

Towards eyegaze-aware analysis and synthesis of audiovisual speech

Frédéric Elisei, Gérard Bailly, Alix Casari, Stephan Raidt

▶ **To cite this version:**

Frédéric Elisei, Gérard Bailly, Alix Casari, Stephan Raidt. Towards eyegaze-aware analysis and synthesis of audiovisual speech. AVSP 2007 - 6th International Conference on Auditory-Visual Speech Processing, Aug 2007, Hilvarenbeek, Netherlands. pp.50-56. hal-00169556

HAL Id: hal-00169556
https://hal.science/hal-00169556

Submitted on 4 Sep 2007

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.

Towards eye gaze aware analysis and synthesis of audiovisual speech

Frédéric Elisei, *Gérard Bailly*, *Alix Casari*, *Stephan Raidt*

Departement "parole et cognition", GIPSA-Lab, Grenoble Universités, France
{frederic.elisei,gerard.bailly,stephan.raidt}@gipsa-lab.inpg.fr

Abstract

Eye gaze plays many important roles in audiovisual speech, especially in face-to-face interactions. Eyelid shapes are known to correlate with gaze direction. This correlation is perceived and should be restored when animating 3D talking heads. This paper presents a data-based construction method that models the user's eyelid geometric deformations caused by gazing and blinking during conversation. This 3D eyelid and gaze model has been used to analyze and automatically reconstruct our German speaker's gaze. This can potentially complement or replace infra-red based eye tracking when it is important to collect not only where the user looks but also how (ocular expressions…). This method may be used as a tool to study expressive speech and gaze patterns related to cognitive activities (speaking, listening, thinking...).

Index Terms: eye gaze, eyelids, face-to-face interaction, talking heads, embodied conversational agents, audiovisual speech corpus analyzing.

1. Introduction

When interacting, people mostly gaze at face and gesturing. While speech is clearly audiovisual [1], facial expressions and gaze also inform us about the physical, emotional and mental state of the interlocutor. Together with speech and gesturing, face and gaze participate in signaling discourse structure, ruling turn taking and maintaining mutual interest.

This paper will focus on the eye region, studying the eyelid shape deformations that we could observe, along with eye gaze changes, in an expressive audiovisual speech corpus used to create a talking head (Figure 1).

Our long term research goal is to build an accurate control model for our next generation of context-aware ECAs (embodied conversational agents). We want them to be able to collaborate on some tasks using voice and pertinent gaze patterns (keeping eye contact, maintaining mutual attention, pointing at objects, displaying feeling of thinking...) with humans in their environment.

Therefore, audiovisual synthesis should be able to provide the eye region of the face with realistic appearance as well as pertinent behavior and control models. In the same interactive loop (imposing *real-time* scene analysis), ECAs should be reactive to the gaze patterns of their interlocutors and implement a complex set of interaction rules, as noted in [2]. For analysis as well as for synthesis, this remains a challenge covered by many researches from various domains.

2. Related research results

2.1. Importance and role of eye gaze

Eye gaze production and perception in human-human interaction has already attracted a lot of psychophysical and psychological studies. In conversation, gaze is involved in the regulation of turn taking, accentuation and organization of discourse [3, 4]. Most data on eye movement of perceivers during audiovisual speech perception have been gathered using non-interactive audiovisual recordings [5]. Several experiments have however shown that gaze patterns in live interaction are significantly different from screening: social rules have in fact a strong impact on communication when interacting face-to-face [6]. Gaze and eye-contact are also important cues for the development of social activity and speech acquisition [7].

Perceived gaze direction is biased by the head orientation [8]. To our knowledge, no study has measured the importance of the eyelid movements for gaze perception, although gaze direction clearly influences eyelid appearance.

2.2. Animating eye gaze or eyelids

Some statistical properties of the studied or measured eye gaze trajectories and blink rates have been evidenced and implemented to drive avatars and ECAs [9-13]. These gaze control models usually distinguish between several states of the user-ECA interaction (at least talking, listening, but sometimes much more complex ones gathered in the literature). While they often use rather realistic-looking 3D face models, the eyelid shapes are often not realistically correlated with the eye gaze direction. For example, Lee et al. [11] report: « During the analysis of eye-tracking images, we noticed a high correlation between the eyes and the eyelid movement which could be incorporated ».

