



HAL
open science

Automatic tool to annotate smile intensities in conversational face-to-face interactions

Stéphane Rauzy, Mary Amoyal

► **To cite this version:**

Stéphane Rauzy, Mary Amoyal. Automatic tool to annotate smile intensities in conversational face-to-face interactions. *Gesture*, 2023, Volume 21 (Issue 2-3), pp.320-364. 10.1075/gest.22012.rau . hal-04194987

HAL Id: hal-04194987

<https://hal.science/hal-04194987>

Submitted on 4 Sep 2023

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.

Automatic tool to annotate smile intensities in conversational face-to-face interactions

Stéphane Rauzy

Mary Amoyal

Aix Marseille Université, CNRS, Laboratoire Parole et Langage UMR 7309
5 avenue Pasteur
BP 80975, 13604 Aix-en-Provence, FRANCE
Tel.: +334-13-55-36-88
Fax: +334-13-55-37-44
E-mail: stephane.rauzy@univ-amu.fr, mary.amoyal@univ-amu.fr

Abstract

This study presents an automatic tool that allows to trace smile intensities along a video record of conversational face-to-face interactions. The processed output proposes a sequence of adjusted time intervals labeled following the *Smiling Intensity Scale* (Gironzetti, Attardo, and Pickering, 2016), a 5 levels scale varying from neutral facial expression to laughing smile. The underlying statistical model of this tool is trained on a manually annotated corpus of conversations featuring spontaneous facial expressions. This model will be detailed in this study. This tool can be used with benefits for annotating smile in interactions. The results are twofold. First, the evaluation reveals an observed agreement of 68% between manual and automatic annotations. Second, manually correcting the labels and interval boundaries of the automatic outputs reduces by a factor 10 the annotation time as compared with the time spent for manually annotating smile intensities without pretreatment. Our annotation engine makes use of the state-of-the-art toolbox OpenFace for tracking the face and for measuring the intensities of the facial Action Units of interest all along the video. The documentation and the scripts of our tool, the SMAD software, are available to download at the HMAD open source project url page <https://github.com/srauzy/HMAD>.

Keywords: Smiling Intensity Scale, Automatic annotation, OpenFace software, Machine Learning, Facial Action Coding System

1 Introduction

Social interactions and specifically conversations involve inherently gestures and speech as an “integrated whole” (McNeill, 1992). Many studies have demonstrated the link between gestures (e.g. hand gestures, head gestures or facial gestures) and the discourse organization (see among others (Alibali, Kita, & Young, 2000; Kendon, 2004; McNeill, 2012)). It has also been shown that the gaze direction and the head trajectory are involved in the conversational process (see for example (Hanna & Brennan, 2007; Holler et al., 2014; Kendon, 1967)). Among facial expressions, smiling is “very frequent” (Kerbrat-Orecchioni & Cosnier, 1987) in conversational interaction and it has been shown that “the face is the most often observed part of the body when we are talking” (Guy, 2013). If smile has mainly been studied to reflect emotions such as joy (Bateson, Winkin, Bansard, Cardoen, & Birdwhistell, 1981; Paul Ekman, 1984), it also can be considered as a “facial gesture” (Bavelas & Gerwing, 2007) that conveys pragmatic and interactive functions (Argyle, 1975). Whether smile is conversational or emotional, this facial expression is involved in the “collaborative process” (Sacks, Schegloff, & Jefferson, 1978) required to communicate. In order to explore the smile role in conversations, we first need to identify and quantify smiles. There is a real need in a reliable and systematic annotation tool that take into account the speech consequences on facial expressions such as the noisy environment due to the activation of facial muscles. Not only presence or absence of smile need to be explored, neither low or high smiles, but smile intensities. In this present study, we focus on smiling annotation using a sign-based approach which describes facial changes based on physiological features. In that aim, our approach relies on the Smiling Intensity Scale (SIS) (Gironzetti et al., 2016), an annotation scale based on the Action Units from the FACS (P. Ekman, Friesen, & Hager, 2002). The solicitation of different muscles and the differences observed between a low or a high smile (El Haddad, Chakravarthula, & Kennedy, 2019) provides evidence that it is important to take into account the different smile intensities. The smiling intensity scale describes smiling into 5 intensities (from neutral face to laughing smile) and provides a guideline for a manual annotation. However, the manual annotation of gesture (hand or facial gesture) is a time-consuming task. In practice, it limits the size of manually annotated corpus available. Moreover, a large amount of data is needed in order to deeply analyze interactions. That is why in our multimodal approach, any alternative solution to the manual annotation is welcome. As far as we know, there is no tool which allows to automatically annotate smile intensities according to the SIS scale. This exploratory study aims to provide such a tool.

