AUDIOVISUAL SPEECH SYNTHESIS. FROM GROUND TRUTH TO MODELS

G. Bailly

Institut de la Communication Parlée, UMR CNRS n°5009, INPG/Univ. Stendhal
46, av. Félix Viallet, 38031 Grenoble Cedex, France
bailly@icp.inpg.fr

ABSTRACT

We present here the main approaches used to synthesize and drive talking faces. Illustrative systems are described. We 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. We then focus on the necessity to capture, model and render with maximum fidelity the intimate coherence of the facial deformations observed on a human face.

1 INTRODUCTION

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 *image-based* systems generate videorealistic animations using simple image processing techniques. This evolution, surprisingly, parallels the evolution of acoustic synthesis, where corpus-based synthesis tends to wipe out decades of research on parametric (articulatory then formant) synthesis. The more direct link between articulation and facial deformation, compared to acoustics, together with the need for giving the gift of speech to virtual non-human creatures help, in case of facial animation, to maintain a balance between the two approaches.

Another key feature of current virtual animations is the quasi general use of motion capture: instead of a hopeless quest of comprehensive laws underlying the biological movements of animals or persons according to the diverse tasks a given organ can accomplish, movements are typically captured on a living actor and scaled to the target virtual creature.

We will first describe some of the main features of these two approaches, trying to distinguish between control and graphic rendering of the face. Then we will comment on the few evaluation results 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 and appearance that take into account underlying articulatory control of the musculo-skeletal system.

2 MODEL-BASED VISUAL SYNTHESIS

The models that will be presented in this section have in common the aim to reproduce 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 [30, 31] software and his particular 3-D talking head. This line of models should be classified as terminal-analog synthesizers in the sense that do not aim at understanding the underlying physiological mechanisms that produce the speech signals and the facial deformations, but only attempt to reproduce them in geometrical terms. We will first describe briefly such a *geometric* approach and then mention some *biomechanical* models of speech articulators that are under development.

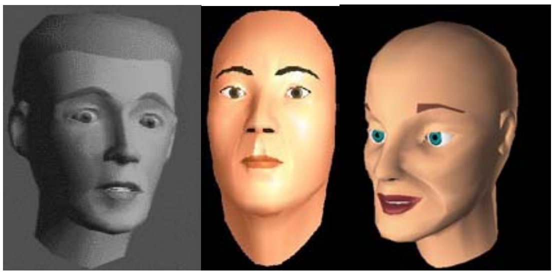

Figure 1: *A gallery of Parke's descendents. From left to right: Sven from KTH, Baldi from PSL, the LCE talking head.*

2.1 *Parke's descendants*

PSL's Baldi [23], KTH's [6, 7] and LCE's [28] Talking Heads, are all 3D computer graphic objects defined by a set of 3D 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 (see Figure 1). Such articulated meshes are often used as generic models in model-based movement tracking systems [17, 42].

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 [16, 36]. The 3D coordinates of the 84 Feature Points (FPs) are controlled by a set of 68 FAPs (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)" [36, p. 33].

2.2 *Control dimensionality*

In the previous models, *geometric* degrees-of-freedom of some characteristic *FPs* of the 3D meshes were considered. Three main problems arise when piloting mesh deformations from such FPs: (a) 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 MEPG4 by the laconic instruction associated with FAP3 *open_jaw* “does not affect mouth opening” [40, p.412]. (b) Most FAPs are low level, and do not take into account speech-specific gestures, which led Vignoli & Braccini [43] to add another layer of control parameters, called APs (Articulatory Parameters), corresponding to mouth height, mouth width, protrusion and jaw rotation, that control the FAPs. (c) Although these APs constitute a more comprehensive set of *articulatory* degrees-of-freedom, they do not solve the problem of extrapolating the movement of tens of vertices starting from the displacement of a single feature point. (d) The influence of each FAP - or even of more realistic action units of the FACS (see below) - on the face generates wrinkles and unforeseen collateral deformations (such as deformation of nose wings) that most ad hoc sculpting procedures forget.

Instead of ad hoc tapering or shape interpolation, we have proposed elsewhere [1, 20, 37] to define APs as *articulatory* degrees-of-freedom delivered by a guided statistical analysis of 3D coordinates of hundreds of facial fleshpoints (see §6).