The eye pupils and eyelids synthesis has gained increased interest in computer vision and image-based rendering with the potential development of audiovisual communication software (video calls and virtual conference systems). Modifying the displayed gaze orientation, and possibly the head orientation, is almost mandatory to restore the eye contact and all the related behavior. Many proposals that worked at correcting the pupil position and appearance have reported that not correcting the eyelids results in inappropriate expressive attitude being perceived [14], for

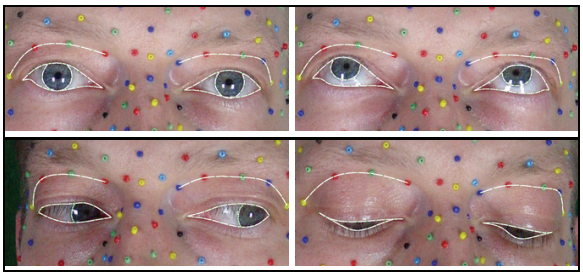


Figure 1: *Evidences from our learning corpus, showing different eyelid shapes and eye gaze directions. The color beads glued on the subject face serve as markers to construct the face model. White lines are 3D curves drawn manually on the images to recover the eye related contours (iris, eyelids…)*

example looking surprised if eyelids remain wide open while the gaze is modified to look downwards. In a shared mixed-reality space, Tateno et al. [15] demonstrated the benefit on gaze awareness of enhancing eyelid shapes.

Many 3D animated models of the eyes region have been used and proposed by computer graphics industry and researchers. Correcting eyelids to avoid interpenetration with eyeballs during the animation is necessary [16], but such algorithms fail to reproduce reality. Some highly detailed 3D models exist [see reviews in 17, 18], but eye and eyelids joint-control is usually not addressed. Moriyama et al. [18], whose model is generic enough to adapt across large subject variability as well as eye movements, uses independent sets of parameters to lower or raise either the eyelids or the eyeball. In [19], a somewhat high-level dynamic controller is proposed for their advanced EMOES model, but eyelid and eyeball controls seem to remain independent. In [20], Itti et al. report that they tried to improve the visual animation of their 3D head by raising eyebrows (effectively acting on eyelids) when gaze is lifted up.

2.3. Tracking eye gaze

Capturing real eye gaze data has never been so easy, with many industrial solutions using infrared light and sensors now available. But the least invasive solutions do not solve the generic problem of recovering the point looked at *in space* by a subject, they only capture the point looked at *on screen* (or any other plane surface) which may be sufficient for mediated face-to-face interaction [21]. More complex but invasive equipments fail to capture ecological face-to-face gazes when subject's face or eyes are partly hidden or look distracting. Some hardware or software relies on tracking *where* the head or the eyes are in space, but they generally do not capture *how* the eyes and eyelids are looking (ocular expressions…)

Computer vision research develops new solutions to compute more accurate/general solutions, but using multiple or single cameras seems promising [22-24].

3. Overview of our proposal

In the real world, eye gaze is important and eyelid deformations (as well as head movements) participate to the elaboration of the displayed and perceived gaze direction. In an ECA context, synthesized or captured gaze patterns should be rendered so that human partners may perceive the intended multimodal deixis and mutual attention. Adequate prediction of eyelid deformations together with gaze direction will reinforce the perception of spatial cognition.

For our audiovisual studies, we chose to capture and model motor activities of the upper face of one subject with a data-based approach. Section 4 presents the labeling of the training dataset, consisting of images that span the space of gaze directions and capture the eyelid deformations, either co-occurring or independent. Section 5 proposes a two-stage nonlinear model that can cope with pure rotations (as undergone by points on iris contour) or more complex transformations. Section 6 presents a preliminary analysis-by-synthesis algorithm, used to collect speech articulation and co-occurring gaze patterns from free speech video sequences.

4. The training dataset

The gaze related eyelid deformation model presented here is extending a 3D articulatory face and lips model of a German speaker. This talking head is presented and evaluated in another paper of these proceedings [25]. The initial model and the extended one were built using a common corpus. In this corpus, 400 colored beads were used as markers, glued on the subject's face and covering the area from one ear to another and from the forehead to the neck. Some of the markers were placed around the eyebrows, as can be seen on Figure 1.

To capture the face model as well as dynamic eye gaze stimuli, three synchronized analog video cameras (720x576 pixels, 50Hz, interlaced) were used to record the subject. In this setup, the subject's chest was hold in place and can be considered as static relatively to the three cameras. Using a calibration object, intrinsic parameters, epipolar geometry and relative positions of the video cameras were extracted. 3D coordinates in the world frame can then be estimated easily.

