INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 6, 331–346, 2003

Audiovisual Speech Synthesis

G. BAILLY, M. BÉRAR, F. ELISEI AND M. ODISIO
Institut de la Communication Parlée UMR CNRS no 5009 INPG/Univ. Stendhal 46, av. Félix Viallet 38031, Grenoble Cedex, France
bailly@icp.inpg.fr
berar@icp.inpg.fr
elisei@icp.inpg.fr
odisio@icp.inpg.fr

Abstract. This paper presents the main approaches used to synthesize talking faces, and provides greater detail on a handful of these approaches. An attempt is made to distinguish between facial synthesis itself (i.e. the manner in which facial movements are rendered on a computer screen), and the way these movements may be controlled and predicted using phonetic input. The two main synthesis techniques (model-based vs. image-based) are contrasted and presented by a brief description of the most illustrative existing systems. The challenging issues—evaluation, data acquisition and modeling—that may drive future models are also discussed and illustrated by our current work at ICP.

Keywords: text-to-speech synthesis, audiovisual synthesis, facial animation, talking faces

1. Introduction

Since the pioneering work of Parke (1975, 1982, 1996), Platt (1981) and Waters (1987, 1992), the computer graphics community has maintained a high level of interest in trying to reproduce realistic facial movements for speech, facial expression, and for activities such as chewing or swallowing. A key event in animation was the film "Tony de Peltrie" (Bergeron and Lachapelle, 1985) produced from the University of Montreal where the animation (speech and expression) of the face of the character was the main way of telling the story. This short film popularized the use of shape interpolation between key frames for facial animation.

For nearly 30 years the conventional approach to synthesize a face has been to model it as a 3D object. In these *model-based* approaches, control parameters are identified that deform the 3D structure using geometric, articulatory or muscular models. Nowadays such comprehensive approaches are challenged by *image-based* systems where segments of videos of a speaker are retrieved and minimally processed before concatenation. This evolution, surprisingly, parallels—during a shorter period—the evolution of acoustic synthesis, where corpus-based synthesis outperforms parametric (articulatory then formant) synthesis. The more direct link between articulation and facial deformation, compared to articulation and acoustics, together with the need for giving the gift of speech to virtual creatures—from speaking pets to speaking objects for which we do not evidently have reference natural stimuli—help, in case of facial animation, to maintain a balance between the two approaches.

We will first describe some of the main features of these two approaches, trying to distinguish between the plant itself—comprising the parameterization of shape and appearance of the face—and its control from phonetic input. Then we will comment on the few evaluation results by comparing the performance in terms of intelligibility, ease of comprehension and general acceptability by end users. We will finally argue for *data-driven* comprehensive 3D models of facial deformation that take into account the articulatory degrees of freedom of the musculo-skeletal system and present the current work conducted at ICP using video-based motion capture data.

2. Model-Based Visual Synthesis

The models that will be presented in this section have in common the aim of reproducing visible 3D facial movements with realistic motions. They differ in the way motion is actually implemented and controlled. Most model-based talking heads used in current text-to-audiovisual speech synthesizers are descendants of Parke's (1972, 1982) software and his particular 3-D talking head. This class of models should be classified as terminal-analog synthesizers in the sense that they do not aim at simulating the underlying physiological mechanisms that produce the speech signals and the facial deformations, but only attempt to reproduce their geometrical consequences. We will first describe briefly such a *geometric* approach and then mention some partial *biomechanical* models of speech articulators that are under development.

2.1. Parke's Descendants

Baldi from the Perceptual Science Laboratory of the University of California, Santa Cruz (PSL) (Massaro, 1998b), Holger, Sven and other characters from the Department of Speech, Music and Hearing of the Royal Institute of Technology—Stockholm (KTH) (Beskow, 1995; Beskow et al., 1997) or the Finnish Talking Head developed at the Laboratory of Computational Engineering of the Helsinki University of Technology (LCE) (Olives et al., 1999) (see Fig. 1) are all 3D computer graphic objects defined by a set of meshes describing the surface geometry of various organs (skin, teeth, eyes, etc . . .) involved in the production of speech. These polygonal surfaces typically connect a few hundred 3D vertices. Such articulated meshes are often used as generic models in model-based movement tracking systems (Li et al., 1993; Tsai et al., 1997; Eisert and Girod, 1998) (see Fig. 2).

Control parameters move vertices (and the polygons formed from these vertices) on the face by simple geometric functions such as rotation (e.g. jaw) or translation of the vertices in one or more dimensions (e.g., mouth opening or widening). Effects of these basic operations are tapered within specified regions of the face and blended into surrounding regions. Interpolation is also used for most regions of the face that change shape (cheekbones, neck, mouth . . .) or for generating facial expressions. Each of these areas is independently controlled between extreme shapes and associated with a parameter value. Eyes are often modeled by a specific procedure that typically accepts parameters for eye position, eyeball orientation and size, iris color and size or pupil size.

Note that these control parameters are quite heterogeneous: they can be the 3-D coordinates of a single point such as lip corners, or they can drive complex articulatory gestures such as the tuck for labiodentals, or more complex facial expressions such as smiling or surprise.

Such a synthesis strategy has become a standard in the context of the industrial *ISO/IEC MPEG-4* norm. An audiovisual-scene in MPEG-4 is divided into

Figure 1. A gallery of Parke's descendants. From left to right: Sven from KTH, Baldi from PSL, the LCE talking head.

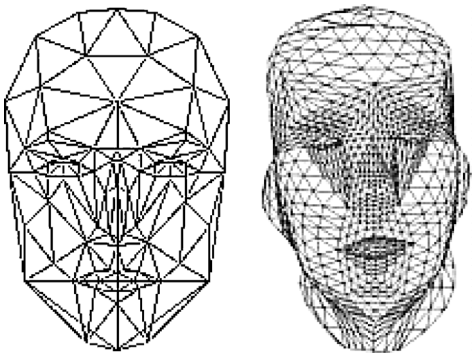


Figure 2. 3D meshes commonly used for tracking head postures and face movements. From left to right: Candide (Rydfalk, 1987) and Eisert's MPEG4 compliant articulated head (Tsai et al., 1997).

different objects. Each of these objects will be encoded separately. The encoded audio/video-objects are multiplexed together with a scene description and transmitted to the viewer. At the viewer-side the datastream is demultiplexed and all the separate audio/video-objects are rendered, according to the scene description, in real-time to the screen. The audio/video objects could also be synthetically generated (*SNHC* Synthetic Natural Hybrid Coding). Besides rigid 3D objects, MPEG-4 defines deformable 3D structures including body and face objects (Doenges et al., 1997; Pockaj et al., 1999). The 3D coordinates of the 84 *FP* (Feature Points) are controlled by a set of 68 *FAP* (Facial Action Parameters) that "are responsible for describing the movements of the face, both at low level (i.e. displacement of a specific single point of the face) or at high level (i.e. reproduction of a facial expression)" (Pockaj et al., 1999, p. 33). More than 30 FAP control lower face movements and are thus directly concerned with speech movements.

2.2. Articulatory Degrees-of-Freedom

In MEPG-4, three main problems arise when piloting mesh deformations from FAP:

1. FAPs are at the same time *geometric* and *articulatory* degrees-of-freedom. The jaw feature point (taken as the mean position of the two lower incisors) acts also as the mean carrier of the lip movements. There is thus a contradiction between an extrinsic geometric control of the lip aperture and the intrinsic *articulatory* control between lips and jaw. This antagonism is solved in MEPG-4 by the laconic instruction associated with FAP3 *open_jaw* that it "does not affect mouth opening" (Tekalp and Ostermann, 2000, p. 412).
2. Most FAPs are low level, and do not take into account speech-specific gestures, which led Vignoli and Braccini (1999) to add another layer of control parameters, called APs (Articulatory Parameters), corresponding to mouth height, mouth width, lips protrusion and jaw rotation, that control the FAPs.
3. Although these APs constitute a more comprehensive set of *articulatory* degrees-of-freedom, the low-level problem of computing the next position of tens of vertices around every displaced feature point remains, and cannot be solved in the articulatory parameter domain. Ad hoc solutions such as simple tri-linear or 3D-spline interpolation, Radial Basis Functions (RBF) or more sophisticated flesh models could be used but these models should be parameterized in order to take into account the non-uniform and non-isotropic changes of the directions of forces applied to the skin by the underlying musculo-skelettal structure.