Recent advances in the field of computer vision and machine learning have given birth to a generation of softwares enable to detect and track a face along a video record and eventually to measure its internal facial movements. Automatic analysis of facial expressions implies in fact several steps of treatment (see for example (Martinez, Valstar, Jiang, & Pantic, 2019) for a recent survey): a pre-processing of video images in order to detect and track the face and its characteristic facial landmarks all along the video capture, the extraction of features describing for example atomic facial muscle actions (i.e. the Action Units (AUs) of the Facial Action Coding System (FACS) (P. Ekman et al., 2002; Paul Ekman & Friesen, 1978)) and a final step allowing to automatically detect facial actions based on the measured features. Various solutions based on different techniques and algorithms have

already led to a bunch of distributed softwares (see table 1 of (Baltrušaitis, Zadeh, Lim, & Morency, 2018) for an overview and a comparison of the respective characteristics of the available tools). The today challenge concerns how facial behavior analysis softwares do perform on in-the-wild¹ videos recording spontaneous facial expressions² (see for example (Cohn & De la Torre, 2014; Dhall, Goecke, Gedeon, & Sebe, 2016; Martinez et al., 2019)). In a prior study (Rauzy & Goujon, 2018) we investigated the detection of eyebrows raising and eyebrows frowning (Goujon, Bertrand, & Tellier, 2015) on spontaneous and in-the-wild video materials. The videos were first treated by the OpenFace toolbox (Baltrušaitis, Mahmoud, & Robinson, 2015; Baltrušaitis, Robinson, & Morency, 2016; Baltrušaitis et al., 2018) in order to track the face and to obtain the facial landmark trajectories before applying our own processing. The results were encouraging but without bringing a clear-cut answer regarding the benefits of using this tool as a help for annotation purpose. The present study approach is similar but concerns the automatic annotation of smile intensities along time on in-the-wild videos.

Automatic smile detection has been already addressed considering different issues and exploring various dimensions (see for example (Whitehill, Littlewort, Fasel, Bartlett, & Movellan, 2009)). The proposed solutions vary indeed whether one wants to detect the presence or the absence of smile (An, Yang, & Bhanu, 2015; Chen, Ou, Chi, & Fu, 2017; Guo, Polania, & Barner, 2018; Shan, 2012; Zhang, Huang, Wu, & Wang, 2015) or rather one wants to estimate smile intensity (Bartlett, Littlewort, Braathen, Sejnowski, & Movellan, 2003; Bartlett et al., 2006; Girard, Cohn, & De la Torre, 2015; Jiang, Coskun, Badokhon, Liu, & Huang, 2019; Shimada, Matsukawa, Noguchi, & Kurita, 2010; Vinola & Vimala Devi, 2019). The methods applied also change if one is interested in classifying single face image (An et al., 2015; Chen et al., 2017; Guo et al., 2018; Jiang et al., 2019; Shan, 2012; Shimada et al., 2010; Zhang et al., 2015) rather than proposing a dynamical annotation of a video recording (Freire-Obregón & Castrillón-Santana, 2015). The main difficulty plaguing in practice the automatic smile intensity estimation task lies however on the lack of a large dataset with manually annotated references (Girard et al., 2015; Guo et al., 2018; Walecki, Rudovic, Pavlovic, & Pantic, 2019). Despite the variety of existing tools, none correspond to the speech-noise conversational data we want to analyze. This is why this study presents a new ad hoc tool that automatically detects smiles in conversation.

This paper is organized as follows. In section 2 we present the OpenFace software which will be pipelined in order to feed the input of our analysis. Section 3 describes our gold standard corpus, a collection of 4 videos in which smile intensities have been manually annotated and coded by two judges following the SIS (Gironzetti et al., 2016). The building up of the model is detailed section 4. We establish the correspondence between the manual annotations and the intensities of the facial Action Units measured by OpenFace. The dynamical probabilistic model is settled in a second step and we afterwards explain how to estimate the parameters of the model from the gold standard data. The output of the

¹The term 'in-the-wild' is used by the computer vision community to describe any realistic settings where the captured face may be far from the frontal head pose, may undergo abrupt head motions, may be masked due to partial occlusions and may be subject to varying illumination conditions.

²The term 'spontaneous facial expressions' stands for the natural facial expressions anybody experiments during everyday-life social interaction, in contrast with facial expressions resulting from posed emotion played by an actor for example.

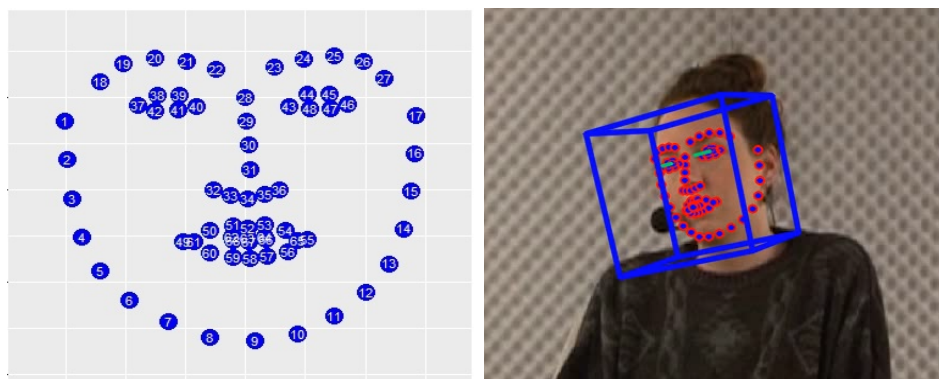


Figure 1. Left panel: The position of the 68 facial landmarks used by the OpenFace software. Right panel: A frame capture of a processed video showing the landmarks position and head pose (the projected blue cube edges).

smile detection engine is finally described in a last step. The performance of the tool is investigated section 5. A first evaluation is performed using the standard metrics (precision, recall, f-measure and Cohen’s κ coefficient). A second evaluation compares the annotation time required for manually correcting the labels and interval boundaries of the automatic outputs versus the time spent for annotating smile intensities without pretreatment. Section 6 contains practical information on how to download the open source scripts of our SMAD tool. The last section is devoted to discussion and concluding remarks.

