Degrees of freedom of facial movements in face-to-face conversational speech

Gérard Bailly, Frédéric Elisei, Pierre Badin & Christophe Savariaux
Institut de la Communication Parlée, UMR CNRS n°5009, INPG/Univ. Stendhal
46, av. Félix Viallet, 38031 Grenoble CEDEX, France
{gerard.bailly,frederic.elisei,pierre.badin,christophe.savariaux}@icp.inpg.fr

ABSTRACT
In this paper we analyze the degrees of freedom (DoF) of facial movements in face-to-face conversation. We propose here a method for automatically selecting expressive frames in a large fine-grained motion capture corpus that best complement an initial shape model built using neutral speech. Using conversational data from one speaker, we extract 11 DoF that reconstruct facial deformations with a average precision less than a millimeter. Gestural scores are then built that gather movements and discursive labels. This modeling framework offers a productive analysis of conversational speech that seeks in the multimodal signals the rendering of given communicative functions and linguistic events.
Author Keywords: Facial movements, model-based face tracking, expressive audiovisual speech

INTRODUCTION
When we interact with each other and even in absence of the interlocutor (e.g. when phoning), facial movements due to speech articulation are often accompanied by head movements, facial expressions and gestures, used by the speaker for underlining the meaning of the speech acts, involving the listener or elements of the environment in the discourse as well as maintaining mutual attention by back channeling. These facial movements can aid the understanding of the message, but also convey a lot of additional information about the speaker, such as his emotional or mental state. Nonverbal components in face-to-face communication have been studied extensively, mainly by psychologists. Studies typically link head and facial movements or gestures qualitatively to speech acts. Many of the more prominent movements are clearly related to the discourse content or to the situation at hand. For example, if the sets and releases of eye contact are of most importance for face-to-face interaction, much of the body language in conversations is used to facilitate turn-taking. Movements also emphasize a point of view. Some movements serve biological needs, e.g. blinking to wet the eyes. Few quantitative results have been published that clearly describe what are the basic components of the facial movements, what are their precise region of action and how they combine, and finally how such head and facial movements correlate with elements of the discourse. Eckman and Friesen studied extensively emotional expressions of faces [10] and also describe non-emotional facial movements that mark syntactic elements of sentences, in particular endings. The appropriate generation of face, hand and body movements is of most importance for Embodied Conversational Agents [5, 6] as well as for Sociable Robots [4]. The rules governing the firing of mimics and the implementation of that mimics are however often set in a very ad hoc way and results generally from intensive labeling of videos recordings with no special focus on fine-grained motion capture.
Face detection, identification and tracking as well as facial movement tracking use generally model-based approaches where speaker-specific appearance and shape models should be learned from training data [8, 9, 12]. The number of free dimensions of these models heavily influences the system's performance: this number should offer a compact search space without sacrificing a good fit with observed movements. Initial models are often trained off-line using limited hand-labeled data. Most models consider either facial expressions [16] or speech-related facial gestures [15], with few attempts that treat the global problem [3].
The work below presents our first effort in characterizing the DoF of the facial deformation of one speaker when involved in face-to-face conversation. So-called expressemes – for expressive visemes – are extracted from life interaction videos for building shape models with minimal training data. A methodology is proposed to incrementally refine these models by automatically selecting pertinent expressemes.

THE CORPORA AND RECORDING PROCEDURE
For six years we have developed a procedure for building speaker-specific fine-grained shape [17] and appearance models [11] for the face and the lips: we glue more than 200 colored beads on the speaker's face to have access to fleshpoints. We also fit generic teeth, eyes and lips models to photogrammetric video data to regularize geometric data of these important but smaller organs. Corpora were generally dedicated to the study of speech coarticulation and limited to read material as our target application was multimodal text-to-speech synthesis [2]. The speech-related facial movements of seven speakers (2 females, 5 males) with different mother languages (Arabic, English, German, French) have been "cloned". Shape models of all speakers are controlled with the 6 articulatory parameters controlling the jaw, lips and laryngeal positioning.
More recently we extended the recorded material to basic