Instead of ad hoc tapering or shape interpolation, we have proposed already (Badin et al., 2000; Revéret et al., 2000; Elisei et al., 2001) to define APs as *articulatory* degrees-of-freedom extracted by a guided statistical analysis of 3D coordinates of hundreds of facial fleshpoints gathered on a human speaker (see Section 6.1).

2.3. Skin and Muscle-Based Facial Animation

A more comprehensive way of addressing the problem of modeling facial deformation due to underlying movements of the speech organs is to simulate the biomechanical properties of skin tissues and of the musculo-skeletal systems.

Instead of geometric control parameters, facial movements are in this kind of model directly controlled by muscular activations that are supposed to be more directly connected to communicative intentions. Ekman and Friesen (1975, 1978) thus established the Facial Action Coding System (FACS) that describes facial expressions by means of 66 muscle actions.

Muscles apply forces to sets of geometric structures representing soft objects, in particular skin tissue. The simplest approach to skin tissue emulation is a collection of strings connected in a network (Platt and Badler, 1981) then organized in layers (Waters, 1987; Terzopoulos and Waters, 1990). These models

Figure 3. Terzopoulos' facial muscle model used by Ishikawa et al. (1998) has the 12 muscles in the forehead area and the 27 muscles in the mouth area. After scaling Terzopoulos' biomechanical model to the speaker's morphology (a), Ishikawa et al., estimate the activations of these muscles so as to reproduce an original facial expression (b), and The resulting 3D reconstruction is shown in (c) (from Ishikawa et al., 1998).

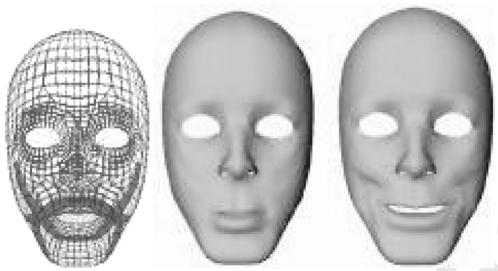


Figure 4. Joint action of the zigomatic and orbicularis oris muscles in a 3D biomechanical model of the facial tissue (from Chabanas and Payan, 2000). From left to right are figured the insertions of the muscles in the subcutaneous layer, the "neutral face" and the result of the joint action for producing a "smile".

distinguish typically three layers: epidermal, dermal and subcutaneous (muscular) layers as in Terzopoulos' model (see Fig. 3). The dermal layer has in charge of simulating the viscoeleastic properties of the skin and of propagating the different forces exerted to the subcutaneous layer by muscles to the visible epidermal layer. Transverse deformation modes, volume conservation or more complex deformation models such as finite-element modeling (see Fig. 4) are also considered. Muscles also connect the subcutaneous layer to the underlying bones (skull, hyoid, jaw . . .). Interaction between rigid structures and the skin model (notably necessary between teeth and mouth when spreading lips) should also be considered.

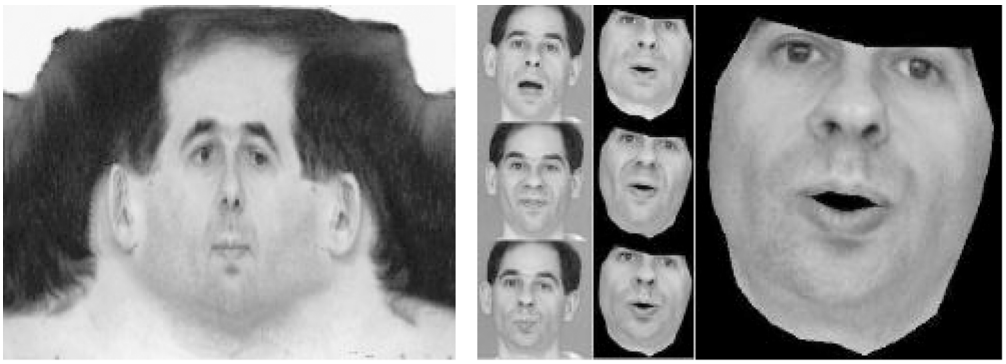
Although such models can potentially separate out the active contribution of muscular activation from the passive contribution of the skin tissues and of the musculo-skeletal structure to the resulting skin deformation, the dimensionality of the control space is very high compared to the degrees-of-freedom (DOF) of the facial geometry effectively used in the task. The muscular system is highly redundant and movements typically recruit a few dozen individual muscles whose actions need to be coordinated, sometimes in a very precise way (see Section 4.2).

2.4. *Videorealistic Rendering*

Such geometric or biomechanical models describe with more or less success the face shape variation according to geometric, articulatory or biomechanical commands. Facial surface variation is described by the 3D displacements of hundreds of points in a coordinate system bound to the skull. To synthesize a complete image of these facial movements, a facial appearance should be generated that accompanies facial movements: head movements as well as skin tissue deformation generate large but also subtle changes in the visual appearance of the face. As the position and normal at each facial fleshpoint changes, the illumination of that point changes. Large or small wrinkles can also appear or disappear according to facial movements.

Rendering procedures generally begin by defining a mesh connecting the facial points. Elementary triangles constituting the mesh are then colorized using different techniques: the simplest one consists of associating each vertex with a color and interpolating between the facial points using standard *shading* procedures (see Fig. 1). Videorealistic appearance could be obtained by applying *texture mapping*: A facial texture is typically obtained by identifying the position of the facial points on photographs of the speaker. Multiple views are typically collected and patched to obtain cylindrical textures (see Fig. 5(a)) that enable free head rotation.

We have demonstrated elsewhere (Revéret et al., 2000) that texture blending is also necessary to model the texture modification. This change of skin appearance is due for example to the appearing/disappearing nasogenian wrinkle—between lip corners and nose wings—when moving from a spread towards a rounded articulation ($[_a f_a]$ vs. $[_u p_u]$ in Fig. 5(b)): in this case, multiple textures are warped towards the target shape and blended according to the distance between the target shape and those from which textures have been extracted. A more general image-based rendering technique called *Statistical Appearance Model* (*SAM*) makes a more systematic use of all available training images (see Section 3.3).



(a) (b)

Figure 5. Texturing 3D models with videorealistic data. (a) A cylindrical texture that enables free head rotation. (b) Warping and blending multiple textures. Three original frames (from top to down, the speaker utters the three cardinal visemes [a], [$_{a}$f$_{a}$], [$_{u}$p$_{u}$]) are first warped towards the target shape then blended.

3. Image-Based Visual Synthesis

In the past decade, a series of new systems using image processing techniques has emerged. These systems consider how the color of each pixel in an image of the face changes according to the sound produced. These image-based systems have the potentiality to generate hyper-realistic images since large sets of natural videos are used and minimal image processing is performed. We will distinguish here between three "families" of systems: (a) systems that select appropriate segments of a large database and patch selected regions of the face on a background image; (b) systems that consider facial or head movements as displacements of pixels; (c) systems that also compute the change of the appearance of each pixel according to facial movements or speaker's appearance.