2 The OpenFace software

The OpenFace toolkit (Baltrušaitis et al., 2015; Baltrušaitis et al., 2016; Baltrušaitis et al., 2018) is an open source project which proposes the full capabilities of facial behavior analysis tool: head tracking, facial landmark detection, head pose estimation, facial action unit recognition and eye-gaze estimation. OpenFace implements a Constrained Local Neural Field algorithm (CLNF, (Baltrušaitis, Robinson, & Morency, 2013a, 2013b)) in order to achieve facial features tracking. Illustration of the OpenFace processed video is shown figure 1. The positions of 68 facial landmarks are provided along time for each video frame as well as global parameters such as the 3-dimensional head pose and the yaw, pitch and roll angles specifying the direction of the head. Reconstruction of the facial landmark movements corrected from head rotation and global head translation is obtained by fitting a head model to the OpenFace output data. A complete description of the OpenFace output and the download instructions can be found at the Tadas Baltrušaitis github url address:

<https://github.com/TadasBaltrusaitis/OpenFace>

2.1 Action Unit detection

The description of atomic facial muscle activities that combines to achieve different facial expressions, i.e. the facial Action Units (AU), are encoded thanks to the Facial Action Coding System (FACS) (P. Ekman et al., 2002; P. Ekman & Friesen, 1975; Paul Ekman &

Table 1

The 7 facial Action Units proposed by the OpenFace output which are related to smile activities (according to the SIS).

Name	Description	Facial muscle involved
AU06	Cheek Raiser	<i>Orbicularis oculi, pars orbitalis</i>
AU07	Lid Tightener	<i>Orbicularis oculi, pars palpebralis</i>
AU10	Upper Lip Raiser	<i>Levator labii superioris</i>
AU12	Lip Corner Puller	<i>Zygomaticus major</i>
AU14	Dimpler	<i>Buccinator</i>
AU25	Lips part	<i>Depressor labii inferioris or relaxation of Mentalis, or Orbicularis oris</i>
AU26	Jaw Drop	<i>Masseter, relaxed Temporalis and internal Pterygoid</i>



Figure 2. Examples of configuration with missing measurements. Two left panels: The face pose is too far from the frontal pose. Two right panels: Partial face occlusion. Pictures extracted from the “Aix-DVD” Corpus (Gorisch & Prévot, 2014).

Friesen, 1978). The approach used by OpenFace to perform facial Action Unit detection and intensity estimation is described in (Baltrušaitis et al., 2015). When dealing with a video record the algorithm performs dynamically a face normalization which improves the performance of AUs prediction. The follow-up of 17 AUs is proposed by OpenFace and 7 of them (AU06, AU07, AU10, AU12, AU14, AU25 and AU26, see the description given table 1) are directly involved during smile activities. Two formats are available in the OpenFace output: The intensity of the AU for each frame on a scale varying continuously from 0 to 5 and a binary variable encoding the presence or the absence of the Action Unit.

2.2 In-the-wild videos

The main challenge for facial behavior analysis softwares is to cope in practice with realistic settings and with spontaneous facial expressions. Some examples of these configurations are presented figure 2. The two left panels show cases where the head pose is too far from the frontal pose, implying the track loss of the head for OpenFace, and thus missing measurements at the landmark positions level. The two right panels of figure 2 show cases of partial occlusions (occlusion due to the second participant in the left panel, self-occlusion for the right panel). Here again, the head track is temporarily lost by the software.

A second and maybe more problematic type of errors arises when the software affects a wrong position to the head without indicating a low confidence in the landmarks detection

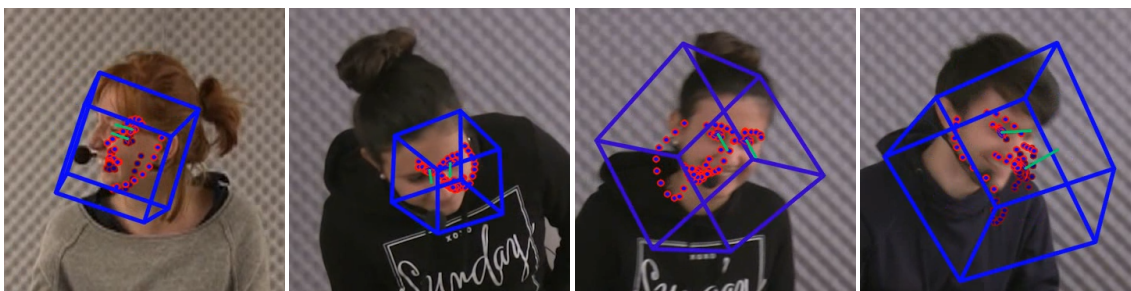


Figure 3. Examples of OpenFace output with problematic measurements. The confidence level of landmarks detection associated with these frames is not given low by the OpenFace software whereas the face tracker clearly fails at detecting the head position.

