Speaking with smile or disgust: data and models

Gérard Bailly, Antoine Bégault, Frédéric Elisei & Pierre Badin

Dept. of Speech & Cognition, GIPSA-Lab, CNRS - Grenoble Universities, France

Gerard.Bailly@gipsa-lab.inpg.fr

Abstract

This paper presents preliminary analysis and modelling of facial motion capture data recorded on a speaker uttering non-sense syllables and sentences with various acted facial expressions. We analyze here the impact of facial expressions on articulation and determine prediction errors of simple models trained to map neutral articulation to the various facial expressions targeted. We show that movement of some speech organs such as the jaw and lower lip are relatively unaffected by the facial expressions considered here (smile, disgust) while others such as the movement of the upper lip or the jaw translation are quite perturbed. We also show that these perturbations are not simply additive, and that they depend on articulation.

Keywords: talking faces, speech, facial expression.

1 Introduction

The overwhelming majority of studies dedicated to the perception of emotion has concentrated on face perception. Since Ekman's pioneer work on facial expressions [7], there is an abundant literature on analysis [13], synthesis [16], perception [11] and recognition [9] of facial expressions displayed by natural and synthetic faces. When requested to determine emotional or affective state of speakers from videos, viewers are likely to give more importance to static and dynamic facial features of the upper face. This has been shown by blending emotion, hiding parts of the face [4, 5] or by changing the viewing angle [10]. Identification performance however depends on the facial expressions considered: if viewers are not able to discriminate angry vs. fear by seeing the lower face [5], facial expressions such as smiling or disgust have a clear visible impact on the lips region. These two facial expressions however interact with speech articulation and no objective nor subjective data are currently available to determine the impact of the joint articulation on the production and identification of both speech and emotion.

For almost two decades an important research effort has also been devoted to the audible dimensions of emotion: objective and subjective impact of emotion on intonation, rhythm or voice quality is now demonstrated. Listeners have been shown to be able to distinguish subtle differences between acted and true expressions, and between mechanical smile and amusement [1, 19].

Facial expressions during speech have clear audible and visible consequences; audiovisual cues are integrated, resulting in cross modal bias [4] and multimodal robustness.

To our knowledge there are however no experiments aiming at quantifying the articulatory strategies implemented by speakers to cope with the concomitant production of speech and facial expressions. Articulatory specifications for phonetic segments and facial expressions should be blended, resulting sometimes in solving contradictory instructions, e.g. spreading lips for smiling when producing rounded vowels or consonants.

The paper presents a preliminary analysis of articulatory reorganization of phonetic segments when facial expressions impacting strongly on the lower face are imposed. Very simple additive models are proposed and evaluated to cope with this articulatory reorganization. Section 2 presents the experimental setup and motion capture data used in this experiment. Section 3 provides details on the methodology that identifies the articulatory degrees-of-freedom (DoF) of the facial deformations observed in neutral and expressive speech. We analyze in Section 4 the main articulatory compensations and reorganizations due to facial expressions. Prediction errors resulting from the mapping of neutral speech to expressive speech by various models are discussed in Section 5.

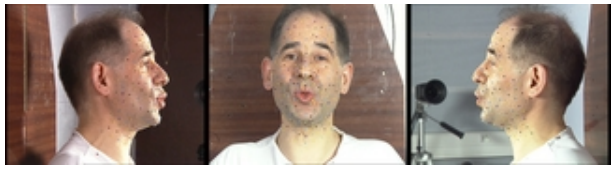

Figure 1: Synchronous multiview recordings of the speaker with 251 colored beads on his face.

Figure 2: The front image corresponding to the first audible frame of a /aba/ sequence uttered with three facial expressions: neutral, smile and disgust.

