HAL
open science

Lip-synching using speaker-specific articulation, shape and appearance models

Gérard Bailly, Oxana Govokhina, Frédéric Elisei, Gaspard Breton

To cite this version:

Gérard Bailly, Oxana Govokhina, Frédéric Elisei, Gaspard Breton. Lip-synching using speaker-specific articulation, shape and appearance models. EURASIP Journal on Audio, Speech, and Music Processing, 2009, Special issue on animating virtual speakers or singers from audio: Lip-synching facial animation, pp.ID 769494. 10.1155/2009/769494 . hal-00447061

HAL Id: hal-00447061

https://hal.science/hal-00447061v1

Submitted on 14 Jan 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2009, Article ID 769494, 11 pages
doi:10.1155/2009/769494

Research Article

Lip-Synching Using Speaker-Specific Articulation, Shape and Appearance Models

Gérard Bailly,[1] Oxana Govokhina,[1,2] Frédéric Elisei,[1] and Gaspard Breton[2]

[1] *Department of Speech and Cognition, GIPSA-Lab, CNRS & Grenoble University, 961 rue de la Houille Blanche-Domaine universitaire-BP 46-38402 Saint Martin d'Hères cedex, France*
[2] *TECH/IRIS/IAM Team, Orange Labs, 4 rue du Clos Courtel, BP 59 35512 Cesson-Sévigné, France*

Correspondence should be addressed to Gérard Bailly, gerard.bailly@gipsa-lab.grenoble-inp.fr

Received 25 February 2009; Revised 26 June 2009; Accepted 23 September 2009

Recommended by Sascha Fagel

We describe here the control, shape and appearance models that are built using an original photogrammetric method to capture characteristics of speaker-specific facial articulation, anatomy, and texture. Two original contributions are put forward here: the trainable trajectory formation model that predicts articulatory trajectories of a talking face from phonetic input and the texture model that computes a texture for each 3D facial shape according to articulation. Using motion capture data from different speakers and module-specific evaluation procedures, we show here that this cloning system restores detailed idiosyncrasies and the global coherence of visible articulation. Results of a subjective evaluation of the global system with competing trajectory formation models are further presented and commented.

Copyright © 2009 Gérard Bailly et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Embodied conversational agents (ECAs)—virtual characters as well as anthropoid robots—should be able to talk with their human interlocutors. They should generate facial movements from symbolic input. Given history of the conversation and thanks to a model of the target language, dialog managers and linguistic front-ends of text-to-speech systems compute a phonetic string with phoneme durations. This minimal information can be enriched with details of the underlying phonological and informational structure of the message, with facial expressions, or with paralinguistic information (mental or emotional state) that all have an impact on speech articulation. A trajectory formation model—called also indifferently articulation or control model—has thus to be built that computes control parameters from such a symbolic specification of the speech task. These control parameters will then drive the talking head (the shape and appearance models of a talking face or the proximal degrees-of-freedom of the robot).

The acceptability and believability of these ECA depend on at least three factors: (a) the information-dependent factors that relate to the relevance of the linguistic content and paralinguistic settings of the messages, (b) the appropriate choice of voice quality, communicative and emotional facial expressions, gaze patterns, and so forth, adapted to situation and environmental conditions; (c) the signal-dependent factors that relate to the quality of the rendering of this information by multimodal signals. This latter signal-dependent contribution depends again on two main factors: the intrinsic quality of each communicative channel, that is, intrinsic quality of synthesized speech, gaze, facial expressions, head movements, hand gestures and the quality of the interchannel coherence, that is, the proper coordination between audible and visible behavior of the recruited organs that enable intuitive perceptual fusion of these multimodal streams in an unique and coherent communication flow. This paper addresses these two issues by (i) first describing a methodology for building virtual copies of speaker-specific facial articulation and appearance, and (ii) a model that captures most parts of the audiovisual coherence and asynchrony between speech and observed facial movements.

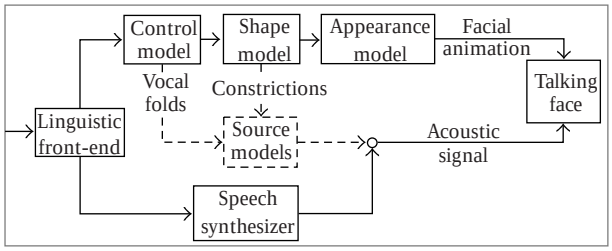

Figure 1: A facial animation system generally comprises three modules: the control model that computes a gestural score given the phonetic content of the message to be uttered, a shape model that computes the facial geometry, and an appearance model that computes the final appearance of the face on screen. The acoustic signal can be either postsynchronized or computed by articulatory synthesis. In this later case the internal speech organs shape the vocal tract (tongue, velum, etc.) that is further acoustically "rendered" by appropriate sound sources.

This "cloning" suite—that captures speaker-specific idiosyncrasies related to speech articulation—is then evaluated. We will notably show that the proposed statistical control model for audiovisual synchronization favorably competes with the solution that consists in concatenating multimodal speech segments.

2. State of the Art

Several review papers have been dedicated to speech and facial animation [1, 2]. A facial animation system generally comprises three modules (cf. Figure 1).

(1) A control model that computes gestural trajectories from the phonetic content of the message to be uttered. The main scientific challenge of this processing stage is the modeling of the so-called coarticulation, that is, context-dependent articulation of sounds. The articulatory variability results in fact not only from changes of speech style or emotional content but also from the under specification of articulatory targets and planning [3].

(2) A shape model that computes the facial geometry from the previous gestural score. This geometry is either 2D for image-based synthesis [4, 5] or 3D for biomechanical models [6, 7]. The shape model drives movements of fleshpoints on the face. These fleshpoints are usually vertices of a mesh that deforms according to articulation. There are three main scientific challenges here: (a) identifying a minimal set of independent facial movements related to speech as well as facial expressions [8] (b) identifying the movement of fleshpoints that are poorly contrasted on the face: this is usually done by interpolating movements of robust fleshpoints (lips, nose, etc.) surrounding each area or regularizing the optical flow [9]; (c) linking control variables to movements, that is, capturing and modeling realistic covariations of geometric changes all over the lower face by independent articulations, for example, jaw rotation, lip opening, and lip rounding all change shape of lips and nose wings.