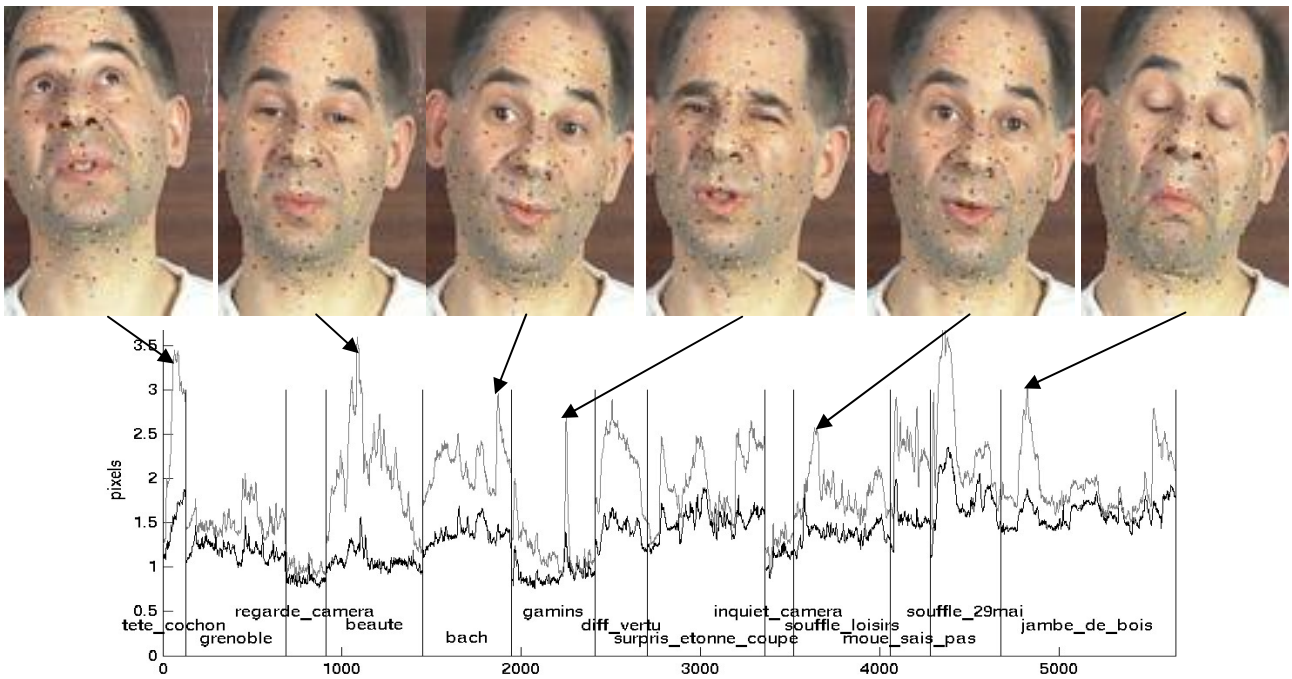

Figure 1: Comparing prediction errors of facial shapes using a model built using 52 speech visemes (light gray) with one incorporating 102 additional expressemes (dark gray), for a series of selected video sequences. The mean error lowers from 1.7 to 1.3 pixels. Frames shown at the top are generating the most important prediction errors of the speech-only model.

acted expressive speech (i.e. smiling, disgust) that most influence lip shape and to free conversation where subjects were asked first to answer to the Proust's questionnaire and then to recall and tell the most enjoyable, the most frightening and the most surprising personal experiences to the experimenter. We study here a corpus of free conversation from one subject lasting approximately 30 minutes. The speaker is filmed with three calibrated PAL cameras (front + both sides). The resulting images have a definition of almost 2 pixels per mm.