estimate. This is illustrated figure 3 for the OpenFace software. Some heuristics have to be developed in order to detect and discard these kind of spurious measurements. An attempt to deal with this weakness will be presented in section 4.5. After having presented the relevant aspects of the OpenFace software, we will now present the constitutive elements of the manual annotation of smiling.

3 Manual annotations of smiles






3.1 The Smiling Intensity Scale (SIS) of Gironzetti et al. (2006)

A recent study (Jensen, 2015) has explored smiles according to the different physiological movements involved in this facial expression. It has been shown that we use at least two muscles (on each side of the face) to produce a smile: the zygomatic and orbicular muscles of the eye. In the literature we can thus distinguish two types of smiles: “authentic” smiles that would be produced with the intervention of these two muscles and “social” smiles where only the zygomatic would be involved (Paul Ekman, Davidson, & Friesen, 1990). The “authentic” smile is also called the “Duchenne smile” named after the French anatomist G.B. Duchenne who worked on this gesture. This terminology has been discussed and tested, particularly on the issue of “simulation” of this authentic smile (Krumhuber, Likowski, & Weyers, 2014). Nevertheless the binary description of smiling in term of “authentic/simulated” or “presence/absence” appears too narrow in order to study the development of this complex facial expression. To overcome this principal limitation, a five levels smiling scale based on the measurements of Action Units detailed by the FACS (Paul Ekman & Friesen, 1978) was proposed in (Harker & Keltner, 2001; Seder & Oishi, 2012). This intensity scale has been improved by Gironzetti et al. (2016) in their *Smiling Intensity Scale* (SIS) which combines the intensities of several Action Units. The SIS measures the intensity of smiles gradually from 0 (neutral facial expression) to 4 (laughing smile). The main characteristics of each level of smiling intensity is detailed table 2 in term of associated facial Action Units. Note that the Duchenne’s dichotomy (spontaneous versus genuine smile) is not addressed by the SIS scale.

The internal consistency of the SIS system was tested by 3 raters that obtained a good inter-rater reliability (Cohen’s $\kappa = 0.89$, see (Gironzetti et al., 2016)). Thus at this stage, this scale allows to factually categorize smile intensities and not to attribute functions to

Table 2

Description of the Smiling Intensity Scale (SIS) of Gironzetti et al. (2016) for annotating smile activity (see appendix A of (Gironzetti, Attardo, & Pickering, 2016)). The illustrations of each level are pictures extracted from the CHEESE! corpus.

<p>Neutral facial expression (S0): No smile, no flexing of the zygomaticus (no AU12), may show dimpling (AU14), but no raised side of the mouth, the mouth may be closed or open (AU25 or AU26).</p> <p>AU concerned : 12,14,25,26</p>	
<p>Closed mouth smile (S1): Flexing of the zygomaticus (AU12), may show dimpling (AU14), may show flexing of the orbicularis oculi (caused by AU6 or AU7).</p> <p>AU concerned : 12 (6,7,14)</p>	
<p>Open mouth smile (S2): Showing upper teeth (AU25), flexing of the zygomaticus (AU12), may show dimpling (AU14), may show flexing of the orbicularis oculi (caused by AU6 or AU7).</p> <p>AU concerned : 25,12 (14,6,7)</p>	
<p>Wide open mouth smile (S3): Showing lower and upper teeth (AU25) or a gap between upper and lower teeth (AU25, AU26), flexing of the zygomaticus (AU12), may show dimpling (AU14) and flexing of the orbicularis oculi (AU6, AU7).</p> <p>AU concerned : 12,6,7,25,26 (14)</p>	
<p>Laughing smile (S4): Jaw dropped (AU25 and AU26 or AU27), showing lower and upper teeth, flexing zygomaticus (AU12), flexing of the orbicularis oculi (AU6 or AU7), dimpling (AU14).</p> <p>AU concerned : 25,26,27,12,6,7,14</p>	

smile. It is worthwhile to mention that smile is a gradual expression. However for practical reasons, the manual annotation of this continuum requires to adopt a discrete scale with several levels.