(3) An appearance model that computes the final appearance of the face on screen. This is usually done by warping textures on the geometric mesh. Most textures are generally a function of the articulation and other factors such as position of light sources and skin pigmentation. The main challenge here is to capture and model realistic covariations of appearance and shape, notably when parts of the shape can be occluded. The challenge is in fact even harder for inner organs (teeth, tongue, etc.) that are partially visible according to lip opening.

Most multimodal systems also synthesize the audio signal although most animations are still postsynchronized with a recorded or a synthetic acoustic signal. The problem of audiovisual coherence is quite important: human interlocutors are very sensitive to discrepancies between the visible and audible consequences of articulation [10, 11] and have expectations on resulting audiovisual traces of the same underlying articulation. The effective modeling of audiovisual speech is therefore a challenging issue for trajectory formation systems and still an unsolved problem. Note however that intrinsically coherent visual and audio signals can be computed by articulatory synthesis where control and shape models drive the internal speech organs of the vocal tract (tongue, velum, etc.). This vocal tract shape is then made audible by the placement and computation of appropriate sound sources.

3. Cloning Speakers

We describe here the cloning suite that we developed for building speaker-specific 3D talking heads that best captures the idiosyncratic variations of articulation, geometry, and texture.

3.1. Experimental Data. The experimental data for facial movements consists in photogrammetric data collected by three synchronized cameras filming the subject's face. Studio digital disk recorders deliver interlaced uncompressed PAL video images at 25 Hz. When deinterlaced, the system delivers three 288×720 uncompressed images at 50 Hz in full synchrony with the audio signal.

We characterize facial movements both by the deformation of the facial geometry (the shape model described below) and by the change of skin texture (the appearance model detailed in Section 5). The deformation of the facial geometry is given by the displacement of facial fleshpoints. Instead of relying on sophisticated image processing techniques—such as optical flow—to estimate these displacements with no make-up, we choose to build very detailed shape models by gluing hundreds of beads on the subjects' face (see Figure 2). 3D movements of facial fleshpoints are acquired using multicamera photogrammetry.

This 3D data is supplemented by lip geometry that is acquired by fitting semiautomatically a generic lip model

(a) Speaker CD

(b) Speaker OC

Figure 2: Two speakers utter here sounds with different make-ups. Colored beads have been glued on the subjects' face along Langer's lines so as to cue geometric deformations caused by main articulatory movements when speaking. Left: a make-up with several hundreds of beads is used for building the shape model. Right: a subset of crucial fleshpoints is preserved for building videorealistic textures.

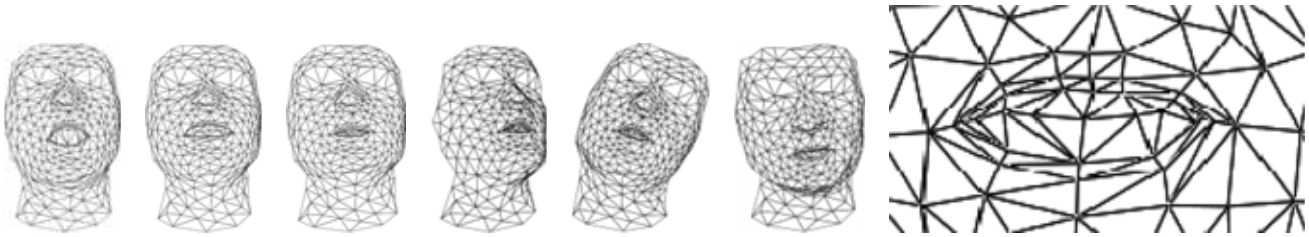

Figure 3: Some elementary articulations for the face and the head that statistically emerge from the motion capture data of speaker CD using guided PCA. Note that a nonlinear model of the head/neck joint is also parameterized. The zoom at the right-hand side shows that the shape model includes a detailed geometry of the lip region: a lip mesh that is positioned semiautomatically using a generic lip model [12] as well as a mesh that fills the inner space. This later mesh attaches the inner lip contour to the ridge of the upper teeth: there is no further attachment to other internal organs (lower teeth, tongue, etc.).

[12] to the speaker-specific anatomy and articulation. This is in fact impossible to glue beads on the wet part of the lips and this would also impact on articulation.

Data used in this paper have been collected for three subjects: an Australian male speaker (see Figure 2(a)), a UK-English female speaker (see Figure 2(b)), and a French female speaker (see Figure 12). They will be named, respectively, by the initials CD, OC, and AA.

3.2. The Shape Model. In order to be able to compare up-to-date data-driven methods for audiovisual synthesis, a main corpus of hundreds of sentences pronounced by the speaker is recorded. The phonetic content of these sentences is optimized by a greedy algorithm that maximizes statistical coverage of triphones in the target language (differentiated also with respect to syllabic and word boundaries).

The motion capture technique developed at GIPSA-Lab [13, 14] consists in collecting precise 3D data on selected visemes. Visemes are selected in the natural speech flow by an analysis-by-synthesis technique [15] that combines automatic tracking of the beads with semiautomatic correction.

Our shape models are built using a so-called guided Principal Component Analysis (PCA) where a priori knowledge is introduced during the linear decomposition. We in fact compute and iteratively subtract predictors using carefully chosen data subsets [16]. For speech movements, this methodology enables us to extract at least six components once the head movements have been removed.

