@unt.edu).

A. Bojorquez, J. L. Castillo, and I. Rudomin are with the Department of Computer Science, ITESM-CEM, Atizapán de Zaragoza, C.P. 52926, México (e-mail: adbojorq@itesm.mx; rudomin@itesm.mx.).

Digital Object Identifier 10.1109/TMM.2004.840611

pitch, duration, or amplitude). Therefore, it is more appealing to construct a direct mapping from speech acoustics (e.g., spectral power) onto facial trajectories, either through control parameters of the facial animation itself [6], three-dimensional (3-D) coordinates of facial points [17], or articulatory parameters [18].

As a result of coarticulation and dynamics, frame-to-frame mappings that neglect the temporal relationship between speech acoustics and facial animation units are not very accurate. It has been shown [19] that the static relationship between speech acoustics and facial configurations can only account for approximately 65% of the variance in facial motion. We considered this result a powerful encouragement to pursue this avenue of investigation. Several methods have been proposed to capture the complex temporal structure by means of tapped delay lines (e.g., a TDNN) or time-series analysis (e.g., an HMM). Lavagetto [18] synthesized mouth articulatory parameters from power-normalized LPC coefficients using TDNNs. Massaro *et al.* [20] used a multilayer perceptron with an explicit input delay line to predict 37 control parameters for an animated talking head (Baldy) from Mel Frequency Cepstral Coefficients (MFCC). Hong *et al.* [21] used a family of multilayer perceptrons, each trained on a specific phoneme, and a seven-unit delay line of MFCCs to capture co-articulation. Brand [17] predicted the 3-D facial trajectories from LPC and RASTA-PLP acoustic features using an entropy-minimization algorithm that learned both the structure and the parameters of an HMM. These methods, however, rely on iterative estimation of nonlinear models that result in computationally intensive training phases. Moreover, our own experience on speech processing shows that the performance of TDNNs strongly depends upon the network architecture and the learning rate [22]. Differences in phoneme classification performance across network architectures can be in the order of 40%, and model selection requires a long trial-and-error process. In light of these considerations, our effort to model basic orofacial motion directly from speech acoustics attempts to overcome the drawbacks of phonetic-level methods while providing a direct noniterative learning method. Our approach uses a perceptual transformation of the speech spectral envelope and prosodic elements to generate an acoustic feature vector that serves as an input to a nearest-neighbor audiovisual mapping.

Starting with Parke's work [23], several types of face models have been developed, which can be classified either as parametric [23]–[25] or muscle-based [26]–[32]. Parametric approaches represent the facial model with specific control parameters and animate by simple mesh or parameter interpolation, whereas muscle-based approaches attempt to represent to some degree of accuracy the anatomic structure of the face and simulate the behavior of muscles to animate the model. Within muscle-based approaches, some of the most commonly used have no representation for the underlying bone, and consist of a single layer spring-mesh where some of the springs are the "muscles" that pull the vertices in the spring mesh. These models are sometimes called pseudo-muscle based to distinguish them from more anatomical muscle models. One could say that, like the parametric models, these pseudo-muscle approaches essentially consider the face as a flexible mask that can be interpolated or pulled by the muscles without major constraints. Anatomy, however, does place some important restrictions since, underneath the skin, there are rigid structures (skull and jaw) that cannot be penetrated. Models that do not take this underlying structure into account are, therefore, limited in their ability to synthesize realistic orofacial dynamics. To address this issue, some methods manage the underlying bone by having several layers of a spring mesh represent the various tissues, and physically restrict the lowest layer to be moved by the bone, directly (i.e., Waters and Terzopoulos [28]). Kähler [31] uses a very detailed muscle model that incorporates different types of muscles, as well as the effects of bulging and intertwining muscle fibers. The influence of muscle contraction onto the skin is simulated using a mass-spring system that connects the skull, muscle, and skin layers. Thus, Kähler also avoids unnatural penetration of the underlying skull-jaw structure. However, his approach requires very anatomically detailed models and extensive simulation of several layers. Our approach is simpler because it is basically a one-layer pseudo-muscle model. Yet, due to the ellipsoidal structure, our model provides reasonably realistic dynamics at higher speeds comparable to the more common methods mentioned above.

Our player was designed to incorporate these considerations, first, by augmenting a nonanatomical one-layer pseudo-muscle approach with an underlying nonpenetrable structure that approximates the skull and the jaw, and second, by developing an optimized implementation that achieves audio-synchronized animation at 60 frames/s (fps) on midrange ($\sim$>600 MHz clock) personal computers with only standard OpenGL acceleration. In addition, an MPEG-4 [25], [33] compliant interface was also implemented. The details of this facial model are described in Section V.

III. Audiovisual System Overview

The vision capture system employed in this research consists of two color cameras (Kodak ES310), and two dedicated personal computers and frame grabbers capable of acquiring 648×484 video directly to a hard drive at 60 fps to include short phonetic phenomena. Speech signals are captured on one of the personal computers using a shotgun microphone (Sennheiser K6/M66 with Symetrix 302 preamplifier) and saved to disk using a proprietary file format that interleaves $1/60$-s of audio between video frames to ensure synchrony. Once the data has been saved to disk, the 3-D coordinates of various FPs are tracked using stereo correspondence, and a number of acoustic features are extracted from the audio track. These procedures are described in detail in the following subsections.

A. Audio Processing

Speech signals are the result of various sources, including message content, noise, speaker dependencies, and prosody (e.g., speed, intonation). Therefore, the unprocessed speech waveform cannot be directly used to predict facial motion. Several frequency-domain processing techniques have been proposed to separate basic speech features from less relevant information. Examples are Filter Bank Analysis, Smoothed Spectrum, Cepstral Analysis, Linear Prediction Coding (LPC) [34], Perceptual Linear Prediction (PLP), and Relative Spectrum (RASTA-PLP) [35]. These algorithms generally use

default parameter values, which are constant for all phones, disregarding the different nature of such segmental information. This results in articulatory attributes whose robustness depends on the extracted preprocessing parameters. Moreover, all the aforementioned algorithms are based on the assumption that formants and their trajectories are the cues for speech perception. Therefore, they extract features that are based on the unambiguous identification of the important peaks in the spectral envelope, and require a robust method for disambiguating these from additional peaks that may occur for a variety of reasons. However, it has been shown that neither the overall level nor the exact location and tilt of the patterns in the frequency domain but the changes in the shapes and relative locations of the major spectral features appear to be important for phonetic perception [36]. Therefore, we decide to preprocess the speech signals with a perceptually-based analysis, which uses concepts from the psychophysics of hearing to obtain an estimate of the auditory spectrum [37], [38].