Using a selected subset of uttered visemes from the calibrated video cameras, we will first build a 3D articulatory face-only model, as described in [26]. In this methodology, the rigid movements of the head defined in reference to the bite plane are accurately subtracted from the measured 3D coordinates. Then, a modified PCA (principal component analysis) approach is applied, that recovers a set of 6 articulatory speech-related dimensions (jaw opening, lips rounding, upper lip control, lower lip control, jaw advance and throat lowering) and how they correlate with the displacements of the beads. In the reconstruction process, the rigid movements of the head, relative to the static chest, are also put back in the model with 6 extra movement parameters. They drive non rigid deformations in the neck area.

4.1. Extra data for the eye region

For the eye-area modeling phase, data with better spatial accuracy was collected using a single numerical photo camera (1200x700 pixels, not interlaced) to grab static postures of the full face (which are also used for texturing the face). In this setup, the subject was sitting on a desk chair that could be rotated: his position relative to the photo camera is not static anymore. The subject was successively asked to hold a constant gaze in nine fixed directions (far front, as much as possible on the right/left/up/down and the remaining corners). Closed eyelids were also recorded. For each hold gaze direction, the seat was slowly rotated to allow high-resolution images to be captured by the photo camera from three viewpoints (left, front and right views). Intrinsic parameters of the photo camera and its projection matrix were also recovered using our calibration object. Of course, 3D coordinates cannot be reconstructed from a single view. as will be explained later.

In our eye-modeling procedure, we first use the built face model to recover the spatial relationship of every coherent image triplets (images featuring the same recorded gaze). Fitting the model recovers the facial articulation (constant for image triplets) as well as the head/chest pose of every image. This is performed with *Matlab,* by minimizing the prediction error of beads positions while optimizing the articulatory and head posture parameters. In practice, this is quite equivalent to having three calibrated photo cameras in different positions for every recorded gaze. More important is the fact that this also recovers the head orientation and computes the head frame in which the modeling has to be conducted.

On every set of three coherent images and with the help of epipolar lines, an interactive editor is used to draw 3D Bezier curves. We outlined the iris contour, the upper and lower eyelids, as well as an eyelid anchor line under the eyebrows (see Figure 1).

To increase the statistical completeness, 24 image triplets were also selected in the dynamic low resolution corpus

(obtained with the three video cameras) in order to span a more complete range of gaze directions. 3D Bezier curves in the eye region were also edited on these images.

The iris contour will be used to estimate the gaze direction. The other contours will shape the 3D model. At this point, the curves are resampled using regularly spaced curvilinear coordinates to get a canonical representation with a constant number of points. All points have been transformed and expressed in the head coordinate system.

5. Modeling the eyeballs/eyelids

5.1. Our modeling hypotheses

Our first set of modeling hypotheses concerns the shape of the iris contour and its relation to the gaze direction. Like many researchers, we assume that this external iris contour is planar and circular. We also assume that the gaze direction is equal to the normal traversing the iris contour circle in its center. From these, we will capture the eye gaze directions and the eyeball rotations by observing the external iris contour.

Our second modeling hypothesis is that the eyelid deformations that correlate with the eyeball rotations can be captured in a two stage non-linear model, as was done for hand articulation in a cued speech context [27]. In the first stage of the model, angular parameters are driving trigonometric functions. The second stage of the model will be linear, predicting 3D coordinates from the trigonometric values. Such a two-stage model can capture pure rotations without a loss, but more complex anisotropic displacements or deformations can also be modeled.

A third modeling hypothesis is that remaining eyelid deformations can be modeled as extra linear contributions that add to the previous deformations. We expect at least one for blinking. This is not a strong assumption as a complex non-linear deformation can always be modeled by several linear contributions (it just adds more parameters than necessary and adds unnecessary complexity to control models).

Our last hypothesis will only affect the eyeballs rendering process: the subject's eyeball geometry can not be measured, so we used an average generic ratio to derive his eyeballs geometry from his iris radius.

