Speaking with smile or disgust: data and models
Gérard Bailly, Antoine Begault, Frédéric Elisei, Pierre Badin

To cite this version:
Gérard Bailly, Antoine Begault, Frédéric Elisei, Pierre Badin. Speaking with smile or disgust: data and models. AVSP 2008 - 7th International Conference on Auditory-Visual Speech Processing, Sep 2008, Moreton Island, Australia. pp.111-116. hal-00333673

HAL Id: hal-00333673
https://hal.science/hal-00333673
Submitted on 15 Nov 2008

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.
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 nonsense 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. 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.

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
explore 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 characterizes the /i/ vs. /y/ opposition; lips2 controls the upper lip raising / lowering movement, useful to realise the labiodental 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 jaw3 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 movements 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), unfrowning 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.

<table>
<thead>
<tr>
<th></th>
<th>Smile vs neutral</th>
<th>Disgust vs neutral</th>
<th>Smile vs predicted smile</th>
<th>Disgust vs predicted disgust</th>
</tr>
</thead>
<tbody>
<tr>
<td>jaw1</td>
<td>0.88 - 0.6</td>
<td>0.95 - 0.4</td>
<td>0.92 - 0.4</td>
<td>0.96 - 0.4</td>
</tr>
<tr>
<td>lips1</td>
<td>0.95 - 0.6</td>
<td>0.94 - 0.4</td>
<td>0.97 - 0.4</td>
<td>0.96 - 0.2</td>
</tr>
<tr>
<td>lips2</td>
<td>0.86 - 0.4</td>
<td>0.85 - 0.4</td>
<td>0.91 - 0.3</td>
<td>0.91 - 0.4</td>
</tr>
<tr>
<td>lips3</td>
<td>0.65 - 1.0</td>
<td>0.69 - 1.3</td>
<td>0.90 - 0.7</td>
<td>0.82 - 0.6</td>
</tr>
<tr>
<td>jaw2</td>
<td>0.26 - 1.2</td>
<td>0.24 - 0.8</td>
<td>0.87 - 1.0</td>
<td>0.90 - 0.9</td>
</tr>
<tr>
<td>lar1</td>
<td>0.77 - 0.6</td>
<td>0.54 - 2.6</td>
<td>0.85 - 0.7</td>
<td>0.63 - 0.6</td>
</tr>
<tr>
<td>sourcil</td>
<td>0.20 - 0.9</td>
<td>0.30 - 5.3</td>
<td>0.58 - 0.7</td>
<td>0.77 - 1.8</td>
</tr>
<tr>
<td>sourire</td>
<td>0.16 - 0.6</td>
<td>0.49 - 0.3</td>
<td>0.77 - 0.3</td>
<td>0.59 - 0.1</td>
</tr>
<tr>
<td>degout</td>
<td>0.49 - 0.1</td>
<td>-0.21 - 0.8</td>
<td>0.75 - 0.1</td>
<td>0.54 - 0.2</td>
</tr>
<tr>
<td>front</td>
<td>-0.17 - 0.3</td>
<td>-0.20 - 0.3</td>
<td>0.53 - 0.1</td>
<td>0.72 - 0.1</td>
</tr>
<tr>
<td>menton</td>
<td>-0.08 - 0.2</td>
<td>-0.11 - 0.4</td>
<td>0.89 - 0.2</td>
<td>0.82 - 0.2</td>
</tr>
</tbody>
</table>

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

**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 a 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 expression. We are currently investigating this issue. The parameters responsible for the shaping of expressive facial blending seems to predict with sufficient objective precision interaction for one single speaker. A simple multilinear

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