The speech signal, sampled at 16 kHz, is processed in frames of $1/60$-s with no overlap between the frames. Overlapping frames were used in the early stages of the research, but they were found not to improve the predictive performance of the models. Each frame is pre-emphasized with an FIR filter $(H(z) = 1 - az^{-1};\ a = 0.97)$ and weighted with a Hamming window to avoid spectral distortions [34]. The pre-emphasized and windowed speech signal is then passed through a filter bank that uses a critical-band resolution of the Fourier spectrum described by

$$\text{Bark} = 13\tan^{-1}\left(\frac{0.76f}{1000}\right) + 3.5\tan^{-1}\left(\frac{f^2}{7500^2}\right) \quad (1)$$

$$\text{BW} = 25 + 75\left[1 + 1.4(f/1000)^2\right]^{0.69} \quad (2)$$

where f, the acoustic frequency (Hz), is mapped onto a perceptual frequency scale referred to as a Critical Band Rate or Bark [39]. The center frequencies of the filter bank are uniformly distributed along the Bark scale, whereas the corresponding bandwidths are defined by (2). From these, a vector of Perceptual Critical Band Features (PCBF) [40] is computed as the log-energy of the acoustic signal:

$$\text{PCBF}_i = \log_{10}\sum_{k=1}^{N}[x_i(k)]^2 \quad (3)$$

where $x_i(k)$ is the kth sample of the speech waveform in the ith frequency band. This crude perceptually-modified spectrum is used to capture the dynamic aspects of the spectral envelope. An example of this acoustic transformation is shown in Fig. 1(b). Note that PLP and RASTA-PLP also contain a perceptual analysis component, but they further preprocess the speech signal by including a predictive modeling (PLP) and a temporal filtering process (RASTA-PLP) to make the extracted features closer to the formant trajectory. In this context, PCBFs can be interpreted as the first stage of PLP or RASTA-PLP analysis. A total of 12 PCBFs corresponding to frequencies $f = \{112, 255, 426, 627, 860, 1147, 1483, 1937, 2452, 3299, 4128, 6811\ \text{Hz}\}$ are extracted from the speech signal and combined with two prosodic cues (fundamental frequency and frame energy) to form a 14-D vector of acoustic inputs to our animation system.

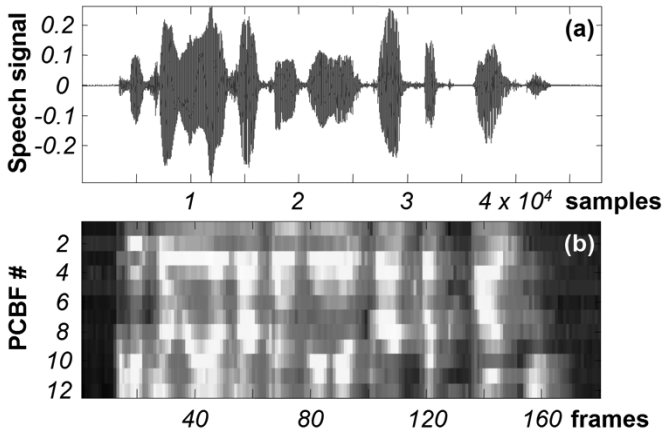

Fig. 1. (a) Audio waveform and (b) PCBFs for the TIMIT sentence "*Pizzerias are convenient for a quick lunch*."

B. Video Processing

To facilitate accurate 3-D tracking of the facial dynamics using a stereo camera pair, 27 markers are placed in the face of the subject at various MPEG-4 FPs [33], as shown in Fig. 4(a). The position of each marker is independently tracked by finding the maximum cross-correlation in a local region of interest centered on the marker's position in the previous frame. The initial position of each marker is manually entered by means of a graphical user interface. This process is performed independently on each of the two stereo sequences. To obtain the 3-D coordinates of the MPEG-4 FPs, we apply a standard calibration procedure comprising a prism with fiduciary markers and a calibration tool to establish the calibration correspondences and calibration matrix [41]. Although the subject is instructed to stand still during data collection, head motion is practically unavoidable due to the natural tendency of human speakers to move their head during speech. Hence, the 3-D coordinates of the tracked FPs contain movements that are associated not only with the production of speech but also with the translational and rotational movements of the head. In order to remove these head movements, head pose is estimated [42] from four facial points (specifically MPEG-4 points 3.11, 3.8, 9.12, and 9.3). Once the head pose of each frame is obtained, the relative 3-D coordinates of the remaining facial points are computed, effectively decoupling facial motion from head motion. Finally, the coordinates of the neutral face in the first frame of each take are subtracted from the remaining frames to yield a vector of relative displacements, as required by the MPEG-4 Facial Animation Parameters. The final result of video processing is a video vector with 81 (27×3) differential measurements, a highly redundant representation since the movement of the various points in the face is highly interdependent. For this reason, the Karhunen–Loéve transform [43] is used to obtain a low-dimensional vector, typically containing 14 dimensions, which captures the principal components of orofacial motion. The details of this processing stage are discussed in Section IV.

C. Mapping Acoustic Dynamics to Visual Articulators

As a result of subphonemic dynamics and coarticulatory effects, it is not feasible to perform a direct frame-to-frame prediction of the video vector $v(t)$ (or its principal components in our case) from the corresponding acoustic vector

$a(t)$, where t denotes the frame index. For this reason, we explicitly encode context by associating each video frame $v(t)$ with the acoustic features from past and future audio frames $a_n(t) = [a(t-n), \cdots a(t), \cdots a(t+n)]$, forming a lookup table of audiovisual dynamic pairs $[a_n,\ v](t)$ from training data. This data can be used to build a tapped delay line model (e.g., a TDNN). In this work, however, a simple and attractive k nearest-neighbor (KNN) procedure is employed. Given an audio window $\hat{a}_n$ from a new voice track, we find the average video configuration $\hat{v}$ of the k closest audio trajectories $a_n(t)$ in the training set using the Euclidean distance. In the present implementation, we use a value of $n = 5$, which corresponds to 83-ms wide windows into the past and the future, and $k = 12$ neighbors, both of which have been empirically optimized [44]. To limit dimensionality, the acoustic dynamics are subsampled 5:1, resulting in a 42-dimensional acoustic vector $a_5(t) = [a(t-5),\ a(t),\ a(t+5)]$. Our experience shows that using a longer window or a finer does not improve performance significantly [44].

D. Audiovisual Database

An audiovisual database of 75 sentences from the TIMIT compact set [45] was collected from a single speaker to evaluate the system. These sentences were spoken by a female American-born (Ohio) speaker according to the following protocol: on each audiovisual recording session the speaker was given a list of sentences and was asked to read them a number of times in random order. To prevent the speaker from becoming fatigued, only five sentences (each repeated five times) were recorded on each session. Therefore, the complete dataset consists of 75×5 (375) sentences. To reduce initialization variations, each sentence/repetition was recorded starting from a neutral facial expression, such as the one shown in Fig. 4(a), which served as a baseline for each audiovisual take. The 75×5 sentences were split into three separate sets, a training set containing 60×4 sentences, a validation set containing the fifth repetition from the previous 60 sentences and a test set containing one repetition from the remaining 15×5 sentences. Separate validation and test sets were utilized to analyze the performance of the KNN mapping both on different takes of training sentences (validation set) and on phonetic sentences not used for training (test set). The training set was used to build the database of audiovisual pairs $[a_n,\ v](t)$, from which video predictions were generated for the validation and test sets according to the KNN rule.