Figure 2: *Gathering flashpoint positions using a photogrammetric method. Here 245 colored beads have been glued on the subject’s face.*

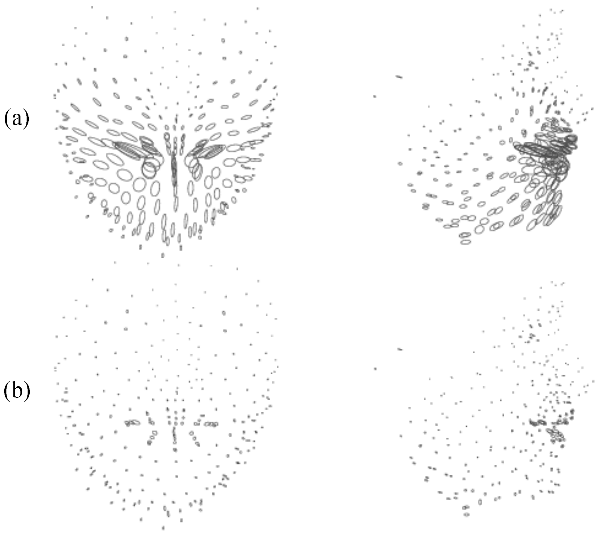

Figure 3: *Dispersion ellipses of the movements of facial fleshpoints of a female subject uttering 56 French visemes: (a) .raw data; (b) after removing the contribution of the articulatory parameters.*

2.3 *Skin and muscle-based facial animation*

Vertices of the previous 3D meshes can in fact be considered as fleshpoints. 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 here directly controlled by muscular activations that are supposed to be more directly connected to communicative intentions. Ekman and Friesen [18, 19] 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 [35] then organized in layers [41, 44]. Instead of the infinitesimally thin surface with no underlying structure considered in geometric models and the simplest muscle-based models, the facial mesh is organized in layers - typically three: epidermal, dermal and subcutaneous (muscular) layers as in Waters & Terzopoulos models - where transverse deformation modes, volume conservation or more complex deformation models such as finite-element modeling [11] are 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 compare 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 §4.2).

2.4 *Videorealistic rendering*

In order to render these facial movements on a computer screen, a facial appearance should be generated: head movements as well as skin tissue deformation generates 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 consists in defining a mesh connecting the facial points. Elementary triangles constituting the mesh are then colorized using different techniques: the most simple consists in associating each vertex with a color and interpolating between the facial points using standard *shading* procedures (see Figure 1). Videorealistic appearance could be obtained by applying texture morphing: 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 that enables free head rotation.

We have demonstrated elsewhere [37] that texture blending is also necessary to model the texture modification (for example see the nasogenian wrinkle when moving from a spread towards a rounded articulation in Figure 4): in this case, multiple textures are morphed 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 morphing/blending technique called *Statistical Appearance Model* (*SAM*) uses more systematically all available training images (see Figure 4).

3 IMAGE-BASED VISUAL SYNTHESIS

In the past decade, a series of new systems based on more simple 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 minimal image processing is performed on large sets of natural videos. We will distinguish here between two “families” of systems: (a) systems consisting in selecting appropriate segments of a large database and patching selected regions of the face on a background image; (b) systems that consider facial or head movements as displacements of pixels; (c) .

3.1 *Overlaying facial regions*

The most illustrative system involving the overlapping of facial regions is VideoRewrite [9]: sequences of mouth shapes are morphed, roto-translated and overlaid with a background video. The morphing smoothes out concatenation artifacts. Then the mouth patch is morphed onto an insertion plane approximating the head orientation. This step is essential (a) for collecting coherent mouth shapes at the training stage, especially when the blending between morphed mouth shapes will be computed and (b) for the perceptual fusion between head and facial movements at synthesis time.

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 while *jaw lines* should be determined to obtain a realistic blending between the background video and mouth shapes.