3.1. Overlaying Facial Regions

The most illustrative system involving the overlapping of facial regions is VideoRewrite (Bregler et al., 1997a): as seen in Fig. 6, sequences of mouth shapes are warped, roto-translated and overlaid with a background video. The warping stage smooths out concatenation artifacts. Then the mouth patch is rototranslated onto an insertion plane approximating the head orientation (see Fig. 7). This step is essential for collecting coherent mouth shapes at the training stage, especially when the blending between warped mouth shapes will be computed, and for the perceptual fusion between head and facial movements at synthesis time.

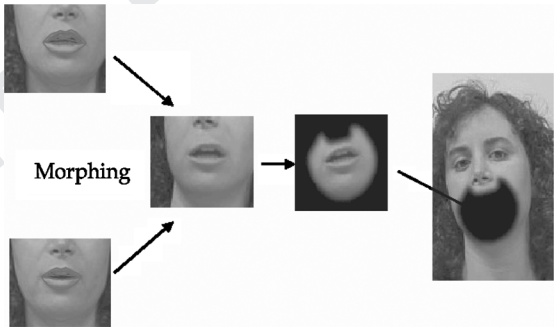


Figure 6. VideoRewrite consists in patching at the right position of the background image a mouth shape obtained by warping images of the database towards the target mouth shape and blending the results (from Bregler et al., 1997a).

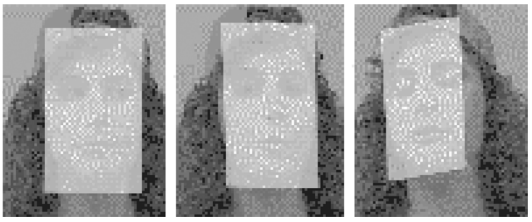


Figure 7. VideoRewrite estimates the best insertion planes for each head posture (from Bregler et al., 1997a).

Although this technique seems to be completely data-driven, VideoRewrite also uses an underlying parameterization of mouth shapes in the selection process: the selection of visual triphones uses dynamic programming where a distance term involves these underlying parameters. *Jaw lines* are also determined to

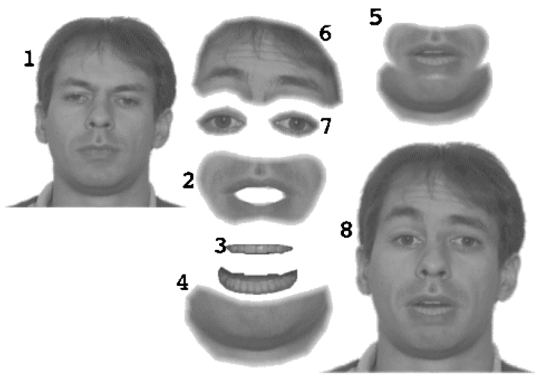


Figure 8. In the sample-based AT&T talking face, the head is decomposed into several facial parts. To generate a novel appearance, base head (1) is combined with mouth (5), eyes (7) and brows (6). The mouth area is generated by overlaying lips (2) on upper teeth (3) and lower teeth + jaw (4). This allows animating a jaw rotation independent of the lip shape. The result of the superposition of ((2), (3) and (4) is displayed in (5). Final image is shown in (8) (from Cosatto and Graf, 1997; Cosatto and Graf, 1998).

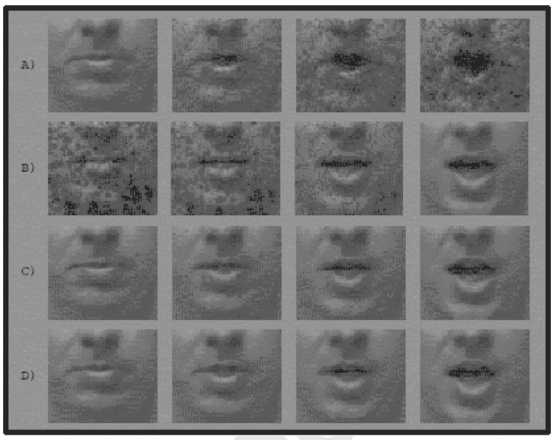


Figure 9. MikeTalk computes and blends two optical flows in order to symmetrically morph between two visemes A and B (respectively here [$_{a}$b$_{a}$] and [a]) to produce the intermediate images. Original frames are at the left of the first top row and at the right of the second one. From top to bottom: forward interpolation from A to B, backward interpolation from B to A, blending the two interpolations and filling the remaining holes.

obtain a realistic blending between the background video and mouth shapes.

This patching principle can also be applied to a more complete decomposition of the face. In the sample-based ATT Talking Face, Cosatto and Graf (1997, 1998) decompose the face into 5 regions (see Fig. 8) comprising the forehead, the eyes, the mouth, the upper teeth and the chin coupled with the lower-teeth if visible. Such a further decomposition reduces the number of parameters needed to describe each region which in turn could be controlled in an independent manner. It is therefore the responsibility of the control model to capture and restore the coordination between the control parameters of the different regions, while bigger regions have the advantage of maintaining coherence despite possible inaccurate estimation of optimal control parameters.

3.2. *Moving Pixels*

Instead of considering the deformation/movement of whole regions of the face, Ezzat (1998) tries to reproduce speech movements by computing displacements of pixels on the screen. MikeTalk computes an optical flow to find where each pixel of a source image projects/moves in a target image. Interpolation between two images A and B—visemes in the case of MikeTalk (see Section 4.1)—is performed by blending results of the optical flow computation from A to B and B to A. Any remaining "holes" in the interpolated images are filled using neighboring pixels (see Fig. 9).

More recently Ezzat et al. (2002) introduced a multidimensional morphable model (MMM) that uses also a basis of N image prototypes $\{I_i, i = 1 \ldots N\}$. Instead of considering only the warping between two images, any image I is characterized and could be re-synthesized from $2 * N$ barycentric parameters that describes pixel flow (α) and appearance change (β) in the image prototypes basis. Alternative methods for building complementary models of shape and appearance changes as well as linear models of these changes have also been proposed.

3.3. *Modeling Shape and Appearance*

A classical method in computer vision for tracking rigid or deformable objects in a complex scene is to build a statistical model of its shape and appearance from multiple views of the object. Medical images and faces are good examples of objects where changes of shape and appearance are complex and often subtle. For face recognition, Turk and Pentland (1991) introduced the *Eigenfaces* technique that uses Principal Components Analysis (PCA) to compute the eigenvectors of the covariance matrix of a training set of images.

When the eigenvectors are displayed, they look like a ghostly face, and are termed eigenfaces. The eigenfaces can be linearly combined to reconstruct any image with minimal distortion when using the first eigenface components.

Whereas eigenfaces model shape and appearance variations jointly but blindly, Statistical Appearance Models (SAM) (Cootes et al., 2001) represent shape and texture variations separately and then build a combined shape and appearance model which controls both shape and texture and takes into account the correlations between them. First a PCA is applied to the position of feature points to describe shape variation. Then *shape-free* image patches are obtained by warping each original image so that its feature points match the mean shape. Cylindrical textures (see Fig. 5(a)) are obvious examples of such shape-free images: where the face texture is projected onto a cylinder centered on the head axis. For SAM, all images of the database will be morphed to a single predefined mean configuration of a pre-defined mesh (see also Section 2.4). An eigenface decomposition is then further applied to these normalized face patches. Finally a combined appearance model is built by concatenating shape and texture eigenvectors and eliminating the correlations existing between them by applying a third PCA on these vectors. Shape and texture eigenvectors are of course weighted before applying PCA to compensate for unit scaling.

Active Appearance Models (AAM) using these SAM have been successfully applied to multi-speaker facial databases (Matthews et al., 2002). Theobald et al. (2001) use a set of 9431 greyscale images from one single speaker. As many as 31 first principal component vectors are necessary to explain 90% of the data variance.