MODELLING SPEECH-RELATED MOVEMENTS

Data-driven shape models are classically built using principal component analysis (PCA). Usually a generic mesh is fitted by hand on a few dozen of representative frames. Shape parameters emerging from a PCA performed on these frames are often very difficult to interpret: they often mirror fortuitous correlations observed in the limited set of training material. Key frames should thus be chosen carefully so as to represent the diversity of facial movements usually involved in the task with maximum statistical coverage. We propose here to combine automatic feature point tracking, frame hand-labeling and statistical modeling to gather these key frames. Furthermore our shape models are built using a so-called guided PCA where a priori knowledge is introduced during the linear decomposition. We in fact compute and iteratively subtract predictors using carefully chosen data subsets [1]. For speech movements, this methodology enable us to extract six components directly related to jaw, proper lip movements and clear movements of the throat linked with underlying movements of the larynx and hyoid bone. We added to these six components two additional "expressive" components involved in our acted corpus of expressions: "smile" and "disgust" gesture that emerge from the analysis of our set of "smiling" and "disgust" visemes respectively.

TRACKING LOCAL FACIAL DEFORMATIONS

The speech-related shape model of the facial movements is then used to guide a multi-view tracker of the beads using correlation-based techniques [14]. The initial shape model only helps us to constrain the search space within proper regions of interest for each vertex of the facial mesh. The entire corpus of free conversation is then tracked. While most beads are tracked using at least two views, which enable 3D constraints to be applied, some beads are only tracked on one view, notably those located on the speaker's profile or in regions with high curvature.

The beads are tracked as patterns of 13x11 pixels. We track usually around 600 patterns per frame (compared to 250 beads on 3 views). The processing time for each frame is typically 2 seconds on a standard 2Ghz PC. We then interpret the reconstruction error of the beads summed up on all views (see Figure 1). We select automatically discourse units that have the most important reconstruction errors. We then retain the most salient frames which are precisely marked by hand, adding here the untracked beads.

ADDING SOME EXPRESSIVE FACIAL MOVEMENTS

The final objective of this selection process is to whiten the error structure by identifying and adding necessary DoF of

Figure 2: Non photorealistic synthetic views showing the effect on shape of each elementary expressive action. The face is rendered using a unique texture. From left to right: neutral, raising eyebrows, lip corners raiser (obtained from the smiling visemes), nose wrinkler (obtained from the visemes uttered with disgust), un-frowning and chin raiser. Note that lip corners raiser, nose wrinkler and chin raiser do affect lip shape.

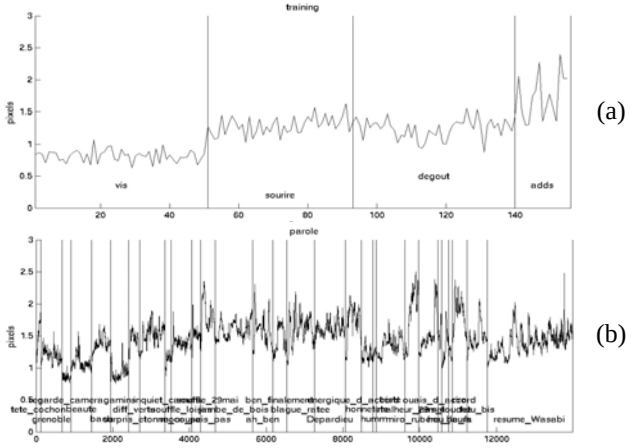

Figure 3: Modeling and tracking errors with the final model. (a) modeling errors of training data (visemes and expressemes). (b) tracking errors for selected conversational speech sequences.

unexplained facial movements. The analysis of the selected frames reveals for example an important residual error in the region of the forehead. Three basic components have been identified and added using first principal components of given regions of selected frames: eyebrows raising/lowering, forehead frowning, and chin raising/lowering. These elementary gestures combine to control facial shapes. The effects of single elementary gestures with reference to the neutral face are shown in Figure 2. The shape model that includes speech-related and expression-related facial movements has finally 11 DoF.