The first one, jaw1 controls the opening/closing movement of the jaw and its large influence on lips and face shape. Three other parameters are essential for the lips: lips1 controls the protrusion/spreading movement common to both lips as involved in the /i/ versus /y/ contrast; lips2 controls the upper lip raising/lowering movement used for example in the labio-dental consonant /f/; lips3 controls the lower lip lowering/raising movement found in consonant / / for which both lips are maximally open while jaw is in a high position. The second jaw parameter, jaw2, is associated with a horizontal forward/backward movement of the jaw that is used in labio-dental articulations such as /f/ for example. Note finally a parameter lar1 related to the vertical movements of the larynx that are particularly salient for males. For the three subjects used here, these components account for more than 95% of the variance of the positions of the several hundreds of fleshpoints for thirty visemes carefully chosen to span the entire articulatory space of each

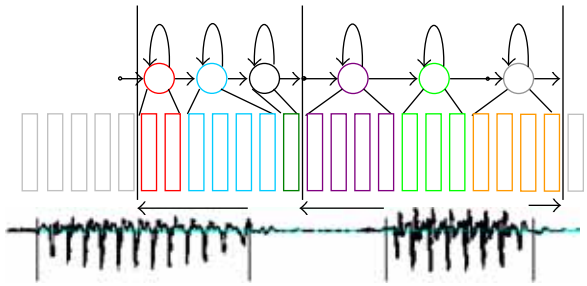

Figure 4: The phasing model of the PHMM predicts phasing relations between acoustic onsets of the phones (bottom) and onsets of context-dependent phone HMM that generate the frames of the gestural score (top). In this example, onsets of gestures characterizing the two last sounds are in advance compared to effective acoustics onsets. For instance an average delay between observed gestural and acoustic onset is computed and stored for each context-dependent phone HMM. This delay is optimized with an iterative procedure described in Section 4.3 and illustrated in Figure 5.

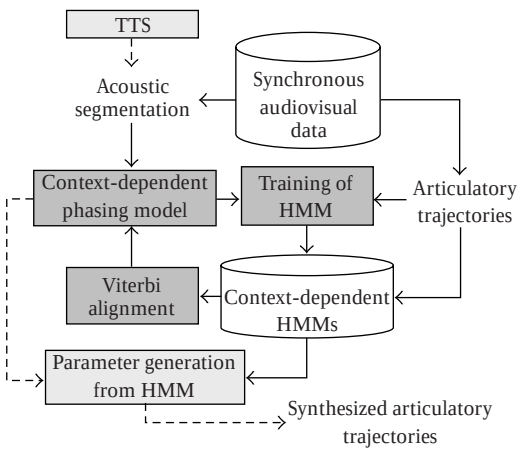

Figure 5: Training consists in iteratively refining the context-dependent phasing model and HMMs (plain lines and dark blocks). The phasing model computes the average delay between acoustic boundaries and HMM boundaries obtained by aligning current context-dependent HMMs with training utterances. Synthesis simply consists in forced alignment of selected HMMs with boundaries predicted by the phasing model (dotted lines and light blocks).

language. The root mean square error is in all cases less than 0.5 mm for both hand-corrected training visemes and test data where beads are tracked automatically on original images [15].

The final articulatory model is supplemented with components for head movements (and neck deformation) and with basic facial expressions [17] but only components related to speech articulation are considered here. The average modeling error is less than 0.5 mm for beads located on the lower part of the face.

4. The Trajectory Formation System

The principle of speech synthesis by HMM was first introduced by Tokuda et al. [18] for acoustic speech synthesis and extended to audiovisual speech by the HTS working group [19]. Note that the idea of exploiting HMM capabilities for grasping essential sound characteristics for synthesis was also promoted by various authors such as Giustiniani and Pierucci [20] and Donovan [21]. The HMM-trajectory synthesis technique comprises training and synthesis parts (see [22, 23] for details).

4.1. Basic Principles. An HMM and a duration model for each state are first learned for each segment of the training set. The input data for the HMM training is a set of observation vectors. The observation vectors consist of static and dynamic parameters, that is, the values of articulatory parameters and their temporal derivatives. The HMM parameter estimation is based on Maximum-Likelihood (ML) criterion [22]. Usually, for each phoneme in context, a 3-state left-to-right model is estimated with single Gaussian diagonal output distributions. The state durations of each HMM are usually modeled as single Gaussian distributions. A second training step can also be added to factor out similar output distributions among the entire set of states, that is, state tying. This step is not used here.

The synthesis is then performed as follows. A sequence of HMM states is built by concatenating the context-dependent phone-sized HMM corresponding to the input phonetic string. State durations for the HMM sequence are determined so that the output probabilities of the state durations are maximized (thus usually by z-scoring). Once the state durations have been assigned, a sequence of observation parameters is generated using a specific ML-based parameter generation algorithm [22] taking into account the distributions of both static and dynamic parameters that are implicitly linked by simple linear relations (e.g., $\Delta p(t) = p(t) - p(t-1)$; $\Delta\Delta p(t) = \Delta p(t) - \Delta p(t-1) = p(t) - p(t-2)$; etc.).

4.2. Comments. States can capture parts of the interarticulatory asynchrony since transient and stable parts of the trajectories of different parameters are not obligatory modeled by the same state. As an example, a state of an HMM model can observe a stable part of one parameter A (characterized by a mean dynamic parameter close to zero) together with a synchronous transient for another parameter B (characterized by a positive or negative mean dynamic parameter). If the next state observes the contrary for parameters A and B, the resulting trajectory synthesis will exhibit an asynchronous transition between A and B. This surely explains why complex HMM structures aiming at explicitly coping with audiovisual asynchronies do not outperform the basic ergodic structure, especially for audiovisual speech recognition [24]. Within a state, articulatory dynamics is captured and is then reflected in the synthesized trajectory. By this way, this algorithm may capture implicitly part of short-term coarticulation patterns and inter-articulatory

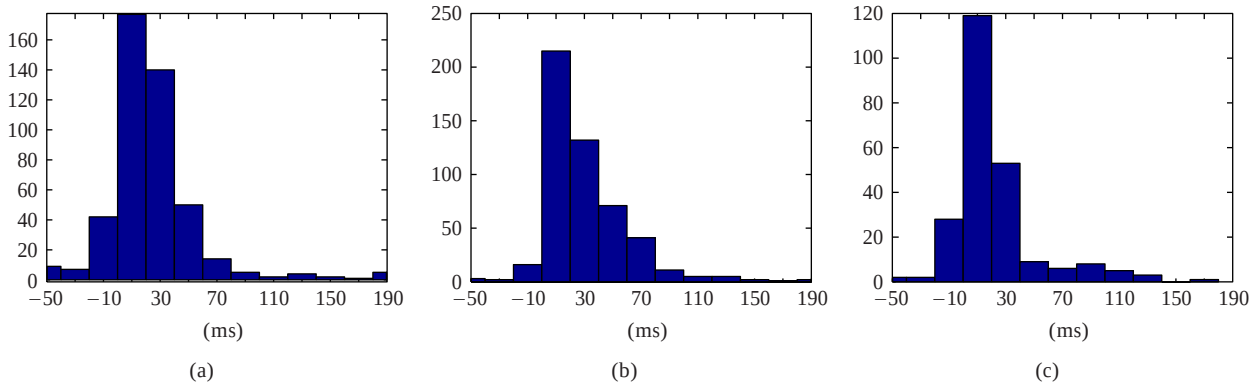

Figure 6: Distribution of average time lags estimated for the HMM bi-phones collected from our speakers. From left to right: CD, OC, and AA. Note that time lags are mainly positive, that is, gestural boundaries—pacing facial motion—are mainly located after acoustic boundaries.

asynchrony. Larger coarticulation effects can also be captured since triphones intrinsically depend on adjacent phonetic context.