Such a rendering technique can be extended to 3D by incorporating views from different viewpoints (Seitz and Dyer, 1996) or by a projection of a unique appearance to a 3D surface (Brooke and Scott, 1998).

4. Control Models

We consider here the problem of how coordination of control parameters of the various shape models proposed so far can be achieved and implemented in practical terms given actual trajectories to be reproduced. We will not address the problem of driving biomechanical models by realistic muscular activations (please refer to the discussion of the equilibrium hypothesis for speech in Perrier et al. (1996)).

4.1. Visemes

The basic control model for speech articulation consists in interpolating between a finite set of visual targets that can be mapped with the center of realizations of phonemes in context. Visemes can thus be defined as allophonic visual realizations of phonemes. Benoit and colleagues (1992) identified 21 visemes that constitute the "labial space" of the French speaker they analyzed. Although such a control strategy, maintaining the facial coherence in the vicinity of targets, is still used in quite a number of systems (especially in image-based synthesis—for example in MikeTalk), it does not take into account asynchronies between movement transitions of different articulators observed in natural speech, since a viseme constitutes per se a unique target posture for the underlying articulatory degrees-of-freedom of all speech articulators (jaw, lips . . .). It is however difficult to identify a unique target for each viseme in each parametric trajectory. One solution is to increase the number of such allophonic variations and increase the complexity of the rule-based control system; another solution is to use a more speech-specific coarticulation model.

4.2. Coarticulation Models

Instead of a nomenclature of all possible (visual) realizations of phonemes in context, coarticulation models specify algorithmically how context-independent targets are combined. The most popular system for driving parametric facial models is Cohen & Massaro's coproduction model (1993): control parameters for each context-independent target are blended spatially and temporally according to weighting factors for each phoneme considered.

Ohman's model (1967), originally applied to lingual coarticulation in occlusives, has also been applied successfully to facial data (Elisei et al., 2001). This model first identifies two groups of gestures on which the coarticulation will operate: a slowly varying vocalic gesture and rapid consonantal gestures that aim at producing certain constrictions given the underlying vocalic gestures. Consonants and vowels thus play asymmetrical roles in the coarticulation model: the vocalic gesture is computed first, then context-sensitive consonantal targets are computed as modulated deviations from the underlying vocalic gesture.

Note that most control models used for more general motor planning identify two or more different

representation spaces for motor planning and control (Bailly, 1998). They distinguish between the control space for movement planning, called the *distal* space, and the control parameters of the plant itself, the *proximal* space. Muscular activations are such proximal commands while lip geometry, coronal contact or even acoustic parameters (formants . . .) can be considered as distal targets. Such control models (Browman and Goldstein, 1990) require an inversion process able to deal with incomplete distal specification and some movement optimization such as minimum force, torque or jerk requirements.

4.3. N-Phones Models

In the approaches described above, parametric trajectories are essentially controlled by target interpolation using predefined transition functions. As video-based movement tracking and motion capture systems become more and more accessible, and video storage for post-processing can be envisaged, it is no longer necessary to use coarticulation models for extrapolating from a limited range of data.

The control parameters for whole trajectories can be stored in a segment dictionary, selected, retrieved and further processed before concatenation. So a new class of visual speech synthesis systems (Bregler et al., 1997b) exploits the same popular data-driven techniques as used for acoustic synthesis . . . and face the same problems of determining the optimal selection criteria and smoothing algorithms: dynamic programming is usually used to find an optimal path though candidate speech segments usually n-phones[1] given "concatenation" and "target" costs (Takeda et al., 1992). Selection criteria may however incorporate audiovisual costs. Moreover simple smoothing procedures can be applied with success to the already smooth articulatory gestures stored in the audiovisual segments: Fig. 10 gives an example of our audiovisual text-to-speech system based on the concatenation of multi-represented audiovisual diphones (diphones recorded in different contexts). The advantage of audiovisual segments is that they capture the intrinsic audiovisual correlations and asynchronies, but at the expense of truncated coarticulation.

Note the kinematic triphone model proposed by Okadome et al. (1999), where the kinematics of actual

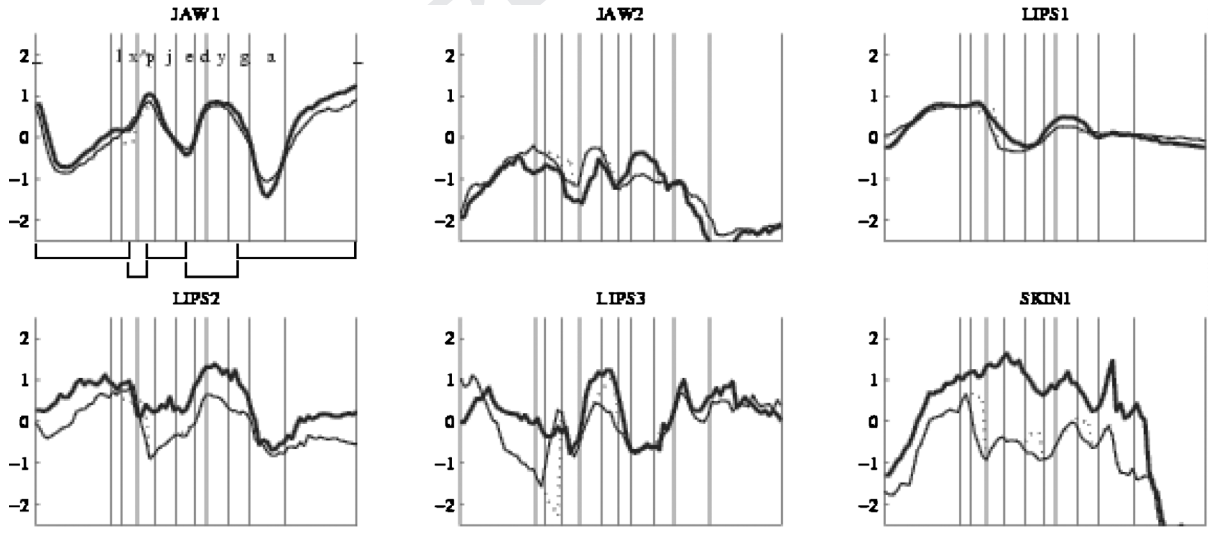


Figure 10. Generating articulatory trajectories using audiovisual diphones. Diphones are scaled and concatenated in order to produce specific sounds and specific sound durations. The target durations are here those of a short French utterance ("le pied du gars") uttered by the same female speaker who recorded the training corpus from which the diphones are extracted. The natural facial movements of the short utterance are plotted in thick lines. 8 diphones from 5 different utterances of the training corpus have been selected and concatenated (intervals at the bottom of the JAW1 caption). Vertical lines cue phoneme boundaries. Results of a pure concatenation is displayed in dotted lines. A simple smoothing procedure is however applied before concatenation (final result in plain lines) consisting in a gradual anticipation of the target right to the boundary by the trajectory within the demi allophone right to the boundary. Note the good match between natural and synthetic trajectories for essential articulatory parameters: jaw aperture (JAW1) and lips rounding/spreading (LIPS1). Natural and synthetic trajectories have been found equivalent in a perceptual test comparing different generation schemes (Bailly et al., 2002).

triphone articulatory tongue movements are characterized by the position and the first derivative of each parameter at each acoustic target of the triphone. Reconstruction is done using a minimum-acceleration constraint. Such a stylization simplifies the inter-triphone smoothing process while demonstrating good reconstruction of velocity profiles and parameter asynchronies (Shaiman and Porter, 1991).