ANALYSING THE EXPRESSIVE CORPUS

The beads positions for each frame of the set of read and conversational speech data has been estimated using the beads tracker. The facial movements should now be explained by the DoF of the final shape model. The Figure 3 presents the modeling and tracking errors for several thousands of frames. The modeling error for the 154 training frames is less than 1 pixel for visemes and around 1,5 pixels for expressemes (Figure 3a). Note that all these frames have been manually marked. The tracking of beads on the entire sequences from which the visemes and expressemes have been extracted reaches almost the same precision. Note that this tracking error is computed using only the positions of tracked beads (usually 75% of the beads set). Despite the fact that the full model reduces the mean tracking error at around 1,3 mm (see Figure 1 and Figure 3b) and tends to decrease the number of error bursts, significant modeling errors still remain that claim for extra DoF to be added to the final shape model (see Figure 3b).

Reliable gestural scores can be built (see Figure 4) that gather the time evolution of the shape parameters together with the speech signal and discourse labels. These gestural scores provide very valuable data on synchronization of multimodal events that participate to the encoding of distinctive communicative functions: these scores provide the necessary receptacle of ground-truth bottom-up events and theory-specific top-down interpretations.

A CASE STUDY: EYEBROWS MOVEMENTS

Eyebrows movements are known to contribute to discourse structuring [7] and are often used as redundant markers of emphasis [13]. Our preliminary analysis of 20 turns of our conversational speech data evidences two distinct eyebrows gestures as displayed in Figure 5: bursts associated with words on emphasis that co-occur with pitch accents and more global gestures coextensive with dialog acts.

COMMENTS

MPEG4/SNHC identifies 64 Facial Animation Parameters. Similarly the well-known FACS individualizes 28 facial elementary gestures – not including eyes, eyelids and head movements - that combine to produce facial mimics. It is still an open question to determine (a) what are the basic synergies between these elementary gestures that are required to encode the complex repertoire of facial mimics; (b) how they effectively combine and how they are controlled; and (c) how speaker-specific strategies implement universal or culture-specific facial attitudes.

We claim here that this repertoire may be learnt using limited resources i.e. recording a limited set of visemes and expressemes and that a dozen of basic gestures is sufficient to reach a prediction error of about one millimeter uniformly distributed all over the face.

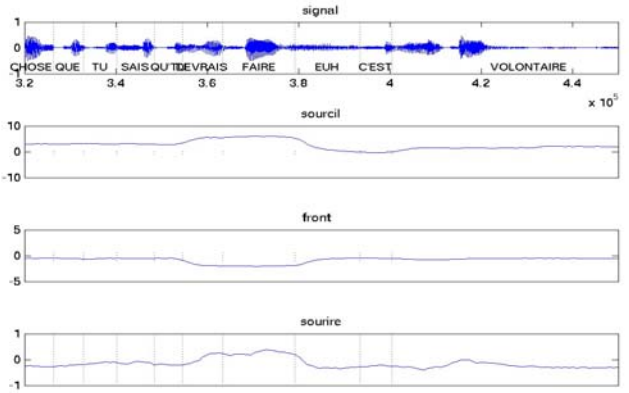

Figure 4: Gestural score for a selected speech act showing a burst of smiling ("sourire" score) during the uttering of "devrais faire". The following hesitation "euh" is also associated with lower eyebrows ("sourcil" score).

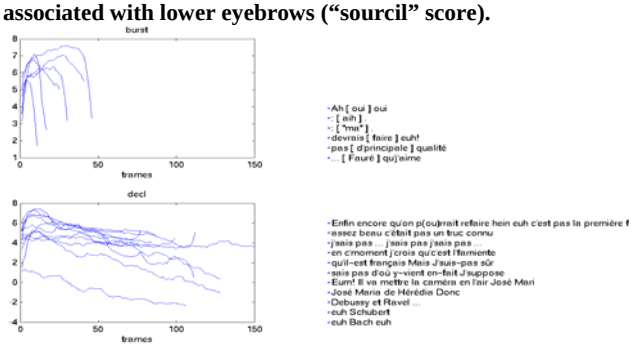

Figure 5: Time course of the eyebrow parameter. Top: bursts associated with words on emphasis. Bottom: initial burst+ declination associated with entire dialog acts.