5.2. Results of the modeling

From the labeled iris contours and for each eye, we compute the 3D coordinates of each iris center, the direction that is normal to the iris plane, as well as an iris radius. Fitting circles is robust enough that we recover an almost constant radius for the two irises across the high resolution images. This value has been used to ease the labeling in the 24 lower resolution images. The normal vector is expressed in the head frame as two angles *a1* (azimuth) and *a2* (elevation). In our two-stage modeling formalism, we are interested in the linear regression between 3D coordinates and a set of trigonometric values. Ideally, we would retain the sine and cosine from *a1*, from *a2*, from their difference and from their sum. These are necessary and sufficient to reconstruct the rotation matrix characterized by *(a1,a2)* by an additive formalism (because cos*(a1)*.cos*(a2)* and other similar product functions are linear combinations of the retained functions). In practice, we do not have enough training data to robustly recover all the model coefficients that would drive the 3D coordinates from the full set of trigonometric values. For robustness, we only retain sin*(a1+a2),* sin*(a1-a2),* cos*(a1)* and cos*(a2)*, as they best reconstruct the eyelid data variance. A linear regression of the centered training 3D coordinates using the trigonometric values as predictors recovers 66% and 61% respectively of the left and right eyelid coordinates variance.

We perform a standard principal component analysis (PCA) on the residual data, to catch extra degrees of freedom. As we expected, the first principal component acts along a blink dimension, narrowing or widening the upper and lower eyelids. This captures 40.9% and 30.8% of the residual data, for the right and left eyelids respectively. As the next principal components do not display clearly interpretable movements, we only retain the first principal component.

Using three parameters per eye (two angular ones and a linear one), 79.9% and 73% of the original data variance is captured (for the right and left eye, respectively).

Figure 2 presents the effects of our three parameters on eyelids geometry and gaze orientation. Both the upper and lower eyelid contours are deforming, especially for horizontal gaze movements. For vertical gaze movements, both are moving. The upper eyelid anchorage is not very active. Globally, the eyelid geometry preserves the vision field while protecting the rest of the eye. A geometric control model can easily generate the four low-level angular parameters from a target point in space to preserve the gaze vergence. Using homogeneous coordinates for the target point conveniently encodes cases with vergence or gazes at optical infinity.

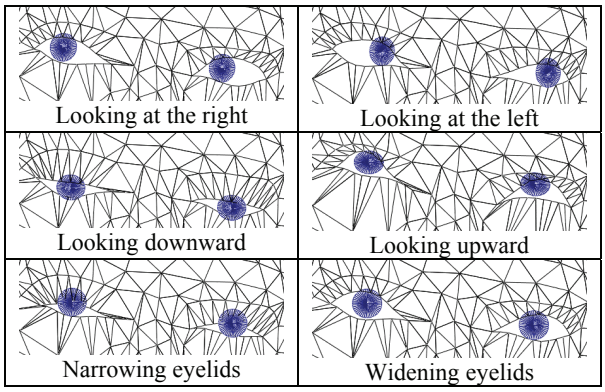


Figure 2: *Effects of each control parameters on the gaze and the eyelid geometry (in the figure, the same value is used for left and right eye parameters)*

5.3. Rendering of the model

Figure 3 presents textured 3D views of the merged model (6 speech-related articulatory parameters, 3 gaze and blink parameters per eye, and 6 head movement and orientation parameters). In the eye region, the crude mesh learned with beads only has been replaced by the animated high-definition eyelid region and 3D eyeballs. The eyelids are textured with a single static image. Rendering is performed using the *OpenGL* library. For each eyeball, we use a two-sphere model whose intersection matches the measured iris circle. The ratio of the radii of each pair of spheres reproduces an average ratio. To avoid 3D collisions between the eyeball and the eyelid geometries, the rendering is performed using *OpenGL* stencil buffer capability: the textured eyelids and face are rendered first and the stencil buffer is cleared. Then, the area comprised between the upper and lower eyelid is filled using invisible triangles to enable the stencil for these pixels. This defines two windows through which eyeball geometries will be rendered and become visible.