2 Motion capture data

For almost a decade we have developed a procedure for building speaker-specific fine-grained shape [17] and appearance models [8] for the face and the lips: we have used this procedure for the present study and glued 251 colored beads on the speaker's face to mark the fleshpoints (see Figure 1). The speaker is filmed with three synchronous calibrated PAL cameras. The resulting images have a definition of almost 2 pixels per mm. We also fit generic teeth, eyes and lips models to photogrammetric video data to complement the geometric data with shapes of these important organs.

More recently we extended the recordings of the speaker to: (1) basic acted expressive speech (i.e. smile, disgust) that most influence lip shape, (2) free conversation where the speaker was asked to answer the Proust's questionnaire and to tell the

experimenter his or her most enjoyable, most frightening and most surprising personal experiences.
We analyze here the corpus of acted expressive speech from one speaker (see Figure 2). Note that this corpus is used for bootstrapping the development of the speaker's clone used to analyze the whole corpus including free conversation [3].

3 Statistical articulatory model

Our statistical shape models are built using a so-called guided PCA where *a priori* knowledge is introduced during the linear decomposition. In fact, we compute and iteratively subtract predictors using carefully chosen data subsets [2]. For speech movements, this methodology enables 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. The face and lips model is controlled by six parameters for neutral speech. 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 that characterises the /i/ vs. /y/ opposition; *lips2* controls the upper lip raising / lowering movement, useful to realise the labio-dental consonant /f/ for instance; *lips3* controls the lower lip lowering / raising movement found in consonant /zh/ 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 instance. Note finally a parameter *lar1* related to a movement of larynx lowering. Altogether, more than 90% of the data variance in the neutral speech learning corpus is taken into account by these components. We have supplemented these six components with two additional "expressive" components involved in our acted corpus of expressions: "smile" and "disgust" gestures that emerge from the analysis of our set of "smile" and "disgust" visemes respectively. These gestures do not really differ from AU12 and AU9 of the FACS system: predictors of the movement are effectively the 1st principal component of the residual movement of fleshpoints along the lip corners raiser and the nose wrinkler. Using our third corpus (natural expressive free speech), an additional analysis-by-synthesis process has been performed to identify additional components [3]. 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 3. The shape model that includes speech-related and expression-related facial movements has finally 11 DoF.
Note that these elementary facial movements are not directly related to particular visemes (contrary to a key frames approach). They emerge from general tendencies observed in data subsets containing several dozen facial configurations.

Figure 3: Showing the effect of each elementary expressive action on the face and lip shapes. The face is rendered using a unique texture, hence the non photorealistic appearance. From left to right, top to bottom: neutral, raising eyebrows, lip corners raiser (obtained from the smile 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.

4 Comparative analysis of articulation for neutral *vs* expressive speech

The statistical articulatory model is then coupled with a computer vision algorithm (pattern matching using normalized cross correlation) to build a robust beads tracker. The movements of the head and face in the 50 Hz corpus are tracked with this tool. Forced alignment of the audio signal of the VCV sequences is performed in order to label automatically the centers of realization of each allophone (silence, V, C, V) for a subset of 90 VCV sequences. We compare the values of the 11 articulatory parameters of these center frames between neutral and expressive production. The global results are displayed in Figure 4. The examination of this comparative study delivers quite interesting features:
Opening/closing movements of the jaw and lower lip are relatively unaffected by these two expressions
Jaw is significantly advanced and upper lip raised in expressive speech in comparison with neutral speech
Disgust is clearly characterized by the activation of the nose wrinkler and lowering of the larynx. Disgust recruits the lip corners raiser but with a smaller amplitude than smile.
Smile strongly decreases the rounding gesture. Production of rounded sounds while smiling are in return clearly decreasing the amplitude of activation of the lip corners raiser (see parameter *sourire* in Figure 5)

5 Mapping neutral to expressive speech

Several proposals have been made for adding facial expressions to talking faces [6, 14, 15]. Most consist in phoneme-dependent or exemplar-based gesture blending procedures. These techniques are very sensitive to data

sparsity and depend on the appropriate selection of the appropriate expressive visemes stored in the database.
We test here the predictive power of multilinear models trained on the small VCV data set used in the present experiment (4 frames for each of the 90 VCVs). Each parameter of an expressive frame is thus computed as a linear combination of the parameters of the associated frame uttered with neutral expression.