IV. Principal Components of Orofacial Motion

As described in Section III-B, the video-processing stage yields 3-D differential displacements for each of the 27 facial markers, for a total of 81 features for each video frame. Clearly, these video features will have a high level of redundancy since the movement of the various points in the face is highly interdependent. For instance, motion in the lower lip can be affected by the opening/closing of the jaw, and motion in the corners of the lips is coupled to motion in the cheeks. To help unveil these relationships, we employ a Karhunen–Loéve decomposition of the 81-dimensional (81-D) feature space that extracts the principal components of orofacial motion. Principal component analysis (PCA) is a classical multivariate statistics technique

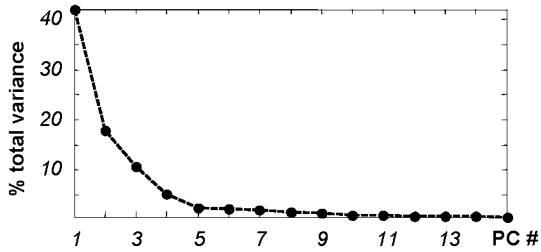

Fig. 2. Contribution of the first 15 PCs to the total variance in the data.

that extracts a low dimensional projection that best represents, in the mean-squared-error (mse) sense, high dimensional data in an orthogonalized multidimensional space. It can be shown that such projection is aligned with the directions of maximum variance of the data [43]. Briefly, PCA approximates a random vector $v \in R^N$ by a linear combination of $M (M < N)$ orthogonal bases, which are the eigenvectors φ_i corresponding to the largest eigenvalues λ_i of the covariance matrix Σ_v of v:

$$u^T = v^T[\varphi_1, \varphi_2, \ldots, \varphi_M, \varphi_{M+1}, \ldots, \varphi_N] \tag{4}$$

where $u = [u_1, u_2, \ldots, u_M]^T$ is the low-dimensional projection of v, the 81-D video vector in our case. When applied to the audiovisual database described in Section III-D, PCA of the 81-D video vector indicates that 90% to 95% of the total variance in facial motion is captured by the first 15 to 24 eigenvectors, respectively. The distribution of variance across the first 15 eigenvalues is shown in Fig. 2. This result clearly indicates that there is a high degree of correlation among the video features, and suggests that a more compact video representation may be obtained by preserving only a few principal components (PCs). In order to analyze the individual contribution of each PC, we reconstruct the 3-D coordinates $\hat{v}$ of the 27 FPs by back-projecting one PC u_k at a time, and substituting the remaining components with their average value $E[u_n]$ in the dataset

$$\hat{v}^T = [E[u_1], \ldots, u_k, \ldots, E[u_N]] \cdot [\varphi_1, \varphi_2, \ldots, \varphi_N]^{-1}. \tag{5}$$

Fig. 3 illustrates the 3-D directions of motion for each of the four largest PCs. The contour of the lips, eyebrows and nose (thin green lines) represents the neutral configuration of the FPs, whereas the (red) arrows represent the principal component of motion at the facial points. The orientation of each arrow represents the direction of maximum motion, whereas the magnitude of the vector is proportional to the range of motion along that direction. As shown in Fig. 3(a), PC_1 is associated with vertical motion in the jaw and the lower lip. Given that these contain some of the most active facial articulators during speech production, it is not surprising that their movements bring the largest contribution to variance in the data. PC_2, in turn, captures motion perpendicular to the image plane, which can be related to the protrusion of the lips. The third PC, however, does not contain information that can be clearly associated with a particular speech articulator. At first, one may be tempted to associate this component with prosodic head motion such as shaking, nodding and rolling, but this interpretation is invalid since head motion is removed beforehand through pose estimation, as described in Section III-B. Therefore, we hypothesize that this component contains variance due to the placement of the markers on