These coarticulation effects are however anchored to acoustic boundaries that are imposed as synchronization events between the duration model and the HMM sequence. Intuitively we can suppose that context-dependent HMM can easily cope with this constraint but we will show that adding a context-dependent phasing model helps the trajectory formation system to better fit observed trajectories.

4.3. Adding and Learning a Phasing Model. We propose to add a phasing model to the standard HMM-based trajectory formation system that learns the time lag between acoustic and gestural units [25, 26], that is, between acoustic boundaries delimiting allophones and gestural boundaries delimiting pieces of the articulatory score observed by the context-dependent HMM sequence (see Figure 4). This trajectory formation system is called PHMM (for Phased-HMM) in the following.

A similar idea was introduced by Saino et al. [27] for computing time-lags between notes of the musical score and sung phones for an HMM-based singing voice synthesis system. Both boundaries are defined by clear acoustic landmarks and can be obtained semiautomatically by forced alignment. Lags between boundaries are clustered by a decision tree in the same manner used for clustering spectral, fundamental frequency, and duration parameters in HMM synthesis. Saino et al. [27] evaluated their system with 60 Japanese children's songs by one male speaker resulting in 72 minutes of signal in total and showed a clear perceptual benefit of the lag model in comparison with an HMM-based system with no lag models.

In our case gestural boundaries are not available: gestures are continuous and often asynchronous [28]. It is very difficult to identify core gestures strictly associated with each allophone. Gestural boundaries emerge here as a by-product of the iterative learning of lags. We use here the term phasing model instead of lag model in reference to work on control: events are in phase when the lag equals 0 and antiphase when the average lag is half the average duration between events. Because of the limited amount of AV data (typically several hundreds of sentences, typically 15 minutes of speech in total), we use here a very simple phasing model: a unique time lag is associated with each context-dependent HMM. This lag is computed as the mean delay between acoustic boundaries and results of forced HMM alignment with original articulatory trajectories.

These average lags are learnt by an iterative process consisting of an analysis-synthesis loop (see Figure 5).

(1) Standard context-dependent HMMs are learnt using acoustic boundaries as delimiters for gestural parameters.

(2) Once trained, forced alignment of training trajectories is performed (Viterbi alignment in Figure 5).

(3) Deviations of the resulting segmentation with acoustic boundaries are collected. The average deviation of the right boundary of each context-dependent HMM is then computed and stored. The set of such mean deviations constitutes the phasing model.

(4) New gestural boundaries are computed applying the current phasing model to the initial acoustic boundaries. Additional constraints are added to avoid collapsing: a minimal duration of 30 milliseconds is guaranteed for each phone.

A typical distribution of these lags is given in Figure 6. For context-dependent phone HMM where contextual information is limited to the following phoneme, lags are mostly positive: gestural boundaries occur latter than associated acoustic ones, that is, there is more carryover coarticulation than anticipatory one.

4.4. Objective Evaluation. All sentences are used for training. A leave-one-out process for PHMM has not been used since a context-dependent HMM is built only if at least 10 samples are available in the training data; otherwise context-independent phone HMMs are used. PHMM is compared with concatenative synthesis using multirepresented diphones [29]: synthesis of each utterance is performed simply by using all diphones of other utterances. Selection

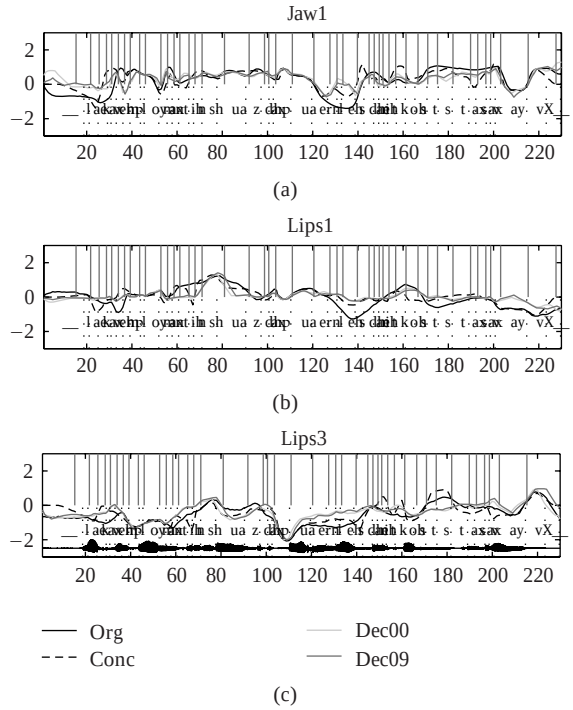

Figure 7: Comparing natural (dark blue) and synthetic trajectories computed by three different systems for the first 6 main articulatory parameters (jaw opening, lip spreading, jaw protrusion, lower and upper lip opening, laryngeal movements) for the sentence "The lack of employment ensures that the poor earn less than it costs to survive." The three systems are concatenation of audiovisual diphones (black), HMM-based synthesis (light blue), and the proposed PHMM (red). Vertical dashed lines at the bottom of each caption are acoustic boundaries while gestural boundaries are given by the top plain lines. Note the large delay of the non audible prephonatory movements at the beginning of the utterance. The trajectories of lower and upper lips for the word "ensures" is zoomed and commented in Figure 8.