CONCLUSIONS AND PERSPECTIVES

A productive analysis of conversational speech should combine two complementary approaches: a top-down approach that seeks in the multimodal signals the rendering of given communicative functions and linguistic events; and a bottom-up approach that reveals multimodal events that emerge from the observation of human partners in action. Combining both approaches will avoid the observer's biases and opens the route towards proper quantitative models of control and negotiation between overlapping scopes of communicative functions. The analysis of multimodal events should also be driven by entropy constraints i.e. implement coherently co-occurring communicative functions and not only result/emerge from global energy-based analysis such as PCA.

We will label gestural scores produced by our model-based gesture-aware tracker with communicative functions in order to study the scope and dynamics of their multimodal gestural instances.

ACKNOWLEDGMENTS

We thank Alain Arnal for his technical help and Ralf Baumbach for his preliminary work on the data. The paper benefited from the comments of 4 reviewers. This work has been financed by a PROCOPE grant between ICP and TUB, the GIS PEGASUS and the Rhône-Alpes region.

REFERENCES

[1] Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., and Savariaux, C. (2002) *Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images.* Journal of Phonetics, **30**(3): p.533-553.

[2] Bailly, G., Bérar, M., Elisei, F., and Odisio, M. (2003) *Audiovisual speech synthesis.* International Journal of Speech Technology, **6**: p.331-346.

[3] Beskow, J. and Nordenberg, M. (2005) *Data-driven synthesis of expressive visual speech using an MPEG-4 talking head*. in *Interspeech*. Lisbon, Portugal. p.793-796.

[4] Breazeal, C. (2000) *Sociable machines: expressive social exchange between humans and robots*. Sc.D. dissertation, in *Department of Electrical Engineering and Computer Science*. MIT: Boston, MA.

[5] Buisine, S., Abrilian, S., and Martin, J.-C. (2004) *Evaluation of multimodal behaviour of embodied agents*, in *From brows to trust: evaluating embodied conversational agents*, Z. Ruttkay and C. Pelachaud, Editors. Kluwer Academic Publishers. p. 217-238.

[6] Cassell, J. and Thórisson, K.R. (1999) *The power of a nod and a glance: envelope vs. emotional feedback in animated conversational agents.* International Journal of Applied Artificial Intelligence, **13**(4-5): p.519-538.

[7] Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., and Espesser, R. (1996) *About the relationship between eyebrow movements and F0 variations.* in *International Conference on Speech and Language Processing*. Philadelphia, PA. p.2175-2178.

[8] Cootes, T.F., Edwards, G.J., and Taylor, C.J. (2001) *Active Appearance Models.* IEEE Transactions on Pattern Analysis and Machine Intelligence, **23**(6): p.681-685.

[9] 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.

[10] 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.

[11] Elisei, F., Bailly, G., Gibert, G., and Brun, R. (2005) *Capturing data and realistic 3D models for cued speech analysis and audiovisual synthesis*. in *Auditory-Visual Speech Processing Workshop*. Vancouver, Canada

[12] Guenter, B., Grimm, C., Wood, D., Malvar, H., and Pighin, F. (1998) *Making faces*. in *SIGGRAPH*. Orlando - USA. p.55-67.

[13] Krahmer, E., Ruttkay, Z., Swerts, M., and Wesselink, W. (2002) *Pitch, eyebrows and the perception of focus*. in *Speech Prosody*. Aix en Province, France. p.443-446.

[14] Lewis, J.P. (1995) *Fast Template Matching.* Vision Interface: p.120-123.

[15] Odisio, M. and Bailly, G. (2004) *Tracking talking faces with shape and appearance models.* Speech Communication, **44**(1-4): p.63-82.

[16] Pighin, F.H., Szeliski, R., and Salesin, D. (1999) *Resynthesizing facial animation through 3D model-based tracking.* International Conference on Computer Vision, **1**: p.143-150.

[17] 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.