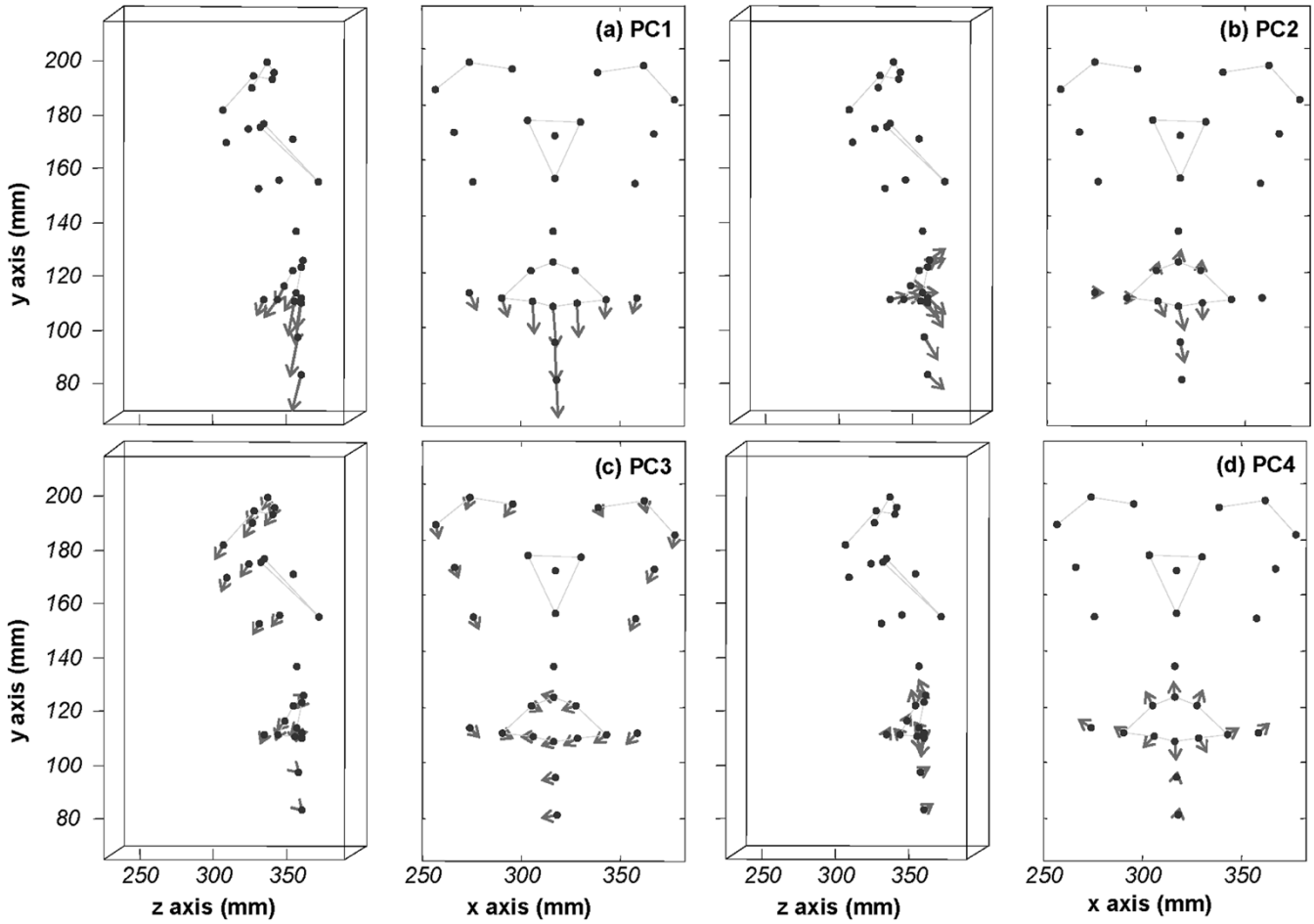

Fig. 3. (a)–(d) Contribution of the first four principal components to facial motion.

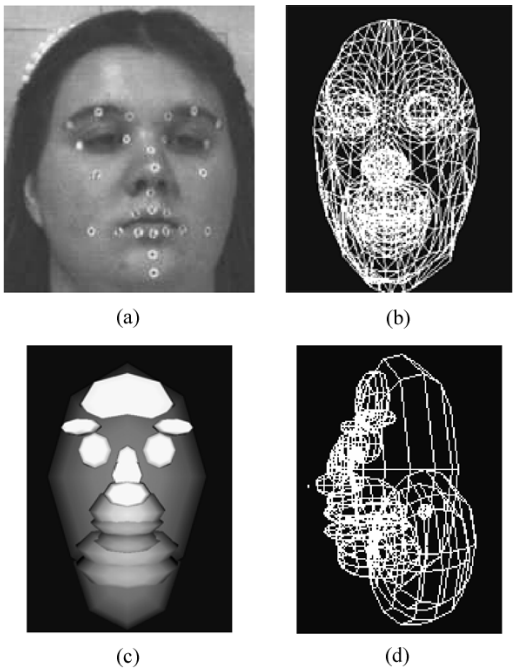

Fig. 4. (a) Neutral face of the subject with visual markers, (b) adjusted wire-frame face, (c) underlying ellipsoids, and (d) wire-frame structure of the ellipsoids, profile view.

the subject's face which, because of lack of obvious reference in the facial features, cannot be precisely matched from one data-collection session to another. Finally, PC_4 captures motion related to the opening and closing of the mouth, as indicated by the low variance around the chin. Subsequent PCs contain low-variance directions of orofacial motion, which do not capture gestures that we can connect to the production of speech, but may nonetheless contribute to the perceptual naturalness of the animation. This issue is explored in Section VI-A.

V. Modeling and Animation

To fully exploit the high-speed audiovisual capture system, we have developed an MPEG-4 compliant player that produces facial animations with realistic dynamics. The player is based on a spring-mesh face with additional springs serving as pseudo-muscles, a fairly standard approach that we have adapted to be directly driven from the 3-D coordinates of the MPEG-4 FPs. In addition, an ellipsoid approximation of the skull and jaw has been implemented to prevent their penetration by the face, thereby giving a sense of volume and a more realistic dynamic behavior.

A. Generic Facial Model

Starting from a basic model by Parke and Waters [46], we have developed a generic facial model with 64 MPEG-4 FPs. The model contains a mesh of 876 triangles and 28 muscles to allow facial expressions and movement. To adjust this generic

model to a particular speaker, the 3-D coordinates of selected MPEG-4 FPs obtained from video are used to fit the vertices of the generic mesh as a particle-spring system with forces, which are adjusted using an Euler ODE solver [47]. Further adjustments in facial proportions as well as facial features (i.e., eyes, mouth, and nose) are performed using MPEG-4 distance fractions and other anthropometric measurements such as distances between the different MPEG-4 FPs.

We associate each of the 28 muscles of the generic face with one or more MPEG-4 FPs (as a head or a tail), a predetermined influence area, and a stiffness constant. The influence area is a list of vertices that are affected by the muscle, as well as the corresponding weight for each vertex, in a manner akin to Kshirsagar [25]. The list of vertices associated to each muscle is determined once, when the generic face model is modified to generate a new model for a particular speaker, and subsequently stored in a data structure. Our approach is a pseudo-muscle approach that attempts to have additional anatomical basis (a solid skull-jaw). The "muscles" are really additional springs that can be contracted or elongated as in standard pseudo-muscle approaches, but that we drive directly from the MPEG-4 FPs. These FPs are moved directly, as predicted from audio. All other points in the muscle's influence area are moved by a combination of the weighted influence of the muscle's FPs, muscle forces calculated from the movement of these FPs, and other spring forces that are related to the elasticity of the skin. This is explained with more detail in the next section.

B. Animation Player

To improve the realism of the animation, the player was augmented by applying an ellipsoidal structure that we originally developed for another application. This structure allowed us to place pieces of clothing over an animated articulated character [48]. The dressed character is approximated using a hierarchy of ellipsoids attached to the character's skeleton, and the pieces of clothing are represented using mass-spring particle systems. First, each particle is associated with the closest ellipsoid. When the character is animated, the ellipsoidal structure is moved accordingly, dragging along the associated particles. Dynamic forces are then applied to the particles. Finally, penetration of the ellipsoids by any particle is corrected. Recent results from this work have shown that real time performance, one of the objectives of our research, can be achieved [49].

When applied to facial animation, this ellipsoid representation avoids unnatural penetration of the underlying skull-jaw structure that is often present in other approaches (both parametric and pseudo-muscle, as explained in Section II). We do not perform an anatomically realistic geometric and physical modeling of the muscles (like Kähler [31]) although this would be possible using the ellipsoids in the system. Instead, we use these ellipsoids to economically represent an approximate geometry of the skull and the jaw that allows us to control collisions. This is far simpler and faster, the latter being our main constraint. If the constraints for nonpenetration of skull and jaw by the vertices in the facial model are not taken into account, the facial dynamics will be incorrect, particularly in parts of the face that are not FPs. As a result of adding this ellipsoidal representation of the skull-jaw, our player is able to model facial dynamics more accurately without incurring in significant costs.

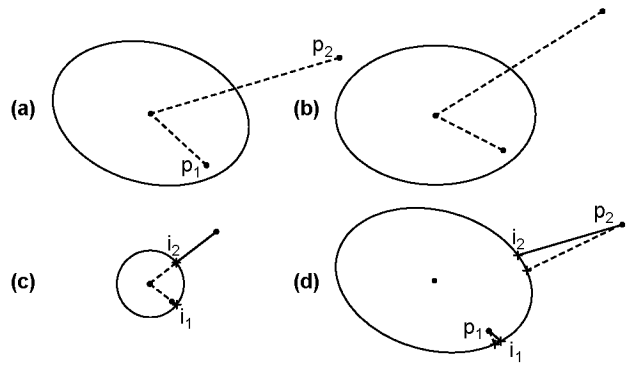

Fig. 5. Efficient collision detection. (a) Two example points near an ellipsoid. (b) Their position after origin-centering and axis-aligning. (c) Their position after scaling, where penetration and points of contact are determined. (d) Intersection points are found.