3.2 Adaptation of the scale

Several theoretical issues appeared when choosing and using the SIS (Gironzetti et al., 2016). On the one hand, in the SIS, smile 3 is originally declined into 2 subsmiles: (3A) showing upper and lower teeth and (3B) showing a space between upper and lower teeth. We did not retain these two classes because criteria for this subcategorization did not seem distinctive enough. On the other hand, this scale also originally dissociates the neutral face (0) and the laughing smile (4). Indeed, in the SIS, the category 4 is a "laughing smile": a smile that immediately precedes a laugh. However defining criteria for distinguishing intensity 4 from a laugh is not straightforward. Therefore, we extended this category to all laughter produced by the participants. In our adaptation of the SIS, we categorized all perceived laughter (vocal or not) as S4. Several arguments motivated this choice. First, in our data we found periods in which laughter was produced without vocalization. According to the original SIS, these periods would not have been annotated as a smile 4. For a laughter including a time laps where the speaker does not produce any sound, it does not seem relevant to annotate the sequence as laughter - laughing smile - laughter. Several studies on the status of laughter in interaction have highlighted that laughter can also have different intensities, as described in the study by El Haddad et al. (2019) (i.e. low, medium and high laughter). Moreover, a laugh can be realized by other gestures or postures (e.g. shoulder movement, orbicular wrinkling, ...) than its canonical manifestations (AUs generally associated with laughter, vocalization).

Laughter and smile do not have the same status, nor the same characteristics and by extension do not have the same interactive functions. In term of frequency, we know that a laughter occurs on average every 2 minutes (Vettin & Todt, 2004) whereas a smile (3 intensities combined) occurs every 40 seconds on average. In term of location, an interval annotated "laughing smile" in the SIS Gironzetti et al. (2016) follows a specific constraint: it has necessarily to precede a laugh which is not the case for the other scale levels (from S0 to S3). The modification that we propose does not place laughter as the highest level of the smiling scale, we distinguish this category as we distinguish the neutral face. We thus obtain a slightly different scale from the SIS (Gironzetti et al., 2016) :

- Neutral face (S0)
- Smiles (S1, S2, S3)
- Laughter (S4)

3.3 The CHEESE! corpus

The CHEESE! corpus is composed of 11 dyadic face-to-face conversation featuring 22 native french students (Priego-Valverde, Bigi, Attardo, Pickering, & Gironzetti, 2018). Each interaction is approximately 15 minutes long, then the corpus lasts on average 3 hours. The aim of the CHEESE! project was to realize a comparative study between French and

American speakers through their smiling reaction at humorous frame (Priego-Valverde et al., 2018). Participants were selected with the criteria that they have a friendship beside their university course. Participants were students from 20 to 30 years old and they did not know the purpose of the initial study nor they did receive any compensation. Participants were seated face-to-face in a soundproof room. Two cameras were positioned behind their back and pointed at the other participant’s face. The scene configuration for two pairs of participants is illustrated figure 4. Both participants were fitted with a micro headset, optimally positioned so as not to hide the mouth while preserving the acoustic signal. Each participant was asked to read a text (a canned joke). After the reading part, participants had 15 minutes to discuss as freely as they wished. The CHEESE! corpus is available on the ORTOLANG website (Open Resources and TOols for LANGuage) at the url <https://www.ortolang.fr/market/corpora/cheese>.

The CHEESE! corpus was manually annotated following the SIS methodology in a previous study (Priego-Valverde et al., 2018). This manual annotation was performed using Elan Software (Brugman, Russel, & Nijmegen, 2004), the manual annotation tool commonly used in Gestures Studies. Each video record was manually sequenced in a series of adjusted time intervals annotated with smile intensity values from 0 to 4. Each time boundary was positioned in a perceptive way without pre-established location. This methodology presents however two minor weaknesses. Smile is a complex facial gesture composed of several physiological features. Defining the smile boundaries is then a difficult task. As a result, the smile annotation produced by two different judges leads to different sets of smiling time intervals. Thus, the standard inter-annotator agreement methods (e.g. Cohen’s κ) are not straightforwardly applicable to this case.

In order to overcome these weaknesses, the time line was divided into a series of 400 milliseconds adjacent time intervals (Amoyal & Priego-Valverde, 2019) (i.e. with no overlap). This ad-hoc time step has been chosen considering that 200 ms is the minimal duration to perceive a complex facial expression such as smile (Heerey & Crossley, 2013; Sanders, 2013). Our time step of 400 ms corresponds to 10 video frames for a video rate of 25 frames per second. Based on this new annotation protocol, a double blind smile annotation has been performed on 2 interactions from the CHEESE! corpus (MA-PC and JS-CL, approximately 17 minutes each). One annotation was performed by an expert of the SIS scale and the second by a naive judge. The mean inter-annotator agreement measured by the Cohen’s κ score reaches 0.88 which confirms that the smile levels and the annotation protocol are reliable. This methodology is therefore appropriate to study smiling during conversation but remains time-consuming since it requires 1 hour to manually annotate 1 minute of video per participant. For example, a whole conversation (i.e. 2 participants during 15 minutes) is manually annotated in 30 hours on average. Thus, the reduction of this manual annotation cost is the main objective of the present study.

3.4 The gold standard

The gold standard is composed of the two conversations mentioned above (i.e. MA-PC and JS-CL). We retain as the final values of the smile annotations the expert ones. Table 3 summarizes the characteristics for each video. The all combined 4 extracts represent about 1 hour of video capture. From now on, we will proceed to two modifications. First, the 5 levels of the SIS system will be labeled as S_0 , S_1 , S_2 , S_3 and S_4 . Second, the adjacent



Figure 4. The scene configuration for the 4 corpus extracts of the gold standard.