The VideoRewrite principle can also be applied to a more complete decomposition of the face. In the sample-based ATT Talking Face [14, 15], Cosatto & Graf decompose the face into 6 regions comprising the eyes, the mouth, the teeth and the chin. 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, MikeTalk [21] 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 - 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 patched using neighboring pixels. Inter-viseme optical flows can be cumulated and a further model of optical flow deformation can also be evaluated using Principal Components Analysis (PCA). First components have a clear articulatory interpretation in terms of jaw and lip movements. Moreover fine details such as lip raising movements as required for the production of labiodentals emerge clearly from the data, showing the excellent and precise job made by the computation of the optical flow.

4 CONTROL MODELS

We will consider here the problem of how coordinative structures of control parameters can be implemented in practical terms given actual trajectories to be reproduced. We will not address the problem of how muscular activations actually drive the articulators (please refer to the discussion of the equilibrium hypothesis for speech in [33]).

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 [4] 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. Consequently it is sometimes difficult to identify a unique target for each viseme in each parametric trajectories. One solution is to increase the number of such allophonic variations and increase the complexity of the rule-based control system or 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 [13]: control parameters for each context-independent target are blended spatially and temporally according to weighting factors for each phoneme considered.

Ohman’s model [26], originally applied to lingual coarticulation in occlusives, has also been applied successfully to facial data [20]. 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 [2]. 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 or coronal contact can be considered as distal targets. Such control models [10] 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 *Triphone models*

In the previous approaches, 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.

Whole control trajectories can be stored into segment dictionaries, selected, retrieved and further processed before concatenation. So a new class of visual speech synthesis systems [Bregler, 1997 #1466] exploit 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.

Note the kinematic triphone model proposed by Okadome et al [27], where the kinematics of actual 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 [38].

4.4 Audiovisual synchrony

Most audiovisual synthesis systems (post)synchronize an acoustic synthesizer with the visual synthesizer via a minimal common input: a phonemic string with phoneme durations. This approach has some clear advantages such as the ability to easily couple two heterogeneous synthesis systems, or to feed visual synthesis with pure acoustic speech recognition results for "lip-sync" [8, 9].

Such phoneme-driven control does not, however, guaranty 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 multiple causes result in very different velocity profiles and thus in different kinematics.

The most evident solution for ensuring coherent audiovisual kinematics is to record synchronously the acoustic signal and visual parameters. Then concatenative synthesis can be performed using selection of audiovisual segments [24], using both segmental and suprasegmental criteria. An interesting approach is to train a Hidden Markov Model (HMM) with audiovisual stimuli [8, 39]. Viterbi decoding of the resulting bimodal HMM will give the most probable set of visual parameters given the acoustic trace [45]. More simple audio-to-acoustics models can also be investigated as proposed by the ATR team [9, 46]

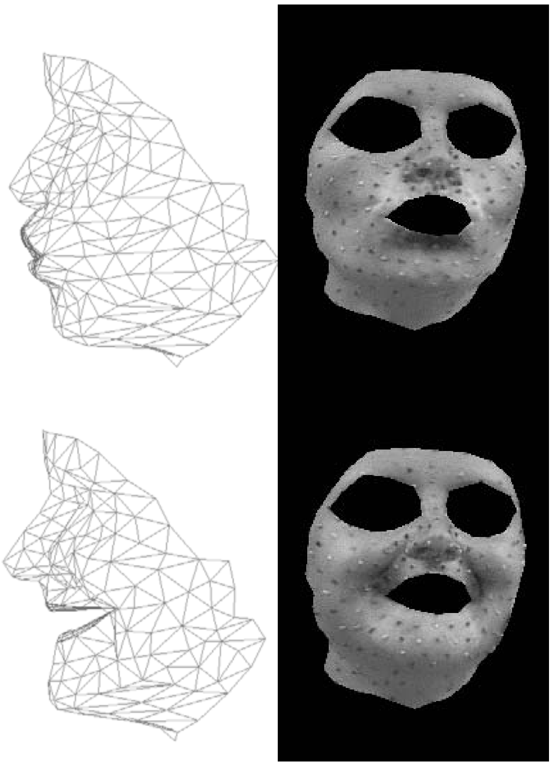

Figure 4: *Shape and appearance associated with extreme variations along the first lip component of the articulatory model of a French female speaker. The appearance model has been trained using original motion capture data where colored beads were used. Note the clear grooving of the cheeks and disappearance of the nasogenian wrinkle when protruding.*

5 EVALUATION

Given that these systems and models have been presented to different scientific communities, its is very difficult to compare the achievements and evaluations of each technique. Most of the time, informal evaluation is performed (with noticeable exceptions [5, 12]), and few evaluations involve direct comparison with "ground-truth" natural motion or video. Brant [8] 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 the unrealistic synthetic face with which observers were presented in Brant's study.