The complete process can be described as follows. First, the 3-D coordinates of MPEG-4 FPs undergo several adjustments before the animation is played. These include adapting the face mesh to the proportions recorded in the video data, translating the center of the face in the video data to the correct coordinate system, calculating rotations and individual adjustments necessary for the FPs, and filtering the depth estimates in order to reduce noise. Once these preprocessing steps are completed, the 27 MPEG-4 FPs can then be used to control the remaining vertices in the mesh in order to produce an overall natural movement during animation playback. This involves four steps.

Step 1) Modify vertices using a set of neighboring FPs. The influence of neighboring FPs on other vertices is determined during startup. This influence is weighted by the distance between each vertex and those FPs within a certain radius. As a result, a vertex may be influenced by more than one FP. An exception is made with the vertices that define the lips, since they closely follow specific FPs.

Step 2) Determine jaw rotation from the 3-D coordinates, and transform the corresponding vertices. The rotation of the jaw is determined from the angular difference between the current position of FPs on the chin and their corresponding position on the neutral face. When the ellipsoids in the jaw are rotated, all the vertices associated to them are transformed accordingly, unless they were already modified in Step 1.

Step 3) Incorporate the effects of spring forces. Springs are defined for every edge in the face mesh and for every muscle. Spring forces are calculated in the standard manner using Hooke's Law. The resulting displacements are finally applied to all the vertices in the model, with the exception of FPs, for which the measured coordinates are known.

Step 4) Correct for ellipsoid penetration. This last step is critical, particularly in the area influenced by mouth movements, because the use of influence areas and springs may cause penetration of the skull-jaw for vertices that are not FPs. Every vertex in the model (except FPs, for which we have measured or predicted coordinates) is checked for ellipsoid penetration, and those in violation of this constraint are moved to the intersection point between the ellipsoid's surface and the vector from the current (trial) to the previous position. One of the main optimization steps in our system turned out to be an efficient point-to-ellipsoid distance calculation. The

procedure we follow is very simple and fast, and assumes that the ellipsoids are scaled spheres. As illustrated in Fig. 5, the method consists of transforming the ellipsoid (and the points being tested for intersection) to be origin-centered, axis-aligned and of radius one. Testing for penetration can then be determined by checking whether the transformed points are within this unit-radius sphere, a very fast computation since the transformations can be performed efficiently with standard OpenGL hardware. This procedure is also used for estimating the distance from the point to the ellipsoid, and to the point on the surface of the ellipsoid where the intersection was found. The line from the transformed point to the origin is intersected with the unit sphere, and the intersection point is then transformed back to the original space. This process does not provide the actual closest point of the untransformed ellipsoid to the untransformed point, but a close approximation that serves its purpose well. It is important to note that, because FPs are moved according to how they have been measured, any noise or jitter that may be visible in the animation is due to the data itself rather than the player. The resulting wireframe face and underlying ellipsoids are shown in Fig. 4(b)–(d).

Maintaining a fixed 60-fps rate during rendering is critical, since every video frame must be processed and displayed in synchrony with the audio stream if phonetic phenomena are to be perceived in correlated time. Thus, runtime speed was one of the main criteria in the design of the animation software. All steps in the above algorithm were carefully optimized, particularly the point-to-ellipsoid distance calculation, and a simple and explicit Euler ODE integration to incorporate the effects of spring forces. To ensure synchrony with the audio track, earlier versions of the player determined if the computer was capable of maintaining a minimum frame rate of 60 fps before the animation was started, and aborted execution otherwise. The most current version of the player uses a graceful degradation scheme (dropping frames) instead. In this case, however, the animation results will be inconsistent with our research goal of accurately reproducing very short phenomena in the orofacial dynamics. On the other end, the player automatically limits the frame rate to a maximum of 60 fps if this speed could be exceeded and, as a result, synchronization is achieved.

VI. Results

Validation of the proposed audiovisual mapping and MPEG-4 compliant animation is performed in two stages using the database of audiovisual TIMIT sentences described in Section III-D. First, an appropriate number of principal components of orofacial motion is determined by means of two perceptual tests with human subjects. Second, the predictive accuracy of the audiovisual mapping is illustrated by tracking a few lip articulators that have been shown to be critical for automatic visual phone recognition.

A. Perceptually-Relevant Principal Components of Orofacial Motion

To verify the hypothesis that the low-variance directions of orofacial motion contribute to the perceived quality of the final animation, two perceptual tests were performed among the members of our research group, for a total of six subjects. Each subject was presented with three different facial animations, containing the first 14, 23, and 80 principal components, or 90%, 95%, and 100% of the total variance, respectively (the third PC was consistently discarded in all of the experiments as it was shown to contain noise). The subjects were asked to identify which animation seemed most pleasant and natural. The results of this informal test indicated that the subjects were not able to express a strong preference between the animations reproduced with 90% and 95% of the variance. More interestingly, the animation with 100% variance was judged to be of lower quality than the other two, since it captures not only speech-related motion but also noise. This result allowed us to conclude that a PCA decomposition of visual data is a valuable tool for identifying the most relevant components of facial motion.

A second experiment was performed to determine if only those PCs whose contribution to 3-D orofacial motion can be directly interpreted (i.e., from Fig. 3) would be sufficient to yield realistic dynamics. In this case the subjects were asked to rate two animations generated with four $(\mathrm{PC}_{1,2,4,5})$ and 14 PCs $(\mathrm{PC}_{1,2,4-15})$. All subjects independently agreed that the animation with four PCs appeared less natural than the one with 14 PCs. This result supports the hypothesis that PCs with lower eigenvalues do encode fine details that are essential to produce a natural looking animation. Thus, it was concluded that $\mathrm{PC}_{1,2,4-15}$ is an appropriate representation for the subsequent analyzes and final implementation in this article.