Table 3

Characteristics for each video: duration of the record in seconds and number of manually annotated smiles on the S0 to S4 scale (SIS convention). The last row and last column shows respectively the altogether number of intervals in each class of smiles and in each corpus.

Participant	Duration	#S0	#S1	#S2	#S3	#S4	ALL
JSCL_CL	990	67	113	126	74	64	444
JSCL_JS	990	72	139	99	70	69	449
MAPC_MA	1044	35	55	30	24	12	156
MAPC_PC	1044	47	60	61	37	14	219
ALL	4068	221	367	316	205	159	1268

time intervals of 400 ms will be merged if they are labeled with the same intensity (e.g. 3 consecutive 400 ms intervals labeled S2 will form a unique S2 interval of 1200 ms). This operation will lead to a series of consecutive intervals with different labels and with variable durations. Using this convention, the gold standard contains 1268 time intervals which are labeled with smile intensities from S0 to S4. The distribution in duration for each class of smile is given table 4 and is illustrated figure 5. During half of the time the participants exhibit a non-smiling facial expression encoded as S0 and the remaining time is distributed between 20% of closed mouth smiles (i.e. S1), 12% of open mouth smiles (i.e. S2), 7% of wide open mouth smiles (i.e. S3) and 10% of laughter (i.e. S4 in the adapted scale). This result shows that participants smile 39% of the time recorded which is consistent with (Kerbrat-Orecchioni & Cosnier, 1987) who shows that smile is a very frequent facial expression in conversations. However these proportions (which indicates the mean distribution) are affected by a large variability depending on the specificity of the interaction (e.g. topic discussed, interactional frame) and of the personal profile characterizing each subject. Figure 6 shows for example that JS-CL interaction contains more than twice smiling active intervals than MA-PC interaction. We also note that the proportion of the smile labels does not follow a general trend but is rather specific of each subject. The proportion of S1 (closed mouth smile) and S2 (open mouth smile) is for example similar for participants CL and PC whereas the S1 class is significantly more represented than the S2 class for speakers JS and MA.

The histogram of the smile duration is illustrated figure 7 for each smiling class

Table 4

Characteristics for each video: Total duration in seconds of the manually annotated smiles intervals (SIS convention) per participant and per smiling class.

Participant	S0	S1	S2	S3	S4	ALL
JSCL_CL	523.2	124.4	138.4	61.6	142.4	990
JSCL_JS	322.8	328.8	100.0	70.4	158.0	990
MAPC_MA	667.2	194.4	80.0	69.2	33.2	1044
MAPC_PC	578.4	144.0	165.6	92.8	63.2	1044
ALL	2091.6	801.6	484.0	294.0	396.8	4068

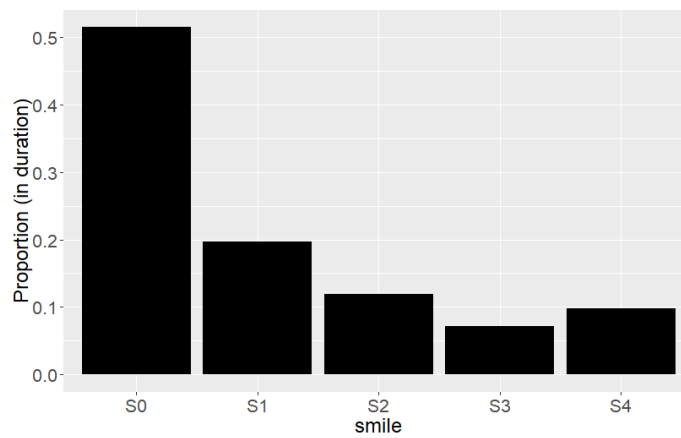


Figure 5. The relative proportion (in duration or in number of video frames) for the 5 classes of smile on the manually annotated corpus.

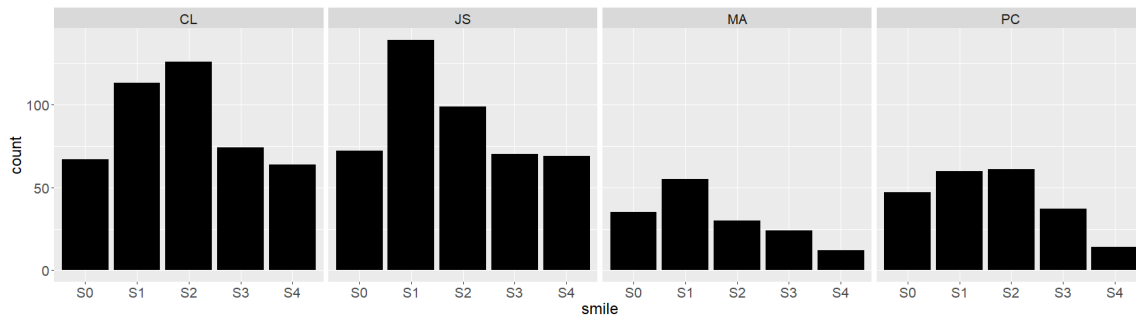


Figure 6. Distribution of the 5 smiling classes for the 4 participants of the manually annotated gold standard.