4.4. Audiovisual Synchrony

Most audiovisual synthesis systems (post)synchronize the visual synthesizer with the acoustic synthesizer with via a minimal common input: a phonemic string with phoneme durations. This approach has some clear advantages such as the ability to couple two heterogeneous synthesis systems easily, or to feed visual synthesis with pure acoustic speech recognition results for "lip-sync" (Bregler et al., 1997a; Brand, 1999) that associates a synthetic animation with the natural acoustic input signal.

Such phoneme-driven control does not, however, guarantee a complete coherence of audiovisual signals, even when synthetic trajectories are obtained by stretching natural ones as mentioned in the preceding section. The lengthening of an allophone can be due to a decrease in speech rate, pre-boundary lengthening, lexical stress or emphatic accentuation: these different causes result in very different velocity profiles and thus in different kinematics.

The most obvious solution for ensuring coherent audiovisual kinematics is to record the acoustic signal and visual parameters synchronously. Then concatenative synthesis can be performed by selection of audiovisual segments (Hällgren and Lyberg, 1998; Minnis and Breen, 1998), using both segmental and suprasegmental criteria (see Fig. 10). An interesting approach is to train a Hidden Markov Model (HMM) with audiovisual stimuli (Brand, 1999; Tamura et al., 1999). Viterbi decoding of the resulting bimodal HMM will give the most probable set of visual parameters given the acoustic trace (Yamamoto et al., 1998).

5. Evaluation

Given that these systems and models have been presented to different scientific communities, it is very difficult to compare the achievements and evaluations of each technique. Most of the time, informal evaluation is performed, and very few evaluations involve direct comparison with "ground-truth" natural motion or video. Brand (1999) for example presented synthesized (via trained audiovisual HMM) versus real facial motion driving the same 3D model to seven observers and found no significant preference rates. However it is very difficult to sort out the relative influences of the quality of the control parameters, and of the unrealistic synthetic face with which observers were presented in Brand's study.

A more systematic evaluation was performed at ATT (Pandzic et al., 1999) on 190 subjects to show the benefit of audiovisual communication. The third experiment of this study aimed at comparing the *appeal* ratings for three different synthetic faces (see Fig. 11) uttering the same set of messages: (a) a standard flat 3D talking head, (b) a texture mapped 3D talking head and (c) a sample-based talking face. Subjects were not particularly seduced by synthetic faces: the best score was obtained by (a) while (c) obtained the worst rating. Surprisingly attempting to increase naturalness

Figure 11. The three talking faces tested by Pandzic et al. (1999).

resulted in inverse satisfaction. These results seem to contradict with the results of the first experiment evaluating the intelligibility of digits in noise where (a) and (c) performed equally well. However actual and estimated times to complete the task were both significantly higher for (c): sample-based faces seem thus to require more cognitive effort and more mental resources. This is also illustrated by the fact that, despite their long-standing experience of audiovisual perception and successful implementation of Baldi, Massaro recognizes that they "failed to replicate the prototypical McGurk[2] fusion effect" (Massaro, 1998a, p. 22), whereas they observed quite a number of combination /bga/ and /gba/ responses. Perceivers thus take into account the two channels of information, as evidenced in the reported performance of *coherent* audiovisual stimuli in noise, but the fusion of this information appears to be more difficult in the case of synthetic stimuli because of the incoherent or impoverished information provided by the two channels.

In most of these perception experiments however the relevance of the control parameters, the adequacy of deformation model and the liability of the rendering technique were tested altogether. It is therefore difficult to diagnose which module was the most deficient. Original *glass box* evaluation procedures and module-specific benchmarks should be proposed to address this problem (see for example (Bailly et al., 2002) for the evaluation of movement generation models).

6. From Data to Models: Cloning Speakers at ICP

As demonstrated by perception experiments on segmental (Pisoni, 1997) and suprasegmental (Ogden et al., 2000) aspects of acoustic synthetic speech, listeners are very sensitive to subtle details of the acoustic structure of speech signals. No doubt, observers also anchor their comprehension of visible speech on the coherence and subtlety of facial deformations induced by the underlying articulatory movements. We believe that this coherence could only be obtained by a careful and precise collection, comprehension and modeling of these articulatory movements and of the global interaction between movements and skin deformation. In fact, movements like lips protrusion or jaw oscillation produce deformation all over the face, while most model-based and image-based systems described above circumscribe influence of control parameters to a limited region using tapering or patching procedures on meshes or images. For example, very few models take into account that the nose wings clearly move during speech production and that some lingual and laryngeal movements have visible consequences on the cheeks and the throat.

Whatever the strategy adopted to render articulatory movements, there is a clear need for precise data on articulatory and geometric DOFs of the facial movements—at least for characterizing or labeling a database. In the following we describe briefly the ICP approach to facial animation that consists of cloning actual speakers: we build *speaker-specific linear shape and appearance models*. Speech gestures can be reliably reproduced by the additive influence of a few parameters (typically 6) that have a clear biomechanical interpretation.

6.1. Statistical Linear Shape Model

Motion capture devices (e.g. Qualisys, Vicon) offer greater and greater spatial and temporal resolution to recover, in real-time, the 3D positions of more and more pellets or beads glued on the subject's face. Although the animation industry now makes intensive use of these tracking systems for animating more and more realistic virtual creatures, research institutes still rely on the quality and efficiency of controlled experiments. Using a very simple photogrammetric method—previously used by Parke to build his initial model (Parke and Waters, 1996)—and up-to-date calibration procedures, we recorded 40 prototypical configurations of a French speaker whose face was marked with 166 glued colored beads (on the cheek, mouth, nose, chin and front neck areas), as depicted in Fig. 12.

Figure 12. Gathering fleshpoint positions using a photogrammetric method. Here the viseme [$_a$f$_a$] is shown.

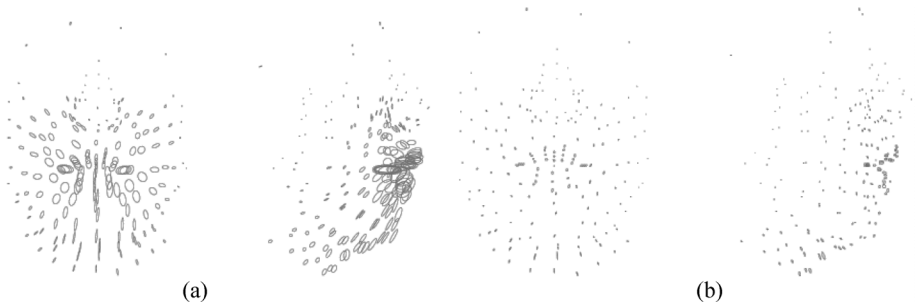


Figure 13. (a) Initial dispersion ellipses of the movements of 197 facial fleshpoints of a subject uttering a series of logatoms (isolated vowels, VCV . . .) spanning the visible articulatory space of its language (French here), (b) Residual dispersion when the contribution of 6 articulatory parameters are subtracted from the data. Each ellipsis is centered here on the mean position of the corresponding facial point (from Elisei et al., 2001).

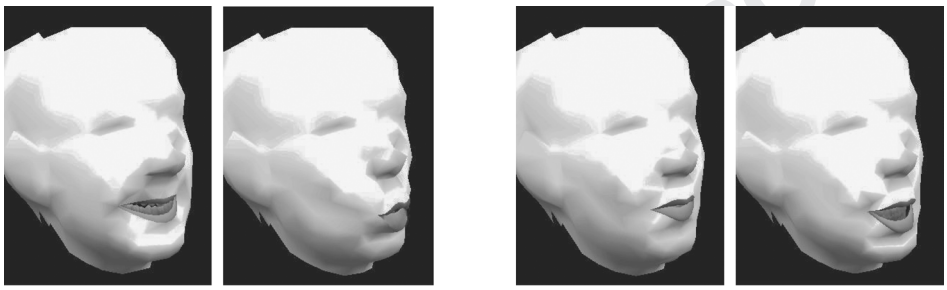


Figure 14. (a) The first statistically significant lip gesture for our French speaker: rounding/spreading. Note the accompanying movement of the nose wings. (b) The second one raises/lowers the lower lip (intrinsic lip movement that does not imply jaw. 6 such elementary gestures are combined (linear superposition) to reproduce any speech movement for that particular speaker.