B. Audiovisual Predictions

The previous perceptual tests allowed us to determine an appropriate low-dimensional projection of the video feature vector where the audiovisual mapping should operate. Given a new audio window $\hat{a}_n$, the KNN algorithm finds the closest audio trajectories $a_n(t)$ in the training set and uses their 14 largest PCs $[u_1, u_2, \ldots, u_{14}]$ to obtain the video prediction $\hat{v}$ through a back-projection:

$$\hat{v}^T = [u_1, \ldots, u_{14}, E[u_{15}], \ldots, E[u_{81}]]\,[\varphi_1, \varphi_2, \ldots, \varphi_{81}]^{-1}. \tag{6}$$

To illustrate the accuracy of the predicted dynamics, we provide trajectories for three separate orofacial articulators that are computed from the predicted 3-D coordinates $\hat{v}$ of a few MPEG-4 FPs: mouth height $(\mathrm{MH} = 8.1{\cdot}y\text{–}8.2{\cdot}y)$, mouth width $(\mathrm{MW} = 8.3{\cdot}x\text{–}8.4{\cdot}x)$ and chin height $(\mathrm{CH} = 9.3{\cdot}y\text{–}2.10{\cdot}y)$ These three parameters capture articulatory motion in the lips and the jaw, where the most important features for automatic visual phone (viseme) recognition occur [50], [51]. Predictions for each of the three articulators, both on validation and test data, are shown in Fig. 6 for the TIMIT sentences: 1) "*Pizzerias are convenient for a quick lunch*," 2) "*The bungalow was pleasantly situated near the shore*," 3) "*Clear pronunciation is appreciated*," and 4) "*Barb's gold bracelet was a graduation present*." The predicted trajectories (thin red line) have been mean-filtered with a 50-ms window to reduce jitter. These results illustrate the good performance of the KNN audiovisual mapping in the validation

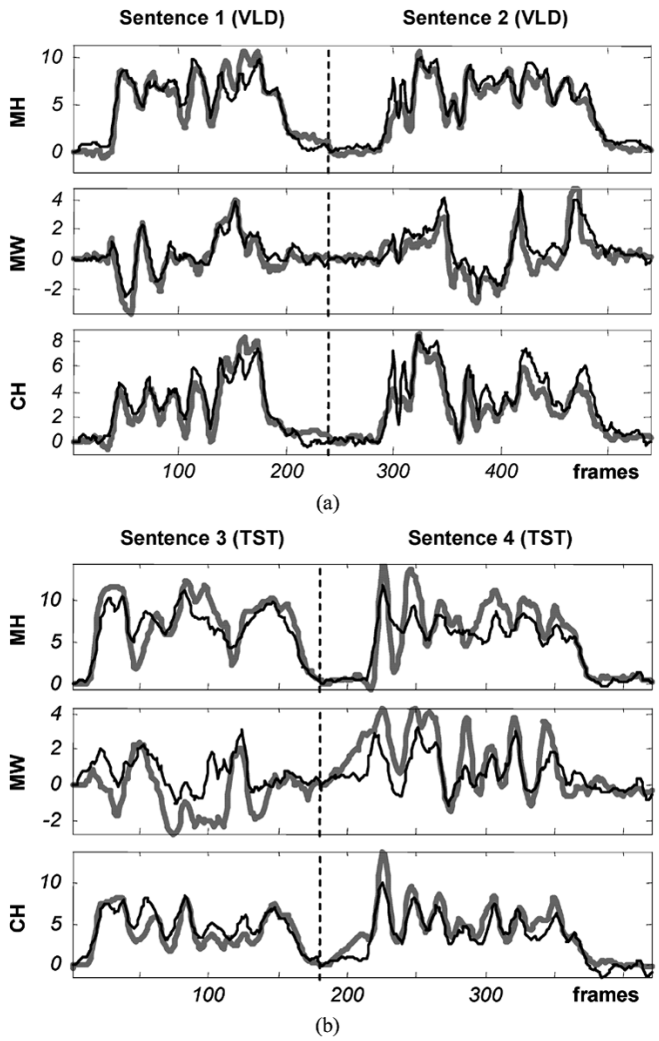

Fig. 6. Predicted (thin red line) versus actual (thick blue line) lip trajectories for the validation and test sentences.

TABLE I
NORMALIZED MSE (ϵ) AND CORRELATION COEFFICIENT (ρ) OF THE KNN AUDIO-VISUAL PREDICTIONS ON VALIDATION AND TEST DATA

Articulatory parameter	VALIDATION		TEST	
	ε	ρ	ε	ρ
MH	0.16	0.92	0.30	0.86
MW	0.32	0.83	0.62	0.64
CH	0.20	0.89	0.29	0.84

data. Although synthesis of test data is less accurate, the KNN mapping is still able to predict the majority of the openings and closures of the three articulators.

Table I shows the average prediction results on the 60 validation and 15 test sentences for the three articulators in terms of the normalized mse ε and correlation coefficient ρ, defined by

$$\varepsilon = \frac{1}{\sigma_p^2 T}\sum_{t=1}^{T}\left(\hat{p}(t)-p(t)\right)^2$$
$$\rho = \frac{1}{T}\sum_{t=1}^{T}\frac{\left(\hat{p}(t)-\mu_{\hat{p}}\right)\left(p(t)-\mu_p\right)}{\sigma_{\hat{p}}\sigma_p} \quad (7)$$

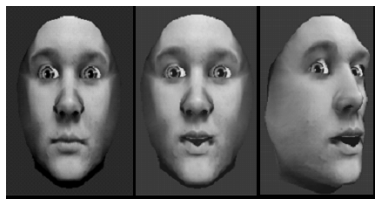

Fig. 7. Some frames of the final animation driven by speech acoustics.

where T is the total number of frames in the dataset, $p(t)$ is the true articulatory parameter at frame $t, \hat{p}(t)$ is its KNN prediction, and μ and σ are their sample mean and standard deviation, respectively. Along with the trajectories in Fig. 6, these results indicate that the vertical motion of the lips, represented by the articulators *MH* and *CH*, can be predicted more accurately than horizontal articulators such as *MW*. These results also show that prediction of orofacial dynamics from the validation set is a simpler task than generalizing to the new phonetic sequences in the test set, as one would reasonably expect. Overall, our results indicate that the KNN mapping is capable of reproducing the majority of the high-level features in the orofacial dynamics.

The animation module can display at fixed frame rate of 60 fps, in synchrony with the audio track. The system has been tested on midrange personal computers with minimal OpenGL acceleration, such as a Pentium III laptop at 650 MHz with a S3 Savage/IX chip. Some frames of the final animation are depicted in Fig. 7. It is difficult to portray in print the dramatic improvement in the realism of the animation dynamics that is obtained by using the ellipsoidal structure.

VII. CONCLUDING REMARKS

We have presented an integral system capable of generating facial animations with realistic dynamics directly from speech. The system employs high-speed stereovision to capture subphonemic phenomena by tracking the 3-D coordinates of 27 MPEG-4 FPs. A PCA decomposition of orofacial motion has also been presented to analyze the perceptually natural components of orofacial motion as well as reduce the dimensionality of the video representation. An acoustic feature vector which combines a perceptually modified spectrum and prosodic cues of speaking rate and intonation is used to predict the complete orofacial dynamics by means of an attractive KNN algorithm operating on a PCA subspace of the 3-D coordinates. Coarticulation and subphonemic phenomena are taken into consideration using a 167 ms-wide tapped delay line on the audio track. Given a newly spoken sentence, the system generates 3-D trajectories directly from speech, bypassing phonetic recognition and many-to-one relationships between phonemes and visemes, thus preserving the realism and synchrony of the movements.