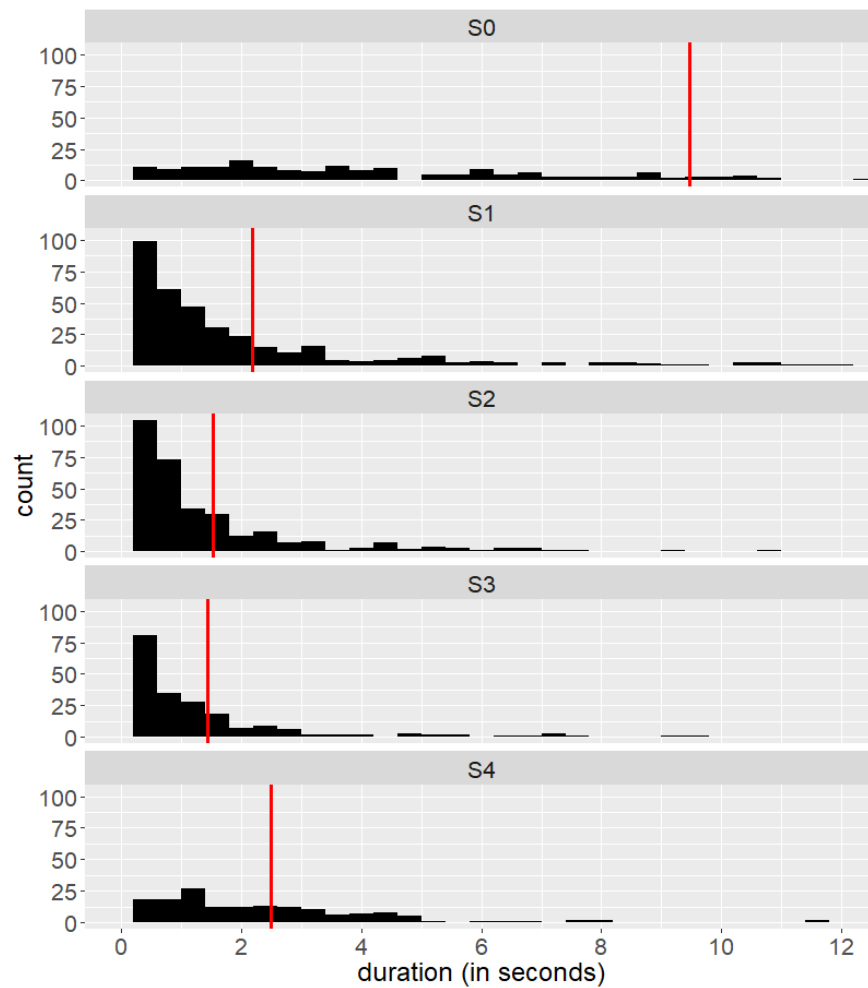


Figure 7. Histogram of the smile durations in seconds for the 5 classes of smiles over the gold standard (all participants combined). The vertical red line shows the mean duration for each class.

(all participants combined). The value of the mean duration is outlined by a vertical red line. The distributions in duration for the three intermediate classes S1, S2 and S3 (i.e. respectively closed, open and wide open mouth smiles) present a gamma like distribution function: a maximum of smile occurrences for short duration which decreases exponentially as the smile duration increases. The mean duration for the three intermediate classes S1, S2 and S3 belong to the range 1 to 2 seconds. The S0 class shows a nearly uniform distribution with an average duration around 10 seconds. As a matter of fact, this difference is not surprising since S0 encodes time interval of neutral facial expression. Similarly the observed duration distribution of laughter suggests that S4 is different by nature compared to the three other smiling classes S1, S2 and S3 (flat distribution and mean duration of 2.5 seconds).

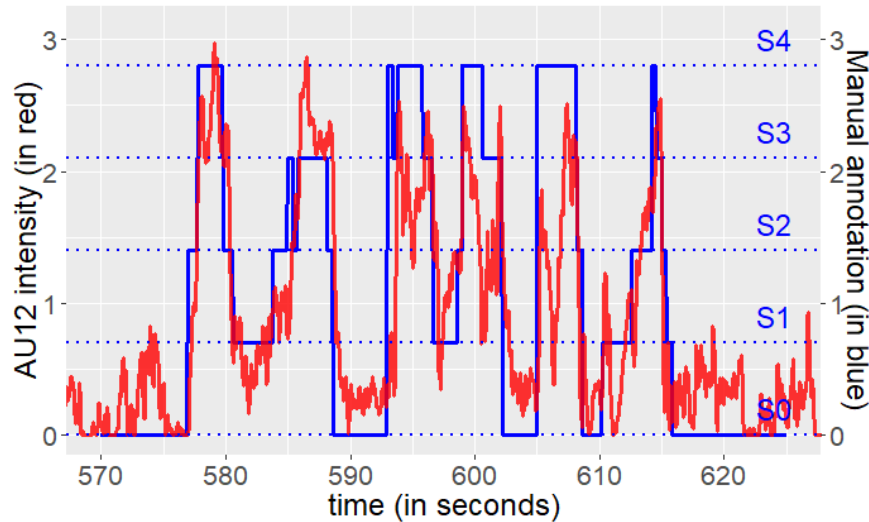


Figure 8. An example of the temporal correlation of the AU intensity (here the AU12 intensity measured by the OpenFace software, the red curve) with the manually annotated smiles on the SIS system (S0 for no smile and S1 to S4 for a gradual scale intensity, blue curve).