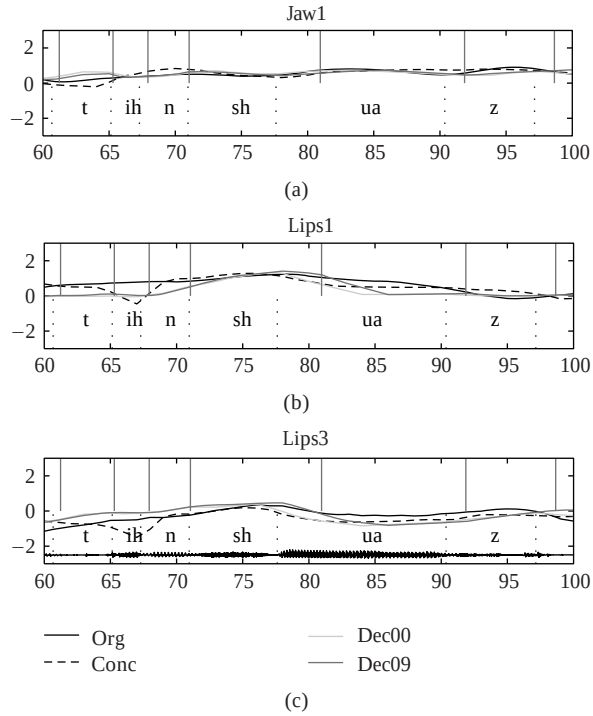

Figure 8: A zoomed portion of Figure 7 evidencing that PHMM (red) captures the original carryover movements (dark blue) of the open consonant [sh] into the [ua] vowel. We plot here the behavior of the lower and upper lip opening. PHMM predicts a protrusion of the lips into half of the duration of the [ua] allophone while both HMM-based (light blue) and concatenation-based (black) trajectory formation systems predict a quite earlier retraction at acoustic onset. In the original stimuli the protrusion is sustained till the end of the word "ensures."

is performed classically using minimization of selection and concatenation costs over the sentence.

Convergence is obtained after typically 2 or 3 iterations. Figures 7 and 8 compare the articulatory trajectories obtained: the most important gain is obtained for silent articulations typically at the beginning (prephonatory gestures) and end of utterances.

Figure 9 compares mean correlations obtained by the concatenative synthesis with those obtained by the PHMM at each iteration. The final improvement is small, typically 4–5% depending on the speaker. We especially used the data of our French female speaker for subjective evaluation because PHMM does not improve objective HMM results; we will show that the subjective quality is significantly different.

We have shown elsewhere [25] that the benefit of phasing on prediction accuracy is very conservative; PHMM always outperforms the HMM-based synthesis anchored strictly on acoustic boundaries whatever contextual information is added or the number of Gaussian mixtures is increased.

5. The Photorealistic Appearance Model

Given the movements of the feature points, the appearance model is responsible for computing the color of each pixel of the face. Three basic models have been proposed so far in the literature.

(1) Patching facial regions [4, 30]: prestored patches are selected from a patch dictionary according to the articulatory parameters and glued on the facial surface according to face and head movements.

(2) Interpolating between target images [9, 31]: the shape model is often used to regularize the computation of the optical flow between pixels of key images.

(3) Texture models [32, 33]: view-dependent or view independent—or cylindrical textures—texture maps are extracted and blended according to articulatory parameters and warped on the shape.

Our texture model computes texture maps. These maps are computed in three steps.

The detailed shape model built using several hundreds of fleshpoints is used to track articulation of faces marked only by a reduced number of beads (see Figure 2). We do not use all available data (typically several dozen thousand frames):

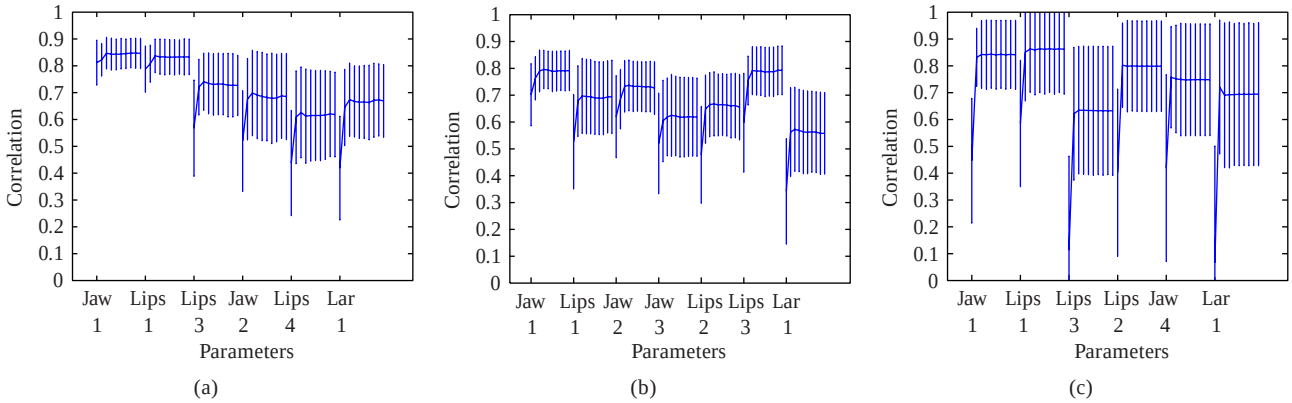

FIGURE 9: Mean correlations (together with standard deviations) between original and predicted trajectories for the main six articulatory parameters (jaw rotation, lip rounding, lower and upper lip opening, jaw retraction and larynx height). For each parameter, correlations for eleven conditions are displayed: the first correlation is for the trajectories predicted by concatenative synthesis using multirepresented diphones (see text); the second correlation is for trajectories predicted by HMM using acoustic boundaries; the rest of the data give results obtained after the successive iteration of the estimations of the phasing model. Asymptotic behavior is obtained within one or two iterations. From left to right: data from speakers CD, OC, and AA.