A more systematic evaluation was performed at ATT [29] 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 driven by the sample synthetic audiovisual control parameters: (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 resulted in inverse satisfaction. These results seem to contradict 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 fusion effect" [22, 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 should be more difficult in the case of synthetic stimuli because of the incoherent or impoverished information provided by the two channels.

In both 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 [3] for the evaluation of movement generation models).

6 MODELS AND DATA

As demonstrated by perception experiments on segmental [34] and suprasegmental [25] 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 lip 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 move clearly during speech production and that some lingual and laryngeal movements have visible consequences. So for our last female speaker, the distance between nose wings has a maximum variation of 3 mm whereas the distance variation between beads fixed at the nostrils bases reaches 4.8 mm. Similarly, maximal variation of her lowest fleshpoint (in the area of the mid throat) is 1.75 cm that is already one third of the maximal variation of the distance between lip corners (4.2 cm)!

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.

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 (from *Final Fantasy* to *Shrek* or *Monsters*), 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 [32] - and up-to-date calibration procedures, we recorded 120 prototypical configurations of a French speaker whose face was marked with 245 glued colored beads (on the cheek, mouth, nose, chin and front neck areas), as depicted in Figure 2. In a coordinate system linked with the bite plane, every viseme is thus characterized by a set of 245 3D points including positions of the lower teeth and of 30 points characterizing the lip shape (for further details see [20, 37]). Although these shapes have potentially 3*245 = 735 geometric DOFs, we show that 6 DOFs already explain 97% of the variance of the data. 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. This influence is sometimes subtle and is sometimes not continuous in geometry, 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, cheeks grooving or movements of the nose wings mentioned above) both for the facial shape and appearance.

7 CONCLUSIONS

Whatever potential vocations this paper may have generated in the audience during the conference or among its readers, 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 call still true? We may draw a (pessimistic?) parallel with results in speech research, where data-driven techniques tend to question the need for more comprehensive models of speech production or intonation.

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, for human speaking faces, to address these perception issues and revise –for pure audio stimuli also?- our evaluation criteria.

We do suggest that this evaluation should be able to separate out the contribution of (a) the control model that computes parametric trajectories from phonetic input, (b) the shape model that specifies how face geometry is affected by the parameters and (c) the appearance model that carries out the final image rendering.

ACKNOWLEDGMENTS

An early version of this paper was presented at the ITRW on Speech Synthesis in Scotland. An extended version has been submitted at the International Journal of Speech Technology. This review benefited from input from my colleagues P. Badin, F. Elisei and M. Odisio. I thank A. Breen, T. Ezzat, Y. Payan, and M. Slaney for providing information about their systems. M. Tabain did not proofread this version but I learned quite a lot from her correction of the previous ones.

BIBLIOGRAPHY

[1] 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*. in *Proceedings of the 5th Speech Production Seminar*. Kloster Seeon - Germany. p. 261-264.

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

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

[4] Benoît, C., Lallouache, T., Mohamadi, T., and Abry, C. (1992) *A set of French visemes for visual speech synthesis*, in *Talking Machines: Theories, Models and Designs*, G. Bailly and C. Benoît, Editors. Elsevier B.V. p. 485-501.

[5] Benoît, C. and Le Goff, B. (1998) *Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP.* Speech Communication, **26**: p. 117-129.

[6] Beskow, J. (1995) *Rule-based Visual Speech Synthesis*. in *Eurospeech*. Madrid, Spain. p. 299-302.

[7] 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*. in *Eurospeech*. Rhodos, Greece. p. 2003-2010.

[8] Brand, M. (1999) *Voice pupperty*. in *SIGGRAPH'99*. Los Angeles, CA. p. 21-28.