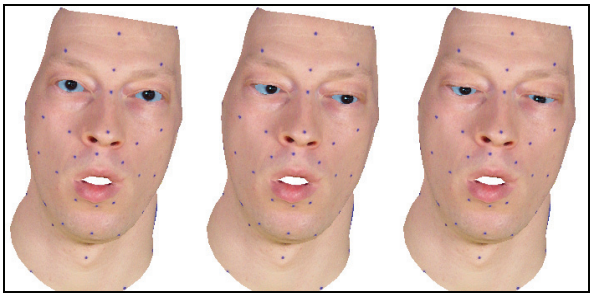
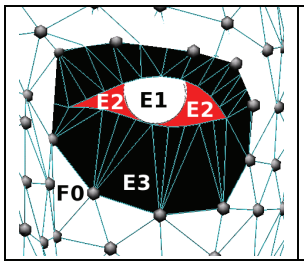


Figure 3: *Textured but non-photorealistic renderings of the model, while looking respectively far in front, on the left and slightly down, ahead of its mouth.*



F0 is the original mesh, connecting beads.
E3 is mainly a skin-colored area, including the eyelids.
E2 is mainly white (sclera from the eyeball), when visible (eyelids opened).
E1 is the visible part of the iris and pupil.

Figure 4: *Details of the extended mesh (left eye)*

Our rendering is not photorealistic yet: eyelashes should be modeled either by their shape or by their appearance; eyelids need multiple textures to capture wrinkles and creases appearance changes, and pertinent glints should be computed for the eyeballs. Despite these drawbacks, subjective tests where the face-model eyes track a moving 3D target show that the face looks more natural and that its gaze is perceived much more accurately than with static eyelids. Of course, objective tests need to be conducted to assess these results.

6. An algorithm for corpus analysis

The model we developed is not to be used in a 3D rendering context only. The motivation of our meticulous modeling is to use it also to analyze recordings of the cloned subject. To illustrate the interest of such a model in this context, we have been experimenting on our currently available data with the very simple four-step algorithm presented here.

This algorithm can capture eye gaze accurately if eyelids are open. It can capture blinks, but fails to robustly capture eye gaze when eyes are almost closed. Experimenting with a bigger corpus in the future, we will try to solve the shortcomings of our actual algorithm.

6.1. Recovering facial articulation

First, we recover both the six head movement parameters and the six articulatory speech parameters using the simpler model (no eyelids). As the rendered 3D model perfectly matches the recorded images, analysis by synthesis could have been used [28]. We actually used a pattern matching algorithm to track the beads displacements first. *Matlab* is used then to minimize the bead positions reconstruction error while optimizing the 12 parameters. At that point, the face model matches the video stimuli by its face orientation and speech articulation, but not by its eye gaze.

6.2. Estimating the blink parameter

It might have been convenient to have markers on the eyelids of our subject (as used in motion capture studios to detect blinks) but our corpus features none there. With a video realistic enough 3D model, we might compare directly each $RGB(x,y)$ pixel at position (x,y) in a camera image with its $R_SG_SB_S(x,y)$ counterpart in the synthesized image. We avoid using a prior segmentation of the images to evade the introduction of coarse segmentation errors. These errors seem very likely with the relatively low coverage of the eye region for a full-face corpus at video camera resolutions.

Instead, our second step uses an *indirect* analysis-by-synthesis loop to estimate the value of each blink parameter (1 per eye). We chose to reinforce the red component of the eyelid region to contrast with the eye ball region, filled by white and iris colors. We propose to represent a pixel at position (x,y) in a camera image by its normalized red component:

$$p(x,y) = \frac{R(x,y)}{\|RGB(x,y)\|} \tag{1}$$

For every triplet value (a_1,a_2,v) of gaze and blink values, the eyelid model and the camera model predict where eyeballs and eyelids are projected (see Figure 4 for the defined regions). For correct parameters values, pixels covered by the *E3* region should have a stronger red bias than pixels in the *E1* and *E2* regions. *E3* is defined so that its external contour does not depend from the gaze and blink parameters: the pixels covered by *E1+E2+E3* depend only from facial articulation and head posture. Using a_1=0 and a_2=0 as initialization, we first optimize v only to minimize:

$$Err_v(a_1,a_2,v) = \sum_{(x,y)\in E3}(1-p(x,y)) + \sum_{(x,y)\in E1\cup E2} p(x,y) \tag{2}$$

The inversion process is to be continued with the two following steps only if the resulting v value does not correspond to a completely closed eyelid.

6.3. Estimating the gaze direction

In the "window" opened by the eyelids, iris and pupil pixels contrast with the white part (sclera). Considering only the red planes (camera image and synthetic image), we found it robust enough to use a *direct* analysis-by-synthesis approach at this point. For this, we chose two suitable colors to draw the synthesized pixels corresponding to E1 and E2. Picking the average color of the two regions in a camera image and amplifying their contrast was successful enough with our subject.