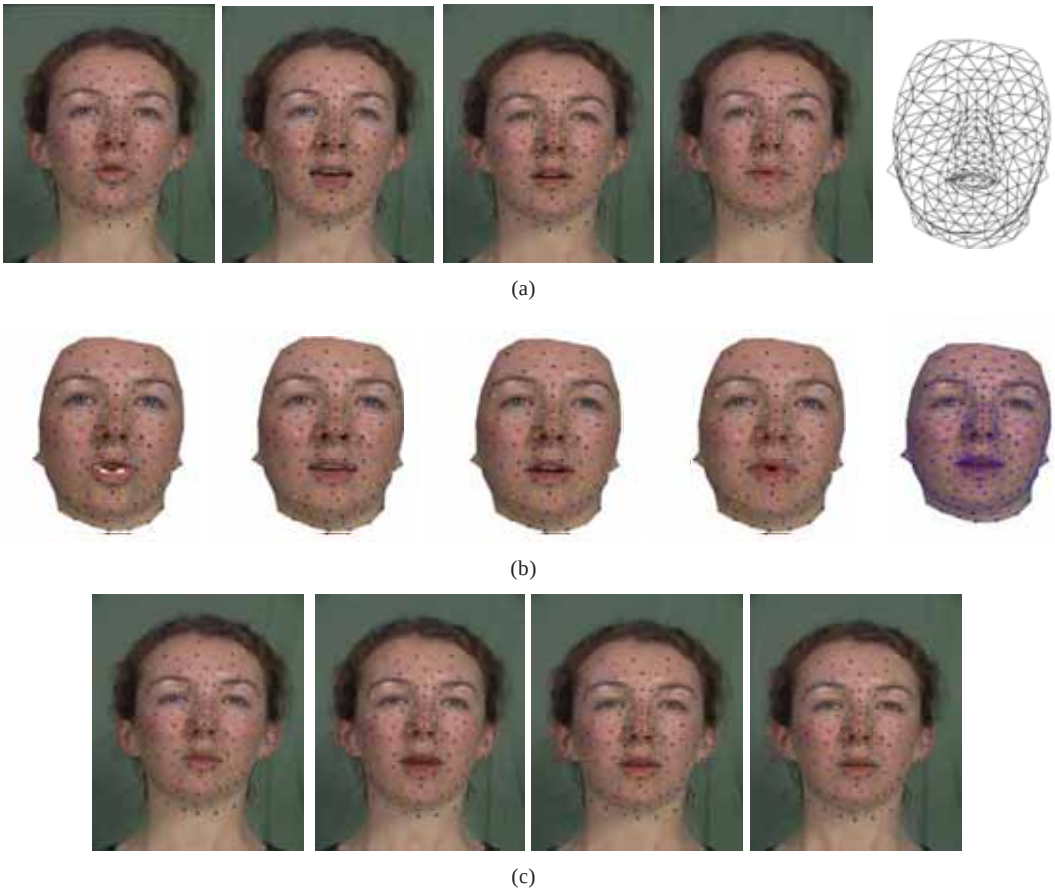

FIGURE 10: Texturing the facial mesh with an appearance model for OC. (a) Original images that will be warped to the "neutral" mesh displayed on the right. (b) shape-free images obtained: triangles in white color are not considered in the modeling process because they are not fully visible from the front camera. The left image displays the mean texture together with the "neutral" mesh drawn with blue lines. (c) resynthesis of the facial animation using the shape and appearance models superposed to the original background video.

Figure 11: Comparison between original images and resynthesis of various articulations for CD. Note the lightening bar at the bottom of the neck due to the uncontrolled sliding of the collar of the tee-shirt during recordings.

Figure 12: Same as Figure 11 for AA whose data have been used for the comparative subjective evaluation described in Section 6.

We only retain one target image per allophone (typically a few thousand frames).

Shape-free images (see [32]) are extracted by warping the selected images to a "neutral shape" (see middle of Figure 10).

A linear regression of the RGB values of all visible pixels of our shape-free images by the values of articulatory parameters obtained in step 1. The speaker-specific shape and appearance models are thus driven by the same articulatory parameters. Instead of the three PCA performed for building Active Appearance Models [32] where independent shape and appearance models are first trained and then linked, we are only concerned here by changes of shape and appearance directly linked with our articulatory parameters. For instance the articulatory-to-appearance mapping is linear but non-linear mapping is possible because of the large amount of training data available by step 1.

The layered mesh-based mapping is of particular importance for the eyes and lips where different textured plans (e.g., iris, teeth, tongue) appear and disappear according to aperture.

Note also that the 3D shape model is used to weight the contribution of each pixel to the regression, for instance, all pixels belonging to a triangle of the facial mesh that is not visible or does not face the camera are discarded (see Figure 10). This weighting can also be necessary for building view-independent texture models: smooth blending between multiview images may be obtained by weighting contribution of each triangle according to its viewing angle and the size of its deformation in the shape-free image.

6. Subjective Evaluation

A first evaluation of the system was performed at the LIPS'08 lipsync challenge [34]. With minor corrections, it winned the intelligibility test at the next LIPS'09 challenge. The trainable trajectory formation model PHMM, the shape and appearance models were parameterized using OC data. The texture model was trained using the front-view images from the corpus with thousands of beads (see left part of Figure 2(b)). The system was rated closest to the original video considering both audiovisual consistency and intelligibility. It was ranked second for audiovisual consistency and very close to the winner. Concerning intelligibility, several systems outperformed the original video. Our system offers the same visual benefit as the natural video is not less not more.

We also performed a separate evaluation procedure to evaluate the contribution of PHMM to the appreciation of the overall quality. We thus tested different control models maintaining the shape and appearance models strictly the same for all animations. This procedure is similar to the modular evaluation previously proposed [29] but with video-realistic rendering of movements instead of a point-light display. Note that concatenative synthesis was the best control model and outperformed the most popular coarticulation models in this 2002 experiment.

6.1. Stimuli. The data used in this experiment are from a French female speaker (see Figure 12) cloned using the same principles as above.

We compare here audio-visual animations built by combining the original sound with synthetic animations driven by various gestural scores: the original one (Nat) and 4 other scores computed from the phonetic segmentation of the sound. All videos are synthetic. All articulatory trajectories are "rendered" by the same shape and appearance models in order to focus on perceptual differences only due to the quality of control parameters. The four control models are the following.

(1) The trajectory formation model proposed here (PHMM).

(2) The basic audio-synchronous HMM trajectory formation system (HMM).