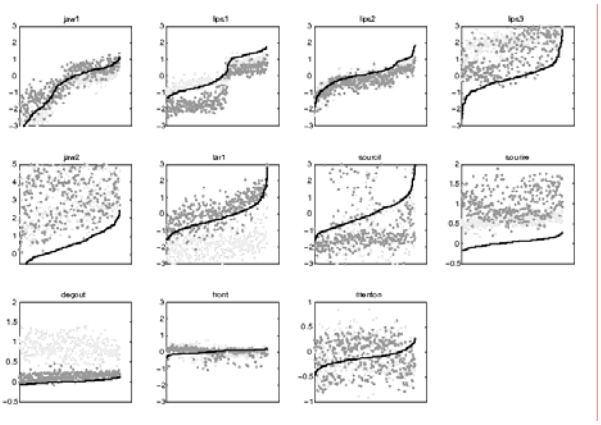

Figure 4: Comparing the 11 articulatory parameters for all target frames of allophones of the VCV sequences. For each parameter, the black line represents the observed values for the neutral phonemes, where phonemes were reordered by ascending parameter value. Light gray dots represent the corresponding observations for the "disgust" corpus, while dark gray dots correspond to the "smile" corpus. If disgust frames set approximately the 9th parameter (*degout* corresponding to nose wrinkler) to 1, both disgust and smile recruit the lip corners raiser (8th parameter *sourire*). Note also the unaffected jaw (1st parameter *jaw1*) and the lower larynx in disgust frames (6th parameter *lar1*).

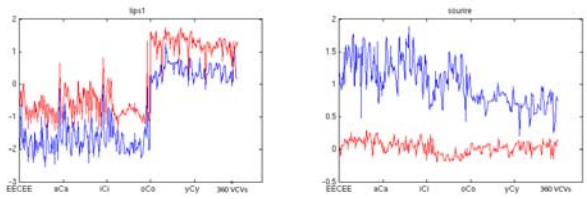

Figure 5: Comparing lip rounding and lip raising parameters for neutral articulations (in red online) *vs* smile articulations (in blue online) frames when considering rounded (rightmost frames) *vs* unrounded sounds (leftmost frames).

Table 1 shows the correlation coefficients between original expressive frames and expressive frames computed from neutral ones. If most parameters that shape speech sounds are satisfactorily predicted, parameters that are specifically recruited for displaying facial expression are still not accurately predicted.
Predictions of the multilinear model can however perform quite satisfactory when applied to complete sentences (see Figure 7). Though the model was learned with hyper-articulated frames, it seems to generalize well when applied in the more co-articulated context as found in sentences.

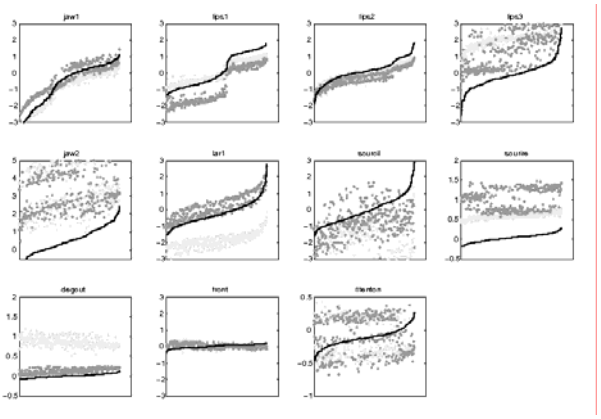

Figure 6: Same as Figure 4 but with parameters for expressive frames computed by a multilinear predictor from neutral ones.