Preserving the previous value of the blink parameter v, together with the defined *E1* and *E2* regions, we optimize for the two eye gaze angles *a1* and *a2* to globally match the image iris contour by minimizing the image reconstruction error (in the red channel):

$$Err_a(a_1,a_2,v) = \sum_{(x,y)\in E1(0,0,v)\cup E2(0,0,v)} |R(x,y) - R'(x,y)| \tag{3}$$

Contrary to v variations that might have subtle influences, *a1* and *a2* usually have easily visible influence in some of the camera images, and non marginal effect on equation (3).

6.4. Final inversion

At this point, the three parameters values are only estimations, not having been optimized at the same time. Using them as initial values to minimize again equation (2) then (3) as in the *linesearch* algorithm leads to better solution.

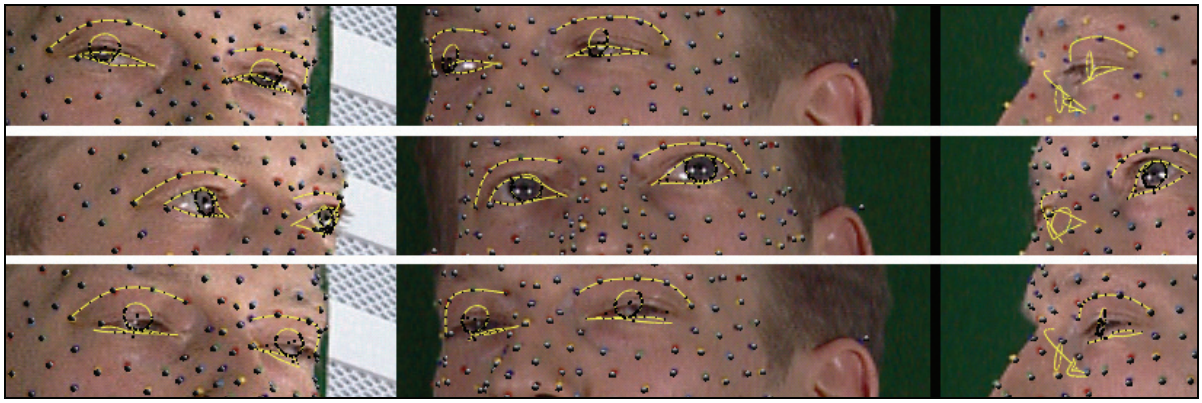


Figure 5: *Tracking results of the proposed algorithm. The three used camera views are shown horizontally. Curves recovered from the model are displayed in bright yellow: eyelids anchorage, eyelids and complete iris circle. The two first lines show succesfull trackings. The last line demonstrates an inversion error, with almost closed eyelids (see text).*

6.5. Results

The proposed algorithm has been implemented using *OpenGL*. Using a PCI x16 3D accelerator, a new set of 6 gaze/blink parameters is recovered every 800ms, using our three views setting. Figure 5 shows some results of our tracking procedure, with a few example frames. The accompanying AVSP DVD (see also [29]) includes videos of analyzed and reconstructed audiovisual recordings. Figure 6 shows the trajectories of the parameters for a subpart of the sequence.

One can notice that eyelashes can lead to over-estimation of the blinking parameter. This is no surprise, as the pixels covered by the dark hairs are classified as closer to the iris color than to the eyelid skin color. This bad classification might also explain the difficulties in recovering the gaze direction when the eyelids are almost closed.

7. Conclusions

We presented a methodology to construct an articulatory model of the shape deformation for the eyelids of a virtual talking head that takes into account the eye gaze direction. A model has been built from the data of a German speaker. It is driven by three low-level parameters per eye: two for gaze direction and one for the proper opening/closing movements of the eyelids.

We will assess the effectiveness of the eyelid deformation model and its positive impact on the perceived gaze in the deictic bootstrapping paradigm already used in [30] with our previous ECA generation.

This model was used to analyze and reconstruct some audiovisual recordings of the cloned subject in a face-to-face interaction. While the inversion algorithm needs to be corrected to cope with closed eyes (post-processing the data in the blinking intervals) or almost closed ones (drawing eyelashes), the model itself predicts satisfactorily the observed deformations. We might also compare the accuracy of the algorithm with that of our infrared based eye-tracker, or check the robustness of the algorithm (to slight head pose changes, image noise or image resolution) with synthetic images.