(3) A system using concatenative synthesis with multi-represented diphones (CONC). This system is similar

to the Multisyn synthesizer developed from acoustic synthesis [35] but uses here an audiovisual database.

(4) A more complex control model called TDA [36] that uses PHMM twice. PHMM is first used to segment training articulatory trajectories into gestural units. They are stored into a gestural dictionary. The previous system CONC is then used to select and concatenate the appropriate multi-represented gestural units. CONC and TDA however differ in the way selection costs are computed. Whereas CONC only considers phonetic labels, TDA uses the PHMM prediction to compute a selection cost for each selected unit by computing its distance to the PHMM prediction for that portion of the gestural score.

The five gestural scores drive then the same plant, that is, the shape textured by the videorealistic appearance model. The resulting facial animation is then patched back with the appropriate head motion on the original background video as in [4, 9].

6.2. Test Procedure and Results. 20 naïve subjects (33 ± 10 years, 60% male) participated in the audio-visual experiment. The animations were played on a computer screen. They were informed that these animations were all synthetic and that the aim of the experiment was to rate different animation techniques.

They were asked to rate on a 5-point MOS scale (very good, good, average, insufficient, very insufficient) the coherence between the sound and the computed animation.

Results are displayed in Figure 13. All ratings are within the upper MOS scale, that is, between average and very good. Three groups can be distinguished: (a) the trajectory formation systems PHMM and TDA are not distinguished from the resynthesis of original movements; (b) the audio-synchronous HMM trajectory formation system is then rated best, and (c) the concatenation system with multi-represented audiovisual diphones is rated significantly worse than all others.

6.3. Comments. The HMM-based trajectory formation systems are significantly better than the data-driven concatenative synthesis that outperforms coarticulation models even when parameterized by the same data. The way we exploit training data has thus made important progress in the last decennia; it seems that structure should emerge from data and not be parameterized by data. Data modeling takes over data collection not only because modeling regularizes noisy data but also because modeling takes into account global parameters such as the minimization of global distortion or variance.

7. Conclusions

We have demonstrated here that the prediction accuracy of an HMM-based trajectory formation system is improved by modeling the phasing relations between acoustic and gestural boundaries. The phasing model is learnt using an analysis-synthesis loop that iterates HMM estimations and forced alignments with the original data. We have shown that this scheme improves significantly the prediction error and captures both strong (prephonatory gestures) and subtle (rounding) context-dependent anticipatory phenomena.

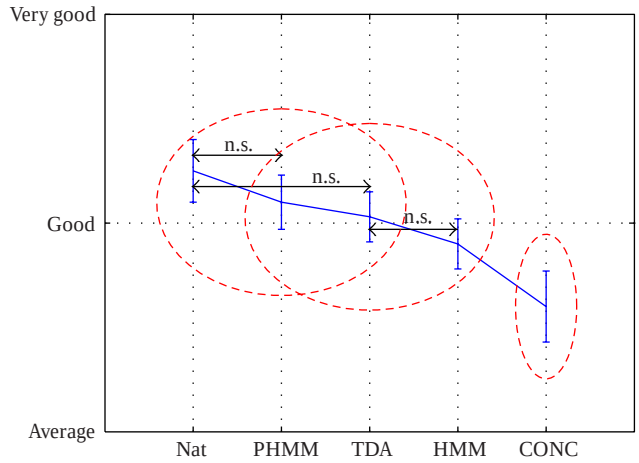

Figure 13: Results of the MOS test. Three groups can be distinguished: (a) the trajectory formation systems PHMM and TDA are not distinguished from the resynthesis of original movements; (b) the audio-synchronous HMM trajectory formation system is then rated best, and (c) the concatenation system with multi-represented audiovisual diphones is rated significantly worse than all others.

The interest of such an HMM-based trajectory formation system is double: (i) it provides accurate and smooth articulatory trajectories that can be used straightforwardly to control the articulation of a talking face or used as a skeleton to anchor multimodal concatenative synthesis (see notably the TDA proposal in [36]); (ii) it also provides gestural segmentation as a by-product of the phasing model. These gestural boundaries can be used to segment original data for multimodal concatenative synthesis. A more complex phasing model can of course be built—using, for example, CART trees—by identifying phonetic or phonological factors influencing the observed lag between visible and audible traces of articulatory gestures.

Concerning the plant itself, much effort is still required to get a faithful view-independent appearance model, particularly for the eyes and inner mouth. For the later, precise prediction of jaw position—and thus lower teeth—and tongue position should be performed in order to capture changes of appearance due to speech articulation. Several options should be tested: direct measurements via jaw splint or EMA [37], additional estimators linking tongue and facial movements [38], or more complex statistical models optimally linking residual appearance of the inner mouth to phonetic content.

Acknowledgments

The GIPSA-Lab/MPACIF team thanks Orange R&D for their financial support as well as the Rhône-Alpes region and the PPF "Multimodal Interaction." Part of this work was also

developed within the PHC Procope with Sascha Fagel at TU Berlin. The authors thank their target speakers for patience and motivation. They thank Erin Cvejic for his work on the CD data.

References

[1] G. Bailly, M. Bérar, F. Elisei, and M. Odisio, "Audiovisual speech synthesis," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 331–346, 2003.

[2] B. J. Theobald, "Audiovisual speech synthesis," in *Proceedings of the 16th International Congress of Phonetic Sciences (CPhS '07)*, pp. 285–290, Saarbrücken, Germany, August 2007.

[3] D. H. Whalen, "Coarticulation is largely planned," *Journal of Phonetics*, vol. 18, no. 1, pp. 3–35, 1990.

[4] C. Bregler, M. Covell, and M. Slaney, "Videorewrite: driving visual speech with audio," in *Proceedings of the 24th Annual Conference on Computer Graphics (SIGGRAPH '97)*, pp. 353–360, Los Angeles, Calif, USA, August 1997.

[5] B. J. Theobald, J. A. Bangham, I. Matthews, and G. C. Cawley, "Visual speech synthesis using statistical models of shape and appearance," in *Proceedings of the Auditory-Visual Speech Processing Workshop (AVSP '01)*, pp. 78–83, Scheelsminde, Denmark, September 2001.