In a coordinate system linked with the bite plane, every viseme is characterized by a set of 197 3D points including positions of the lower teeth and of 30 points characterizing the lip shape (for further details see Revéret et al., 2000; Elisei et al., 2001). Although these shapes have potentially $3 * 197 = 591$ geometric DOFs, we show that 6 DOFs already explain 97% of the variance of the data (see Fig. 13). Of course jaw opening, lip protrusion and lip opening are part of these DOFs, but more subtle parameters such as lip raising, jaw advance or independent vertical movements of the throat clearly emerge. These control parameters emerge from statistical analysis and their influence on facial deformation is additive. These parameters clearly influence independently the movements of the whole lower face (see for example the grooving of the nasogenian wrinkles and the expansion of the nose wings accompanying lip spreading in Fig. 14(a) and the grooving of the chin when pulling down the lower lip in Fig. 14(b)). These influences are sometimes subtle and are not always geometrically continuous, but should not be neglected. Although its crude linear assumptions do not take into account, for now, saturation due to tissue compression, this multilinear technique renders nicely the subtle interaction between speech organs and facial parts (such as formation of wrinkles or movements of the nose wings mentioned above).

6.2. *Statistical Linear Appearance Model*

Such a statistical linear shape model can easily be completed by a statistical appearance model (see Section 3.3). We use here a particular SAM, where the standard blind Principal Components Analysis (PCA) of the pixel colors of *shape-free* image patches is replaced by a multiple linear regression using the articulatory parameters of the linear shape model. We therefore consider only the changes of appearance strictly due to articulatory movements. Thus instead of decomposing and recomposing shape and appearance from raw image data by applying successive PCA, both shape

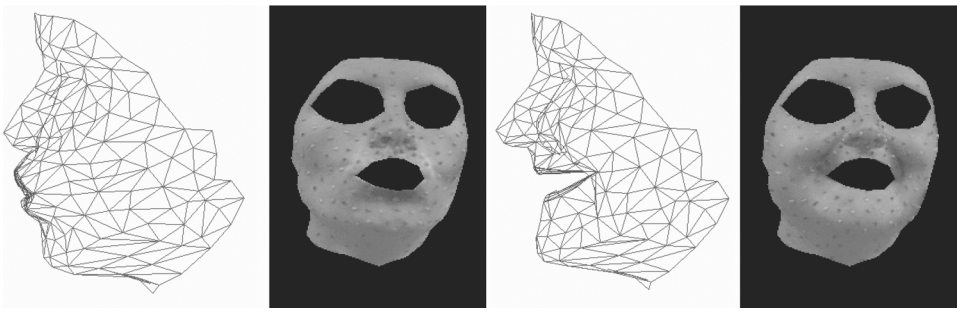


Figure 15. Shape and appearance associated with extreme variations along the first lip component (rounding/spreading) of the articulatory model for a female speaker. The appearance model has been trained using original motion capture data where colored beads were used.

and appearance are ruled by *common* articulatory parameters that have a clear biomechanical interpretation. Figure 15 shows for example the change of shape and appearance associated with the lip rounding gesture for one of our female speakers.

6.3. Controlling Articulatory Degrees-of-Freedom

Such joint linear shape and appearance models offer a unique articulatory control over videorealistic talking heads. A compromise between model-based and image-based techniques has been found that makes intensive uses of statistical analysis. Both shape and appearance models are data-driven. Fine shape and appearance changes due to articulatory movements are captured by using a dense facial mesh. Very few articulatory parameters are used to describe the elementary degrees-of-freedom of facial deformation caused by speech gestures. Typically 6 such parameters are sufficient for the four speakers studied so far. Although these speakers were uttering different language-specific visemes (we recorded a French, a German and an Arabic male speaker and a bilingual French/English female speaker), these parameters have similar effects on the face: these articulatory parameters are weakly correlated and clearly associated with true biomechanical movements and with phonetic/articulatory features. We thus expect this strategy to ease the task of the trajectory formation model. Moreover since the linear shape and appearance model captures the articulatory degrees-of-freedom of a human speaker, this model can be used in an analysis-by-synthesis process for estimating the articulatory parameters given a video source (Revéret et al., 2000; Elisei et al., 2001; Odisio et al., to appear). "Ground truth" articulatory trajectories can thus be collected on large video corpora. They can then be used to tune speech-specific coarticulation models or stored in an audiovisual database (Bailly et al., 2002) for concatenative synthesis that, while not production-based, could be termed production-aware.

Au: Pls. update Elisei et al.

6.4. Towards a Generic Talking Face

The parameters of all our speaker-specific models have a common semantics: open/close or advance/retract jaw, spread/round lips . . . The way and the extend they affect face shape is speaker-dependent but their number and their main actions is universal since we share the facial musculo-skelettal structure. Speakers and languages differ in the way they exploit and synchronize these elementary gestures. We can thus use PARAFAC analysis (Harshman and Lundy, 1984) or more directly multilinear regression to determine the speaker's specific scaling of these universal commands.

Prior to this analysis, each speaker-specific shape model should be characterized not only by the same number of commands but also drive the same mesh structure with the same number of vertices. Now the number of fleshpoints recorded during a motion-capture session (see Fig. 12) is limited to a few hundred and do not entirely cover the whole head. Using a modified mesh-matching algorithm (Couteau et al., 2000), we are able to scale a generic *high-definition* talking face (see Fig. 16(a)) to the *low-resolution* surface defined by the fleshpoints characterizing each viseme of a session (see Fig. 16(b)). An example of the result of a 3D to 3D matching is shown Fig.16(c).

Simply using parameters of the *low-resolution* motion-capture data as linear predictors of the deformation of the *high-definition* mesh sketches the first step towards a generic talking face where conformation and animation parameters (analogue to the MPEG-4 FDP

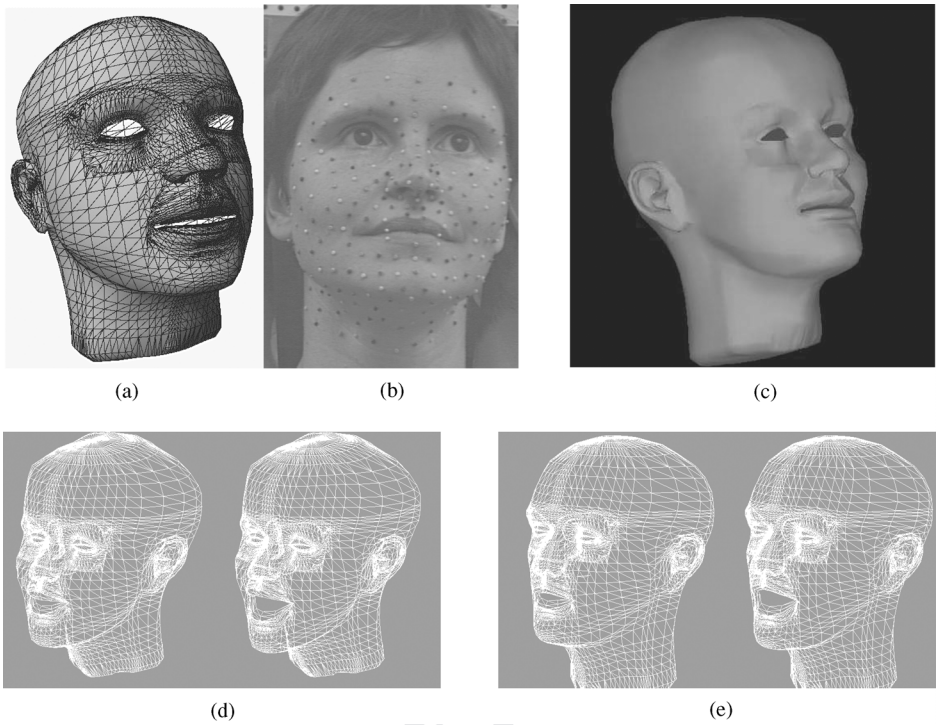


Figure 16. Building a generic talking face. Using an original 3D to 3D matching algorithm (Couteau et al., 2000) a unique generic "high definition" face mesh (a) (here developed by Pighin et al., 1998) is scaled to multiple "low definition" motion capture data from one female speaker: (c) shows the resulting geometry for the viseme shown in (b). A "high definition" articulated clone for each speaker is then developed: (d) shows the shape deformation resulting from setting to +1 the "jaw opening" parameter. (e) same for a male speaker. All shape models are so characterized by the same mesh, have quasi-identical fleshpoints and are driven by comparable parameters.