The recovered eye movements will complement the gestural score driving our ECAs. The automatic generation of such scores is particularly interesting for studying the gestural encoding of linguistic and paralinguistic information in discourse production with less manual labeling [31].

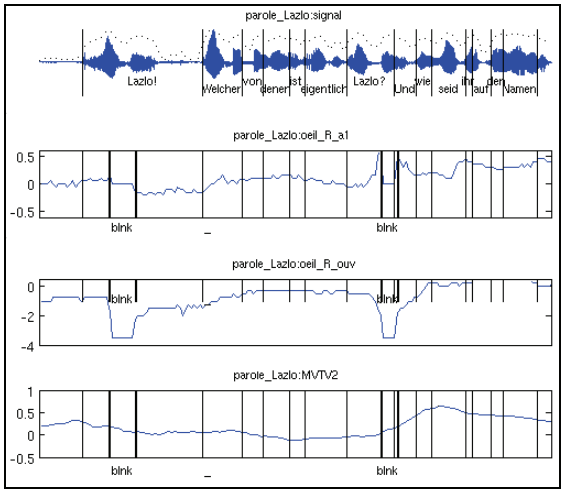


Figure 6: *Analyzing an extract of conversational speech. From top to bottom, using the same time-scale: acoustic signal with orthographic transcription, azimuth angle for the right eye, blink parameter for the right eye and head nodding parameter. The articulatory scores were all produced by automatic inversion processes from the images (3 views) and the models. Two blinks are enlightened in the parametric tracks. They cue end of utterances in order to signal turn keeping. Note that gaze patterns and head motion are quite nicely synchronized with phrases.*

Creating such a 3D model per subject is quite tedious, especially as it needs cloning the full face also. While we believe this data-based approach to be worth the effort when the accurate measurement or analysis of audiovisual stimuli is needed, we plan to reuse the already built model in a 3D synthesis context: with proper scaling and moderate change of orientation, the two eye regions can be merged on another talking agent and improve the perception of gaze.

8. Acknowledgements

Parts of this work were supported within the PROCOPE program, the GIS PEGASUS, the ELESA research federation and the Presence project of the Cluster Rhône-Alpes InfoLog. We want to thank Christophe Savariaux and Alain Arnal for their technical help in the capture process and Ralf Baumbach who helped with annotating the data. We are also grateful to Sacha Fagel for accepting to be our subject.

9. References

[1] D. G. Stork and M. E. Hennecke, *Speechreading by Humans and Machines*. Berlin, Germany: Springer, 1996.

[2] K. Thórisson, "Natural turn-taking needs no manual: computational theory and model from perception to action," in *Multimodality in language and speech systems*, B. Granström, D. House, and I. Karlsson, Eds. Dordrecht, The Netherlands: Kluwer Academic, 2002, pp. 173–207.

[3] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22-63, 1967.

[4] M. Argyle and M. Cook, *Gaze and mutual gaze*. London: Cambridge University Press, 1976.

[5] E. Vatikiotis-Bateson, I.-M. Eigsti, S. Yano, and K. G. Munhall, "Eye movement of perceivers during audiovisual speech perception," *Perception & Psychophysics*, vol. 60, pp. 926-940, 1998.

[6] M. Gullberg and K. Holmqvist, "Visual attention towards gestures in face-to-face interaction vs on screen," presented at International Gesture Workshop, London, UK, 2001.

[7] M. Carpenter and M. Tomasello, "Joint attention, cultural learning and language acquisition: Implications for children with autism," in *Communicative and language intervention series. Autism spectrum disorders: A transactional perspective*, vol. 9, A. M. Wetherby and B. M. Prizant, Eds. Baltimore: Paul H. Brooks Publishing, 2000, pp. 30–54.

[8] S. R. H. Langton, "The mutual influence of gaze and head orientation in the analysis of social attention direction," *Quarterly Journal of Experimental Psychology*, vol. 53A, pp. 825-845, 2000.

[9] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, "ANIMATED CONVERSATION: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents," presented at ACM SIGGRAPH, Orlando, Florida, 1994.

[10] R. A. Colburn, M. F. Cohen, and S. M. Drucker, "The Role of Eye Gaze in Avatar Mediated Conversationnal Interfaces," Microsoft Research, Technical report MST-TR-2000-81, 2000.