An MPEG-4 compliant player has been developed to generate facial animations directly from the 3-D coordinates of FPs. The animation, based on a pseudo-muscle model, has been augmented with an underlying nonpenetrable ellipsoid structure to approximate the skull and the jaw, providing the model with a sense of volume for improved realism. The player is capable of generating realistic dynamics at 60 fps on a notebook without special hardware acceleration.

A. Future Work

Improvements in predictive accuracy for test data can be obtained by employing a larger dataset of continuous speech containing hundreds of sentences with a rich balance of phonetic and prosodic cues. For large audiovisual databases, it is possible that the KNN procedure will become impractical in terms of storage requirements and search time for real-time applications. Vector quantization procedures are currently being investigated to compress our dataset to a reduced codebook of audiovisual pairs and to evaluate the potential loss of predictive accuracy. Additional work needs to be performed on the animation module to improve lip movement and incorporate eye movement, most likely by implementing sphincter (circular) muscles for both the mouth and the eyes. Improvements in the ellipsoidal approximation of the skull, as well as modeling of muscles with ellipsoids, will allow us to achieve a more anatomically faithful animation. The addition of skin wrinkles, hair and improved rendering without compromising real-time performance will also be investigated in future versions of the player.

ACKNOWLEDGMENT

N. Balan, J. Garringer, and T. Husain are gratefully acknowledged for their devoted efforts in data collection.

REFERENCES

[1] D. W. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press, 1998.

[2] L. E. Bernstein and C. Benoit, "For speech perception by humans or machines, three senses are better than one," in *Proc. Int. Conf. Spoken Language Processing*, vol. 3, 1996, pp. 1477–1480.

[3] A. Rogozan and P. Deléglise, "Adaptive fusion of acoustic and visual sources for automatic speech recognition," *Speech Commun.*, vol. 26, pp. 149–161, 1998.

[4] C. Benoît and B. Le Goff, "Audio-visual speech synthesis from french text: Eight years of models, designs, and evaluation at the ICP," *Speech Commun.*, vol. 26, pp. 117–129, 1998.

[5] I. S. Pandzic, J. Ostermann, and D. Millen, "User evaluation: Synthetic talking faces for interactive services," *Visual Comput.*, vol. 15, pp. 330–340, 1999.

[6] M. Cohen and D. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, N. M. Thalmann and D. Thalmann, Eds. Tokyo, Japan: Springer-Verlag, 1993, pp. 141–155.

[7] E. Cosatto and H. P. Graf, "Sample-based synthesis of photo-realistic talking heads," in *Proc. Computer Animation*, Philadelphia, PA, Jun. 8–10, 1998, pp. 103–110.

[8] J. Ostermann, "Animation of synthetic face in MPEG-4," in *Proc. Computer Animation*, Philadelphia, PA, Jun. 8–10, 1998, pp. 49–51.

[9] Overview of the MPEG-4 Standard. ISO/IEC JTC1/SC29/WG11 M2725, 1999.

[10] S. Morishima and H. Harashima, "A media conversion from speech to facial image for intelligent man-machine interface," *IEEE J. Sel. Areas Commun.*, vol. 9, no. 4, pp. 594–600, May 1991.

[11] K. Waters and T. M. Levergood, "DECface: An Automatic Lip Synchronization Algorithm for Synthetic Faces," DEC Cambridge Research Lab., Cambridge, MA, Tech. Rep. CRL 93/4, 1993.

[12] C. Pelachaud, N. I. Badler, and M. Steedman, "Generating facial expressions for speech," *Cogn. Sci.*, vol. 20, pp. 1–46, 1996.

[13] J. Beskow, "Rule-based visual speech synthesis," in *Proc. Eurospeech*, Madrid, Spain, 1995.

[14] L. M. Arslan and D. Talkin, "Codebook based face point trajectory synthesis algorithm using speech input," *Speech Commun.*, vol. 27, no. 2, pp. 81–93, 1999.

[15] T. Ohman, "An Audio-Visual Speech Database and Automatic Measurements of Visual Speech,", Stockholm, Sweden, Tal Musik Hörsel—Quaterly Progress Status Rep., 1998.

[16] E. Yamamoto, S. Nakamura, and K. Shikano, "Lip movement synthesis from speech based on Hidden Markov models," *Speech Commun.*, vol. 26, no. 1–2, pp. 105–115, 1998.

[17] M. Brand, "Voice Puppetry," *SIGGRAPH 1999*, pp. 21–28, 1999.

[18] F. Lavagetto, "Converting speech into lip movements: A multimedia telephone for hard of hearing people," *IEEE Trans. Rehab. Eng.*, vol. 3, no. 1, pp. 90–102, Mar. 1995.

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

[20] D. W. Massaro, J. Beskow, M. M. Cohen, C. L. Fry, and T. Rodriguez, "Picture my voice: Audio to visual speech synthesis using artificial neural networks," in *Proc Int. Conf. on Auditory-Visual Speech Processing*, Santa Cruz, CA, Aug. 1999, pp. 133–138.

[21] P. Hong, Z. Wen, and T. S. Huang, "Real-time speech-driven face animation with expressions using neural networks," *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 916–927, Jul. 2002.

[22] A. Esposito, G. Aversano, and F. Quek, "Optimal parameters in neural network models for speech phoneme characterization," in *Perspective in Neural Computing, Neural Nets, Proc. Wirn 2001*, Salerno, Italy, May 21–24, 2001, pp. 178–186.

[23] F. I. Parke, "Computer Generated Animation of Faces," M.S. thesis, Tech, Rep, UTEC-CSC-72-120, Dept. Comput. Sci., Univ. Utah, Salt Lake City, UT, 1972.

[24] ——, "Parameterized models for facial animation," *IEEE Comput. Graph. Applicat.*, vol. 2, no. 9, pp. 61–68, 1982.

[25] S. Kshirsagar, S. Garchery, and N. Magnenat-Thalmann, "Feature point based mesh deformation applied to MPEG-4 facial animation," in *Deformable Avatars*. Norwell, MA: Kluwer, 2001, pp. 24–30.

[26] L. Nedel and D. Thalmann, "Real time muscle deformations using mass-spring systems," in *Proc. Computer Graphics International*, Hannover, Germany, Jun. 1998, pp. 156–165.

[27] S. M. Platt and N. Badler, "Animating facial expressions," in *Proc. SIGGRAPH*, 1981, pp. 245–252.

[28] K. Waters and D. Terzopoulos, "A physical model of facial tissue and muscle articulation," in *Proc. First Conf. Visualization in Biomedical Computing*, Atlanta, GA, 1990, pp. 77–82.

[29] K. Waters, "A muscle model for animating three-dimensional facial expression," in *Proc. SIGGRAPH*, 1987, pp. 17–24.