Table 1: Correlation coefficients and mean errors between articulatory parameters of original or predicted frames. Expressive frames are predicted from neutral frames by multilinear regression. Coefficients coefficients above 0.8 and errors below 0.5 (standard deviation) are in bold.

	Smile *vs* neutral	Disgust *vs* neutral	Smile *vs* predicted smile	Disgust *vs* predicted disgust
jaw1	**0,88** - 0,6	**0,95 - 0,4**	**0,92 - 0,4**	**0,96 - 0,4**
lips1	**0,95** - 0,6	**0,94 - 0,4**	**0,97 - 0,4**	**0,96 - 0,2**
lips2	**0,86 - 0,4**	**0,85 - 0,4**	**0,91 - 0,3**	**0,91 - 0,4**
lips3	0,65 - 1,0	0,69 - 1,3	**0,90** - 0,7	**0,82** - 0,6
jaw2	0,26 - 1,2	0,24 - 0,8	**0,87** - 1,0	**0,90** - 0,9
lar1	0,77 - 0,6	0,54 - 2,6	**0,85** - 0,7	0,63 - 0,6
sourcil	0,20 - 0,9	0,30 - 5,3	0,58 - 0,7	0,77 - 1,8
sourire	0,16 - 0,6	0,49 - **0,3**	0,77 - **0,3**	0,59 - **0,1**
degout	0,49 - **0,1**	-0,21 - 0,8	0,75 - **0,1**	0,54 - **0,2**
front	-0,17 - **0,3**	-0,20 - **0,3**	0,53 - **0,1**	0,72 - **0,1**
menton	-0,08 - **0,2**	-0,11 - **0,4**	**0,89 - 0,2**	**0,82 - 0,2**

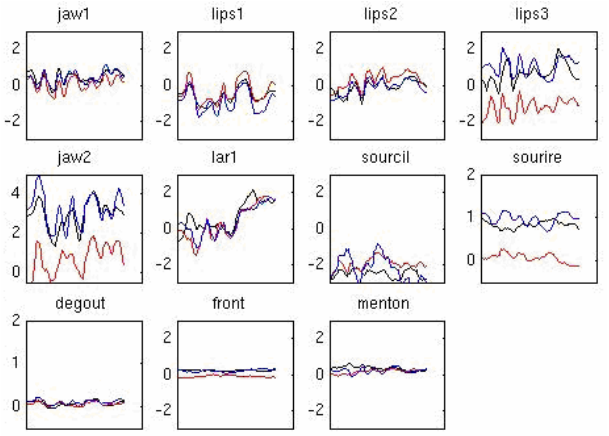

Figure 7: Analyzing with parameters for smile frames of a test sentence computed by a multilinear predictor from a neutral version (blue lines). Red lines show the parameter values for the neutral corpus, black lines show recorded values from the smile corpus. Neutral and target utterances are time-aligned by time-warping acoustics signals.

Conclusions

We have shown that articulatory parameters responsible for the control of facial expressions and speech interact. A first model has been proposed to take into account this complex interaction for one single speaker. A simple multilinear blending seems to predict with sufficient objective precision the jaw movements and lips shapes. Other articulatory parameters responsible for the shaping of expressive facial displays are still not predicted precisely. This asymmetry in terms of objective prediction should have an impact on the subjective identification of phonetic segments and facial expression. We are currently investigating this issue. The proper rendering of the computed movements by adequate textures should be solved before any subjective test may be conducted.

For instance, linear texture models such as promoted by Active Appearance models [20] are quite satisfactory but have a real problem with the mouth opening (see Figure 8).

We report here results obtained on one subject. Strategies are known to vary between subjects and between acted and spontaneous conditions [18]. Once shape and texture modeling have been satisfactorily solved, the full data recorded for our subjects will allow us to investigate both issues.

Figure 8. Modeling the texture by a multilinear regression of shape free images [12]. From left to right: mean texture then variation due to lip rounding, smiling and disgust. Note the appropriate variation of wrinkles and lightening. The final shape-free texture is obtained by blending these components according to the articulation. It is then morphed to the appropriate shape driven by the same articulatory parameters.