and FAP) are separated out. Figure 16(d) and (e) show for example the action of the *speaker-independent* jaw rotation parameter on two *speaker-specific* models.

7. Conclusions

The animation industry clearly drives the progress in facial animation and we should draw some lessons from its history. The panel session on facial animation at Siggraph'97, which involved the participation of such notable researchers as D. Terzopoulos, M. Cohen, F. Parke, D. Sweetland and K. Waters, discussed almost exclusively model-based approaches. Most of the speakers expressed a need for more data acquisition facilities, and a reliance on the progress of models incorporating true biomechanics and aerodynamics. Is this still true?

We may draw a (pessimistic?) parallel between facial animation and the field of acoustic speech synthesis, where data-driven techniques tend to question the need for more comprehensive models of speech production or intonation. We do however believe that comprehensive models could be tuned from ground-truth data using a careful use of statistical analysis driven and constrained by the a priori knowledge of biomechanical and aeroacoustics phenomena governing speech production.

Terzopoulos concluded his discussion: "An intriguing avenue for future work is to develop brain and perception models that can imbue artificial faces with some level of intelligent behavior", while Waters added: "As the realism of the face increases, we become much less forgiving of imperfections in the modeling and animation: If it looks like a person we expect it to behave like a person ..." Evidence suggests that our brains are even "hard-wired" to interpret facial images. If cartoons can use characters that have non-human characteristics, such as dogs, cats, ants or monsters, to speak, we are compelled to address these perception issues and revise our evaluation criteria. We may need to do so for acoustic synthesis as well. We do strongly believe that current intelligibility tests are insufficient

for estimating the cognitive load placed on the subject when perceiving synthetic audiovisual stimuli.

Acknowledgments

This review benefited from input from our colleague P. Badin. We thank A. Breen, T. Ezzat, Y. Payan, and M. Slaney for providing information about their systems. M. Tabain did not proofread this last paragraph but we learnt quite a lot from her correction of the previous ones. We also acknowledge the perspicacious comments made by C. Shadle and two other anonymous reviewers. In collaboration with Yohan Payan (TIMC), Maxime Bérar is responsible for adapting the 3D to 3D matching algorithm (provided by Praxim SA) for building high definition speaking heads.

Notes

1. These segments are typically *n*-phones: speech segments that start with the center of realization of one phoneme to the center of realization of the *n*th next phoneme. Diphones or triphones are usually used. Besides characteristics of the speech signal (pitch, inter-phones boundaries . . .) and the signal itself, accompanying information can also be stored such as video or trajectories of facial parameters (as in Bregler et al., 1997a; or in Okadome et al., 1999).
2. The McGurk effect (McGurk and MacDonald, 1976) involves a situation in which an auditory /ba/ is paired with a visible /ga/ and the perceiver reports hearing /da/.

References

Badin, P., Borel, P., Bailly, G., Revéret, L., Baciu, M., and Segebarth, C. (2000). Towards an audiovisual virtual talking head: 3D articulatory modeling of tongue, lips and face based on MRI and video images. *Proceedings of the 5th Speech Production Seminar*, Germany: Kloster Seeon, pp. 261–264.

Bailly, G. (1998). Learning to speak. Sensori-motor control of speech movements. *Speech Communication*, *22*(2/3):251–267.

Bailly, G., Gibert, G., and Odisio, M. (2002). Evaluation of movement generation systems using the point-light technique. *IEEE Workshop on Speech Synthesis*, Santa Monica, CA.

Benoît, C., Lallouache, T., Mohamadi, T., and Abry, C. (1992). A set of French visemes for visual speech synthesis. In G. Bailly and C. Benoît (Eds.), *Talking Machines: Theories, Models and Designs*. Elsevier B.V., pp. 485–501.

Bergeron, P. and Lachapelle, P. (1985). Controlling facial expression and body movements in the computer-generated short "Tony de Peltrie". *SIGGRAPH, Advanced Computer Animation Seminar Notes*, San Francisco, CA.

Beskow, J. (1995). Rule-based Visual Speech Synthesis. Madrid, Spain, Eurospeech, pp. 299–302.

Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.-E., and Öhman, T. (1997). The Teleface project—multimodal speech communication for the hearing impaired. Rhodos, Greece: Eurospeech, 2003–2010.

Brand, M. (1999). Voice pupperty. *SIGGRAPH'99*, Los Angeles, CA, pp. 21–28.

Bregler, C., Covell, M., and Slaney, M. (1997a). VideoRewrite: Driving visual speech with audio. *SIGGRAPH'97*, Los Angeles, CA, pp. 353–360.

Bregler, C., Covell, M., and Slaney, M. (1997b). Video rewrite: Visual speech synthesis from video. *International Conference on Auditory-Visual Speech Processing*, Rhodes, Greece, pp. 153–156.

Brooke, N.M. and Scott, S.D. (1998). Two- and three-dimensional audio-visual speech synthesis. *International Conference on Auditory-Visual Speech Processing*, Terrigal, Australia, pp. 213–218.

Browman, C.P. and Goldstein, L.M. (1990). Gestural specification using dinamically-defined articulatory structures. *Journal of Phonetics*, *18*(3):299–320.

Chabanas, M. and Payan, Y. (2000). A 3D finite element model of the face for simulation in plastic and maxillo-facial surgery. *International Conference on Medical Image Computing and Computer-Assisted Interventions*, Pittsburgh, USA, pp. 1068–1075.

Cohen, M.M. and Massaro, D.W. (1993). Modeling coarticulation in synthetic visual speech. In D. Thalmann and N. Magnenat-Thalmann (Eds.), *Models and Techniques in Computer Animation*. Springer-Verlag: Tokyo, pp. 141–155.

Cootes, T.F., Edwards, G.J., and Taylor, C.J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(6):681–685.

Cosatto, E. and Graf, H.P. (1997). Sample-based synthesis of photo-realistic talking-heads. *SIGGRAPH'97*, Los Angeles, CA, pp. 353–360.

Cosatto, E. and Graf, H.P. (1998). Sample-based of photo-realistic talking heads. *Computer Animation*, Philadelphia, Pennsylvania, pp. 103–110.

Couteau, B., Payan, Y., and Lavallée, S. (2000). The Mesh-Matching algorithm: An automatic 3D mesh generator for finite element structures. *Journal of Biomechanics*, *33*(8):1005–1009.

Doenges, P., Capin, T.K., Lavagetto, F., Ostermann, J., Pandzic, I., and Petajan, E. (1997). MPEG-4: audio/video and synthetic graphics/audio for real-time, interactive media delivery. *Image Communications Journal*, *9*(4):433–463.