[6] M. Chabanas and Y. Payan, "A 3D Finite Element model of the face for simulation in plastic and maxillo-facial surgery," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Interventions (MICCAI '00)*, pp. 1068–1075, Pittsburgh, Pa, USA, 2000.

[7] D. Terzopoulos and K. Waters, "Analysis and synthesis of facial image sequences using physical and anatomical models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 569–579, 1993.

[8] P. Ekman, "What we have learned by measuring facial behavior," in *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, P. Ekman and E. Rosenberg, Eds., pp. 469–485, Oxford University Press, New York, NY, USA, 1997.

[9] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 388–398, 2002.

[10] N. F. Dixon and L. Spitz, "The detection of audiovisual desynchrony," *Perception*, vol. 9, pp. 719–721, 1980.

[11] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[12] T. Guiard-Marigny, A. Adjoudani, and C. Benoît, "3D models of the lips and jaw for visual speech synthesis," in *Progress in Speech Synthesis*, J. P. H. van Santen, R. Sproat, J. Olive, and J. Hirschberg, Eds., pp. 247–258, Springer, Berlin, Germany, 1996.

[13] F. Elisei, M. Odisio, G. Bailly, and P. Badin, "Creating and controlling video-realistic talking heads," in *Proceedings of the Auditory-Visual Speech Processing Workshop (AVSP '01)*, pp. 90–97, Scheelsmind, Denmark, 2001.

[14] L. Revéret, G. Bailly, and P. Badin, "MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation," in *Proceedings of the International Conference on Speech and Language Processing (ICSLP '00)*, pp. 755–758, Beijing, China, July-August 2000.

[15] G. Bailly, F. Elisei, P. Badin, and C. Savariaux, "Degrees of freedom of facial movements in face-to-face conversational speech," in *Proceedings of the International Workshop on Multimodal Corpora*, pp. 33–36, Genoa, Italy, 2006.

[16] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images," *Journal of Phonetics*, vol. 30, no. 3, pp. 533–553, 2002.

[17] G. Bailly, A. Bégault, F. Elisei, and P. Badin, "Speaking with smile or disgust: data and models," in *Proceedings of the Auditory-Visual Speech Processing Workshop (AVSP '08)*, pp. 111–116, Tangalooma, Australia, 2008.

[18] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '95)*, pp. 660–663, Detroit, Mich, USA, 1995.

[19] M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi, "Text-to-audio-visual speech synthesis based on parameter generation from HMM," in *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH '99)*, pp. 959–962, Budapest, Hungary, September 1999.

[20] M. Giustiniani and P. Pierucci, "Phonetic ergodic HMM for speech synthesis," in *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH '91)*, pp. 349–352, Genova, Italy, September 1991.

[21] R. Donovan, *Trainable Speech Synthesis*, Department of Engineering, University of Cambridge, Cambridge, UK, 1996.

[22] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '00)*, vol. 3, pp. 1315–1318, Istanbul, Turkey, 2000.

[23] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," in *Proceedings of the 5th ISCA Speech Synthesis Workshop (SSW '04)*, pp. 191–196, Pittsburgh, Pa, USA, June 2004.

[24] T. J. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 1082–1089, 2006.

[25] O. Govokhina, G. Bailly, and G. Breton, "Learning optimal audiovisual phasing for a HMM-based control model for facial animation," in *Proceeding of the ISCA Speech Synthesis Workshop (SSW '07)*, Bonn, Germany, August 2007.

[26] O. Govokhina, G. Bailly, G. Breton, and P. Bagshaw, "A new trainable trajectory formation system for facial animation," in *Proceedings of the ISCA Workshop on Experimental Linguistics*, pp. 25–32, Athens, Greece, August 2006.

[27] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "An HMM-based singing voice synthesis system," in *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH '06)*, pp. 2274–2277, Pittsburgh, Pa, USA, September 2006.

[28] T. Okadome, T. Kaburagi, and M. Honda, "Articulatory movement formation by kinematic triphone model," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '99)*, vol. 2, pp. 469–474, Tokyo, Japan, October 1999.

[29] G. Bailly, G. Gibert, and M. Odisio, "Evaluation of movement generation systems using the point-light technique," in *Proceedings of the IEEE Workshop on Speech Synthesis*, pp. 27–30, Santa Monica, Calif, USA, 2002.

[30] E. Cosatto and H. P. Graf, "Sample-based of photo-realistic talking heads," in *Proceedings of the Computer Animation (CA '98)*, pp. 103–110, Philadelphia, Pa, USA, June 1998.

[31] T. Ezzat and T. Poggio, "MikeTalk: a talking facial display based on morphing visemes," in *Proceedings of the Computer Animation (CA '98)*, pp. 96–102, Philadelphia, Pa, USA, 1998.

[32] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[33] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, "Synthesizing realistic facial expressions from photographs," in *Proceedings of the 25th Annual Conference on Computer Graphics (SIGGRAPH '98)*, pp. 75–84, Orlando, Fla, USA, July 1998.

[34] B.-J. Theobald, S. Fagel, G. Bailly, and F. Elisei, "LIPS2008: Visual speech synthesis challenge," in *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH '08)*, pp. 2310–2313, Brisbane, Australia, September 2008.

[35] R. A. J. Clark, K. Richmond, and S. King, "Festival 2—build your own general purpose unit selection speech synthesiser," in *Proceeding of the ISCA Speech Synthesis Workshop (SSW '04)*, pp. 173–178, Pittsburgh, Pa, USA, June 2004.

[36] O. Govokhina, G. Bailly, G. Breton, and P. Bagshaw, "TDA: a new trainable trajectory formation system for facial animation," in *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH '06)*, pp. 2474–2477, Pittsburgh, Pa, USA, September 2006.

[37] P. Badin, F. Elisei, G. Bailly, and Y. Tarabalka, "An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data," in *Proceedings of the Articulated Motion and Deformable Objects (AMDO '08)*, vol. 5098 of *Lecture Notes in Computer Science*, pp. 132–143, Springer, Mallorca, Spain, July 2008.

[38] O. Engwall and J. Beskow, "Resynthesis of 3D tongue movements from facial data," in *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, pp. 2261–2264, Geneva, Switzerland, September 2003.