[11] S. P. Lee, J. B. Badler, and N. Badler, "Eyes alive," *ACM Transaction on Graphics*, vol. 21, pp. 637-644, 2002.

[12] M. Garau, M. Slater, V. Vinayagamoorthy, A. Brogni, A. Steed, and M. A. Sasse, "The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual Environnment," presented at SIGCHI, Lauderdale, Florida, USA, 2003.

[13] M. Bilvi and C. Pelachaud, "Communicative and statistical eye gaze predictions," presented at International conference on Autonomous Agents and Multi-Agent Systems (AAMAS), Melbourne, Australia, 2003.

[14] J. Gemmell, K. Toyama, L. C. Zitnick, T. Kang, and S. Seitz, "Gaze awareness for video-conferencing: A software approach," *IEEE Multimedia*, vol. 7, pp. 26 - 35, 2000.

[15] K. Tateno, M. Takemura, and Y. Ohta, "Enhanced eyes for better gaze-awareness in collaborative mixed reality," presented at International Symposium on Mixed and Augmented Reality, Vienna, Austria, 2005.

[16] F. I. Parke and K. Waters, "Section 6.3, Implementation of a Direct Parametrized Model," in *Computer Facial Animation*. Wellesley, MA, USA: A.K. Peters, 1996.

[17] M. A. Sagar, D. Bullivant, G. D. Mallinson, and P. J. Hunter, "A virtual environment and model of the eye for surgical simulation," presented at SIGGRAPH, Orlando, FL, 1994.

[18] T. Moriyama, J. Xiao, J. Cohn, and K. T., "Meticulously detailed eye model and its application to analysis of facial image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28(5), pp. 738-752, 2006.

[19] N. Adamo-Villani, G. Beni, and J. White, "EMOES: Eye Motion and Ocular Expression Simulator," *International Journal of Information Technology*, vol. 2(3), pp. 170-176, 2005.

[20] L. Itti, N. Dhavale, and F. Pighin, "Photorealistic Attention-Based Gaze Animation," presented at IEEE International Conference on Multimedia and Expo, Toronto, Canada, 2006.

[21] S. Raidt, G. Bailly, and F. Elisei, "Mutual gaze during face-to-face interaction," presented at Auditory-visual Speech Processing, Hilvarenbeek, The Netherlands, 2007.

[22] K.-H. Tan, D. J. Kriegman, and N. Ahuja, "Appearance-based Eye Gaze Estimation," presented at the 6th IEEE Workshop on Applications of Computer Vision, Orlando, Florida, 2002.

[23] J.-G. Wang, E. Sung, and R. Venkateswarlu, "Eye Gaze Estimation from a Single Image of One Eye," presented at the 9th IEEE International Conference on Computer Vision, Nice, France, 2003.

[24] C. Djeraba, "State of the art of Eye tracking," LIFL, Lille, Publication interne 07, 2005.

[25] S. Fagel, G. Bailly, and F. Elisei, "Intelligibility of natural and 3D-cloned German speech," presented at Auditory-visual Speech Processing, Hilvarenbeek, The Netherlands, 2007.

[26] G. Bailly, F. Elisei, P. Badin, and C. Savariaux, "Degrees of freedom of facial movements in face-to-face conversational speech," presented at International Workshop on Multimodal Corpora, Genoa - Italy, 2006.

[27] G. Gibert, G. Bailly, D. Beautemps, F. Elisei, and R. Brun, "Analysis and synthesis of the 3D movements of the head, face and hand of a speaker using cued speech," *Journal of Acoustical Society of America*, vol. 118, pp. 1144-1153, 2005.

[28] M. Odisio and G. Bailly, "Tracking talking faces with shape and appearance models," *Speech Communication*, vol. 44, pp. 63-82, 2004.

[29] "http://www.icp.inpg.fr/~elisei/AVSP07 (Paper-related videos)."

[30] S. Raidt, G. Bailly, and F. Elisei, "Does a virtual talking face generate proper multimodal cues to draw user's attention towards interest points?," presented at Language Ressources and Evaluation Conference (LREC), Genova, Italy, 2006.

[31] J.-C. Martin, G. Caridakis, L. Devillers, K. Karpouzis, and S. Abrilian, "Manual Annotation and Automatic Image Processing of Multimodal Emotional Behaviours: Validating the Annotation of TV Interviews," presented at Language Ressources and Evaluation Conference (LREC), Genoa, Italy, 2006.