[9] Bregler, C., Cowell, M., and Slaney, M. (1997) *VideoRewrite: driving visual speech with audio.* in *SIGGRAPH'97*. Los Angeles, CA. p. 353-360.

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

[11] Chabanas, M. and Payan, Y. (2000) *A 3D Finite Element model of the face for simulation in plastic and maxillo-facial surgery.* in *International Conference on Medical Image Computing and Computer-Assisted Interventions.* Pittsburgh, USA. p. 1068-1075.

[12] Cohen, M.M., Beskow, J., and Massaro, D.W. (1998) *Recent developments in facial animation: an inside view.*

in *ETRW on Auditory-Visual Speech Processing*. Terrigal-Sydney, Australia. p. 201-206.

[13] Cohen, M.M. and Massaro, D.W. (1993) *Modeling coarticulation in synthetic visual speech*, in *Models and Techniques in Computer Animation*, D. Thalmann and N. Magnenat-Thalmann, Editors. Springer-Verlag: Tokyo. p. 141-155.

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

[15] Cosatto, E. and Graf, H.P. (1998) *Sample-based of photo-realistic talking heads*. in *Computer Animation*. Philadelphia, Pennsylvania. p. 103-110.

[16] 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): p. 433-463.

[17] 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): p. 70-78.

[18] 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.

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

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

[21] Ezzat, T. and Poggio, T. (1998) *MikeTalk: a talking facial display based on morphing visemes*. in *Computer Animation*. Philadelphia, PA. p. 96-102.

[22] Massaro, D. (1998) *Illusions and issues in bimodal speech perception*. in *Auditory-Visual Speech Processing Conference*. Terrigal, Sydney, Australia. p. 21-26.

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

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

[25] 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): p. 177-210.

[26] Öhman, S.E.G. (1967) *Numerical model of coarticulation.* Journal of the Acoustical Society of America, **41**: p. 310-320.

[27] Okadome, T., Kaburagi, T., and Honda, M. (1999) *Articulatory movement formation by kinematic triphone model*. in *IEEE International Conference on Systems Man and Cybernetics,*. Tokyo, Japan. p. 469-474.

[28] Olives, J.-L., Möttönen, R., Kulju, J., and Sams, M. (1999) *Audio-Visual Speech Synthesis for Finnish*. in *Auditory-visual Speech Processing Workshop*. Santa Cruz, CA. p. 157-162.

[29] Pandzig, I., Ostermann, J., and Millen, D. (1999) *Users evaluation: synthetic talking faces for interactive services.* The Visual Computer, **15**: p. 330-340.

[30] Parke, F.I. (1972) *Computer generated animation of faces*. University of Salt Lake City: Salt Lake City.

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

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

[33] 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**: p. 365-377.

[34] Pisoni, D.B. (1997) *Perception of synthetic speech*, in *Progress in Speech Synthesis*, J.P.H.V. Santen, et al., Editors. Springer Verlag: New York. p. 541-560.

[35] Platt, S.M. and Badler, N.I. (1981) *Animating facial expressions.* Computer Graphics, **15**(3): p. 245-252.

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

[37] 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.* in *International Conference on Speech and Language Processing*. Beijing - China. p. 755-758.

[38] 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**: p. 3000-3007.

[39] Tamura, M., Kondo, S., Masuko, T., and Kobayashi, T. (1999) *Text-to-Audio-Visual Speech Synthesis based on Parameter Generation from HMM*. in *European Conference on Speech Communication and Technology*. Budapest, Hungary. p. 959-962.

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

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

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

[43] Vignoli, F. and Braccini, C. (1999) *A text-speech synchronization technique with applications to talking heads*. in *Auditory-Visual Speech Processing Conference*. Santa Cruz, California, USA

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

[45] Yamamoto, E., Nakamura, S., and Shikano, K. (1998) *Lip Movement Synthesis from Speech based on Hidden Markov Models.* Speech Communication, **26**(1-2): p. 105-115.

[46] Yehia, H., Kuratate, T., and Vatikiotis-Bateson, E. (2000) *Facial animation and head motion driven by speech acoustics*. in *5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.* Kloster Seeon, Germany. p. 265-268.