Eisert, P. and Girod, B. (1998). Analyzing facial expressions for virtual conferencing. *IEEE Computer Graphics & Applications: Special Issue: Computer Animation for Virtual Humans*, *18*(5):70–78.

Ekman, P. and Friesen, W.V. (1975). *Unmasking the Face*. Palo Alto, California: Consulting Psychologists Press.

Ekman, P. and Friesen, W. (1978). *Facial Action Coding System (FACS): A Technique for the Measurement of Facial Action*. Palo Alto, California: Consulting Psychologists Press.

Elisei, F., Odisio, M., Bailly, G., and Badin, P. (2001). Creating and controlling video-realistic talking heads. *Auditory-Visual Speech Processing Workshop*, Scheelsminde, Denmark, pp. 90–97.

Ezzat, T. and Poggio, T. (1998). *MikeTalk: A Talking Facial Display Based on Morphing Visemes*. Philadelphia, PA: Computer Animation, pp. 96–102.

Ezzat, T., Geiger, G., and Poggio, T. (2002). Trainable videorealistic speech animation. *ACM Transactions on Graphics*, *21*(3):388–398.

Hällgren, Å. and Lyberg, B. (1998). Visual speech synthesis with concatenative speech. *Auditory-Visual Speech Processing Conference*, Terrigal-Sydney, Australia, pp. 181–183.

Harshman, R.A. and Lundy, M.E. (1984). The PARAFAC model for three-way factor analysis and multidimensional scaling. In H.G. Law, C.W. Snyder, J.A. Hattie, and R.P. MacDonald (Eds.), *Research Methods for Multimode Data Analysis*. New-York: Praeger, pp. 122–215.

Ishikawa, T., Sera, H., Morishima, S., and Terzopoulos, D. (1998). Facial image reconstruction by estimated muscle parameter. *International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, pp. 342–347.

Li, H., Roivanen, P., and Forchheimer, R. (1993). 3D motion estimation in model-based facial image coding. *IEEE Transactions on PAMI*, *15*(6):545–555.

Massaro, D. (1998a). Illusions and issues in bimodal speech perception. *Auditory-Visual Speech Processing Conference*, Terrigal, Sydney, Australia, pp. 21–26.

Massaro, D.W. (1998b). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press.

Matthews, I., Cootes, T.F., and Bangham, J.A. (2002). Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(2):198–213.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *26*:746–748.

Minnis, S. and Breen, A.P. (1998). Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis. *ICSLP*, Beijing, China, pp. 759–762.

Odisio, M., Elisei, F., Bailly, G., and Badin, P. (to appear). 3D talking clones for virtual teleconferencing. *Annals of Telecommunications*.

Au: Pls. update.

Ogden, R., Hawkins, S., House, J., Huckvale, M., Local, J., Carter, P., Dankovicová, J. and Heid, S. (2000). ProSynth: An integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Computer Speech and Language*, *14*(3):177–210.

Öhman, S.E.G. (1967). Numerical model of coarticulation. *Journal of the Acoustical Society of America*, *41*:310–320.

Okadome, T., Kaburagi, T., and Honda, M. (1999). Articulatory movement formation by kinematic triphone model. *IEEE International Conference on Systems Man and Cybernetics*, Tokyo, Japan, pp. 469–474.

Olives, J.-L., Möttönen, R., Kulju, J., and Sams, M. (1999). Audio-visual speech synthesis for finnish. *Auditory-Visual Speech Processing Workshop*, Santa Cruz, CA, pp. 157–162.

Pandzic, I., Ostermann, J., and Millen, D. (1999). Users evaluation: Synthetic talking faces for interactive services. *The Visual Computer*, *15*:330–340.

Parke, F.I. (1972). Computer generated animation of faces. *ACM National Conference*, Salt Lake City, pp. 451–457.

Parke, F.I. (1975). A model for human faces that allows speech synchronized animation. *Journal of Computers and Graphics*, *1*(1):1–4.

Parke, F.I. (1982). A parametrized model for facial animation. *IEEE Computer Graphics and Applications*, *2*(9):61–70.

Parke, F.I. and Waters, K. (1996). *Computer Facial Animation*. Wellesley, MA, USA, A.K. Peters.

Perrier, P., Ostry, D.J., and Laboissière, R. (1996). The equilibrium point hypothesis and its application to speech motor control. *Journal of Speech and Hearing Research*, *39*:365–377.

Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., and Salesin, D.H. (1998). Synthesizing realistic facial expressions from photographs. *Proceedings of Siggraph*, Orlando, FL, USA, pp. 75–84.

Pisoni, D.B. (1997). Perception of synthetic speech. In J.P.H.V. Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg (Eds.), *Progress in Speech Synthesis*. Springer Verlag: New York. pp. 541–560.

Platt, S.M. and Badler, N.I. (1981). Animating facial expressions. *Computer Graphics*, *15*(3):245–252.

Pockaj, R., Costa, M., Lavagetto, F., and Braccini, C. (1999). MPEG-4 facial animation: An implementation. *International Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging*, Santorini, Greece, pp. 33–36.

Revéret, L., Bailly, G., and Badin, P. (2000). MOTHER: A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. *International Conference on Speech and Language Processing*, Beijing, China, pp. 755–758.

Rydfalk, M. (1987). CANDIDE, a parameterized face. Sweden, Dept. of Electrical Engineering, Linköping University: LiTH-ISY-I-866.

Seitz, S.M. and Dyer, C.R. (1996). View morphing. *ACM SIGGRAPH*, New Orleans, Louisiana, pp. 21–30.

Shaiman, S. and Porter, R.J. (1991). Different phase-stable relationships of the upper lip and jaw for production of vowels and diphthongs. *Journal of the Acoustical Society of America*, *90*:3000–3007.

Takeda, K., Abe, K., and Sagisaka, Y. (1992). On the basic scheme and algorithms in non-uniform unit speech synthesis. In G. Bailly and C. Benoît (Eds.), *Talking Machines: Theories, Models and Designs*. Elsevier B.V., pp. 93–105.

Tamura, M., Kondo, S., Masuko, T., and Kobayashi, T. (1999). Text-to-audio-visual speech synthesis based on parameter generation from HMM. *European Conference on Speech Communication and Technology*, Budapest, Hungary, pp. 959–962.

Tekalp, A.M. and Ostermann, J. (2000). Face and 2-D Mesh animation in MPEG-4. *Signal Processing: Image Communication*, *15*:387–421.

Terzopoulos, D. and Waters, K. (1990). Physically-based facial modeling, analysis and animation. *The Journal of Visual and Computer Animation*, *1*:73–80.

Theobald, B.J., Bangham, J.A., Matthews, I., and Cawley, G.C. (2001). Visual speech synthesis using statistical models of shape and appearance. *Auditory-Visual Speech Processing Workshop*, Scheelsminde, Denmark, pp. 78–83.

Tsai, C.-J., Eisert, P., Girod, B., and Katsaggelos, A.K. (1997). Model-based synthetic view generation from a monocular video sequence. *Proceedings of the International Conference on Image Processing*, Santa Barbara, California, pp. 444–447.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1):71–86.

Vignoli, F. and Braccini, C. (1999). A text-speech synchronization technique with applications to talking heads. *Auditory-Visual Speech Processing Conference*, Santa Cruz, California, USA, pp. 128–132.

Waters, K. (1987). A muscle model for animating three-dimensional facial expression. *Computer Graphics*, *21*(4):17–24.

Waters, K. and Terzopoulos, D. (1992). The computer synthesis of expressive faces. *Philosophical Transactions of the Royal Society of London (B)*, *335*:87–93.

Yamamoto, E., Nakamura, S., and Shikano, K. (1998). Lip movement synthesis from speech based on Hidden Markov Models. *Speech Communication*, *26*(1–2):105–115.