Acknowledgments

We thank Christophe Savariaux for the technical support. This work has been financed by the project *Presence* of the Rhone-Alpes Cluster ISLE and by the PPF “Multimodal Interaction” supported by Grenoble Universities.

References

[1] Aubergé, V. and M. Cathiard, *Can we hear the prosody of smile ?* Speech Communication, 2003. **40**: p. 87-97.

[2] Badin, P., et al., Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images. Journal of Phonetics, 2002. **30**(3): p. 533-553.

[3] Bailly, G., et al. Degrees of freedom of facial movements in face-to-face conversational speech. in International Workshop on Multimodal Corpora. 2006. Genoa - Italy. p. 33-36.

[4] de Gelder, B., J. Vroomen, and P. Bertelson. Cross-modal bias of voice tone on facial expression: Upper versus lower halves of a face. in Auditory Visual Speech Processing (AVSP). 1998. Terrigal, Australia. p. 93-96.

[5] de Gelder, B., J. Vroomen, and T. Popelier, *Facial expressions: Do part play a full role?* Journal of the International Neuropsychological Society, 1996. **2**(206).

[6] Deng, Z. and U. Neumann. eFASE: expressive facial animation synthesis and editing with phoneme-isomap controls. in Symposium on Computer Animation. 2006. Vienna - Austria. p. 251-260.

[7] Ekman, P. and H. Oster, *Facial expressions of emotion.* Annual Reviews of Psychology, 1979. **30**: p. 527-554.

[8] Elisei, F., et al. Capturing data and realistic 3D models for cued speech analysis and audiovisual synthesis. in Auditory-Visual Speech Processing Workshop. 2005. Vancouver, Canada.

[9] Kakadiaris, I.A., et al., Three-dimensional face recognition in the presence of facial expressions: an annotated deformable model approach. IEEE Trans. on PAMI, 2007. **29**(4): p. 640-649.

[10] Kappas, A., et al., Angle of regard: The effect of vertical viewing angle on the perception of facial expressions. Journal of Nonverbal Behavior, 1994. **18**(4): p. 263-280.

[11] Kätsyri, J., et al. Identification of synthetic and natural emotional facial expressions. in Auditory-visual Speech Processing. 2003. St Jorioz - France. p. 239-244.

[12] Odisio, M. and G. Bailly. Shape and appearance models of talking faces for model-based tracking. in Audio Visual Speech Processing. 2003. St Jorioz, France. p. 105-110.

[13] Pantic, M. and L.J.M. Rothkrantz, *Automatic analysis of facial expression: the state of the art.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000. **22**(12): p. 1424-1445.

[14] Pearce, A., et al. Speech and expression: A computer solution to face animation. in Graphics Interface. 1986. Calgary, Canada. p. 136-140.

[15] Pelachaud, C. and I. Poggi, *Subtleties of facial expressions in embodied agents.* Journal of visualization and computer animation, 2002. **13**: p. 301-312.

[16] Pighin, F., et al. Synthesizing realistic facial expressions from photographs. in Proceedings of Siggraph. 1998. Orlando, FL, USA. p. 75-84.

[17] Revéret, L., G. Bailly, and P. Badin. 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. 2000. Beijing, China. p. 755-758.

[18] Schmidt, K. and J. Cohn. Dynamics of facial expression: Normative characteristics and individual differences. in IEEE International Conference on Multimedia and Expo (ICME). 2001. Tokyo. p. 728 - 731.

[19] Tartter, V.C. and D. Braun, *Hearing smiles and frowns in normal and whisper registers.* Journal of the Accoustical Society of America, 1994. **96**: p. 2101-2107.

[20] Theobald, B.-J., et al. Evaluation of a talking head based on appearance models. in Auditory-visual Speech Processing Workshop. 2003. St Jorioz, France. p. 187-192.