[30] S. Morishima, "Face analysis and synthesis," *IEEE Signal Process. Mag.*, vol. 18, no. 3, pp. 26–34, May 2001.

[31] K. Kähler, J. Haber, and H. Seidel, "Geometry-based muscle modeling for facial animation," in *Proc. Graphics Interface*, 2001, pp. 27–36.

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

[33] F. Lavagetto and R. Pockaj, "The facial animation engine: Toward a high-level interface for the design of MPEG-4 compliant animated faces," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 2, pp. 277–289, Mar. 1999.

[34] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[35] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[36] D. H. Klatt, "Prediction of perceived phonetic distance from critical band spectra: A first step," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1982, pp. 1278–1281.

[37] E. Zwicker, "Suddivision of the audible frequency range into critical bands (Frequenzgruppen)," *J. Acoust. Soc. Amer.*, vol. 33, p. 248, 1961.

[38] E. Terhardt, "On the perception of spectral information in speech," in *Hearing Mechanisms and Speech*, O. Creutzfeld, O. Scheich, and C. Schreiner, Eds. Berlin, Germany: Springer-Verlag, 1979, pp. 26–28.

[39] J. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, vol. 81, no. 9, pp. 1215–1247, Sep. 1993.

[40] G. Aversano, A. Esposito, and M. Marinaro, "A new text-independent method for phoneme segmentation," in *Proc. IEEE Midwest Symposium on Circuits and Systems*, Dayton, OH, 2001.

[41] P. Kakumanu, R. Gutierrez-Osuna, A. Esposito, R. Bryll, A. Goshtasby, and O. N. Garcia, "Speech driven facial animation," in *Proc. Workshop on Perceptive User Interfaces*, Orlando, FL, Nov. 15–16, 2001.

[42] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-D point sets," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, no. 5, pp. 698–700, May 1987.

[43] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2000.
[44] P. Kakumanu, "Audio-Visual Processing for Speech Driven Facial Animation," M.S. Thesis, Comput. Sci. Dept., Wright State Univ., Dayton, OH, 2002.
[45] J. S. Garofolo *et al.*, "Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database," NIST, Gaithersburg, MD, 1988.
[46] F. I. Parke and K. Waters, *Computer Facial Animation*. Wellesley, MA: A. K. Peters, 1996.
[47] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge, U.K.: Cambridge Univ. Press, 1992.
[48] I. Rudomin, R. Perez-Urbiola, M. Melon, and J. Castillo, "Multi-layer garments using isosurfaces and physics," *J. Visualiz. Comput. Animat.*, vol. 12, no. 4, pp. 215–226, 2001.
[49] I. Rudomin and J. Castillo, "Distance fields applied to character animation," *Eurographics*, Sep. 9–10, 2002.
[50] A. Montgomery and P. Jackson, "Physical characteristics of lips underlying vowel lipreading performances," *J. Acoust. Soc. Amer.*, vol. 73, no. 6, pp. 2134–2144, 1983.
[51] K. Finn, "An Investigation of Visible Lip Information to be Used in Automatic Speech Recognition," Ph.D. dissertation, Dept. Comput. Sci., Georgetown Univ., Washington, DC, 1986.

R. Gutierrez-Osuna (M'00) received the B.S. degree in electronics engineering from the Polytechnic University of Madrid, Spain, in 1992, and the M.S. and Ph.D. degrees in computer engineering from North Carolina State University, Raleigh, in 1995 and 1998, respectively.

From 1998 to 2002, he served on the faculty at Wright State University, Dayton, OH. He is currently an Assistant Professor of computer engineering at Texas A&M University, College Station. His research interests include pattern recognition, machine olfaction, biological cybernetics, speech-driven facial animation, computer vision, and mobile robotics.

P. K. Kakumanu (S'03) received the M.S. degree in computer science from Wright State University (WSU), Dayton, OH, in 2002. Currently, he is pursuing the Ph.D. degree in computer science at WSU, where he is an Instructor.

His research interests are in the areas of human–computer interaction, speech animation, and pattern recognition.

A. Esposito is an Associate Professor in the Department of Psychology, Second University of Naples, Italy. She is also affiliated with the International Institute for Advanced Scientific Studies, Naples, and the Speech Communication Group, Massachusetts Institute of Technology, Cambridge. From 1999 to 2002, she was a Research Associate at Wright State University, Dayton, OH. Her research interests are on speech segmentation, acoustic features of visual speech, cross modal analysis of speech, gestures and gaze, and neural networks.

O. N. Garcia (M'58–SM'71–F'84) received the B.S. and M.S. degrees from North Carolina State University, Raleigh, and the Ph.D. degree from the University of Maryland, College Park.

He is currently Founding Dean of Engineering at the University of North Texas, Denton. Previously, he was NCR Distinguished Professor and Chair in the CS&E Department of Wright State University, Dayton, OH. He served at the National Science FOundation as Program Officer in the CISE and EHH Directorates for several years. He was Full Professor of electrical engineering and computer science at George Washington University, Washington, DC, and has been Charter Chair of the Department of Computer Science and Engineering at the University of South Florida, Tampa.

Dr. Garcia received the Merwin Award of the Computer Society and the Emberson Award of the IEEE. He is a Fellow of the American Association for the Advancement of Science. He is a past president of the IEEE Computer Society and has served on the IEEE Board of Directors. He is currently Division V Director-Elect to the IEEE Board of Directors.

A. Bojorquez received the B.S. degree in computer systems engineering from the Instituto Tecnologico y de Estudios Superiores de Monterrey (ITESM-CEM), Mexico, in 1997 and the M.S. degree in computer science from the same institution in 2001.

From 1997 to 1999, she worked in the Information Systems division of the physical plant at ITESM-CEM, developing software, as a webmaster and giving technical support. She is now a Professor of computer science at ITESM-CEM. Her research interests are in the areas of facial modeling and animation.

J. L. Castillo received the B.S. degree in computer systems engineering from the Instituto Tecnologico de Hermosillo, Mexico, in 1996, and the M.S. degree in computer science from the Instituto Tecnologico y de Estudios Superiores de Monterrey (ITESM-CEM), Mexico, in 2002.

From 1997 to 2000, he developed software to graphically locate and query power distribution equipment information for CFE, the Mexican power agency. He is currently back in Hermosillo, combining his teaching at the local ITESM campus with his work as a freelance developer of graphics, web, and client-server software. His research interests are in the areas of videogames, virtual environments, cloth simulation, and face animation.

I. Rudomin received the Ph.D. degree in computer science from the University of Pennsylvania, Philadelphia, in 1990. His dissertation's topic was "Simulating Cloth using a Mixed Geometrical-Physical Method," under advisor Dr. Norman Badler.

Since 1991, he has been a professor at the Instituto Tecnologico y de Estudios Superiores de Monterrey (ITESM-CEM), Mexico. His interests are in human, facial, and cloth modeling and animation, as well as multiuser virtual environments.