4 Automatic detection of smile activity

4.1 Link between the OpenFace AUs intensities and the SIS manual annotations

The OpenFace software proposes the follow-up of a subset of 17 facial AUs reported for each frame as an intensity measurement on a continuous scale spanning the range 0 (absence of the AU) to 5 (presence of the AU at a maximum intensity). Among this subset 7 facial action units (see table 1) are directly involved during smile activities. Some AUs present in the SIS guideline are not measured by OpenFace (e.g. AU27) and conversely some OpenFace AU measurements are not directly mentioned in the SIS guideline (e.g. AU10). A full match between the two systems is however not crucial since the AU intensity measurements are found to be highly dependent on each other.

The temporal correlation between OpenFace AU intensities and the manual annotations obtained by following the SIS guidelines is illustrated figure 8. The action unit AU12 is associated with the contraction of the zygomaticus major muscle (lip corner puller) and is emblematic of smile activity. Figure 8 demonstrates that the measured AU12 intensities trace remarkably well the manually annotated data and delimit accurately the areas of smile activity. There is thus no doubt that the automatic detection of smile activity from the OpenFace output is feasible. However, two issues need to be addressed. The first one concerns the way to combine the different AUs together in order to optimize the smile detection. The second one deals with the best strategy to adopt in order to model and predict the smile intervals.

The first issue can be solved by investigating the different contributions of the AUs of interest to each level of the smile scale. Figure 9 details for each AU the time variation

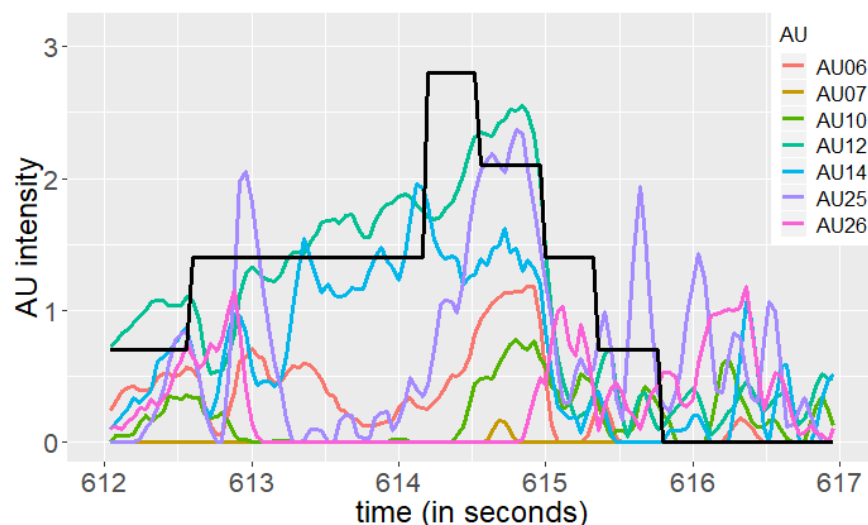


Figure 9. An example of the temporal correlation between the intensity of the 7 facial AU potentially activated during smiles. The SIS manually annotated smiles (5 levels from S_0 to S_4) is also drawn as the black curve.

of their intensities on a 5 seconds time interval. The SIS manually annotated smiles are also drawn. It appears that the intensities of each AU show different trends (for example for the time location of their maxima) and that the link between the AU intensities and the SIS annotations is more subtle than a single correlation. Indeed, the description of the SIS system (see table 2) suggests that some AU components will only be present at some level of the scale and will be absent otherwise. When combining the AUs intensities together, the model has to consider this information in order to predict the appropriate SIS level corresponding to the recorded smile activity.

Figure 10 shows the boxplot distributions for each AU and each class of smile. The median of the intensities distribution is drawn as an inside band mark and the lower and upper hinges of the boxplot stand for the first and third quartile of the distribution. The task of distinguishing between two classes will become much easier if the two medians show significant difference (i.e. the two boxplots do not overlap). Figure 10 shows for example that the AU12 intensity is a good candidate for discriminating between the S_0 class representing no smile areas (the red boxplot) and the remaining classes (S_1 , S_2 , S_3 and S_4) for which the median intensities are much higher. By contrast, the intensity of the AU26 measurement does not present any significant difference between the 5 classes of smile. The AU26 variable will therefore not be included in the model.

4.2 Specification of the model

There are obviously various ways to model the process and the solution we propose herein is one among others allowing to tackle efficiently the problem. We would like to outline that up to now there is no tool allowing to automatically annotate a video record following the SIS scale. Our goal is thus herein to propose one method allowing to achieve this task, the question of whether any other machine learning algorithms (Support Vector

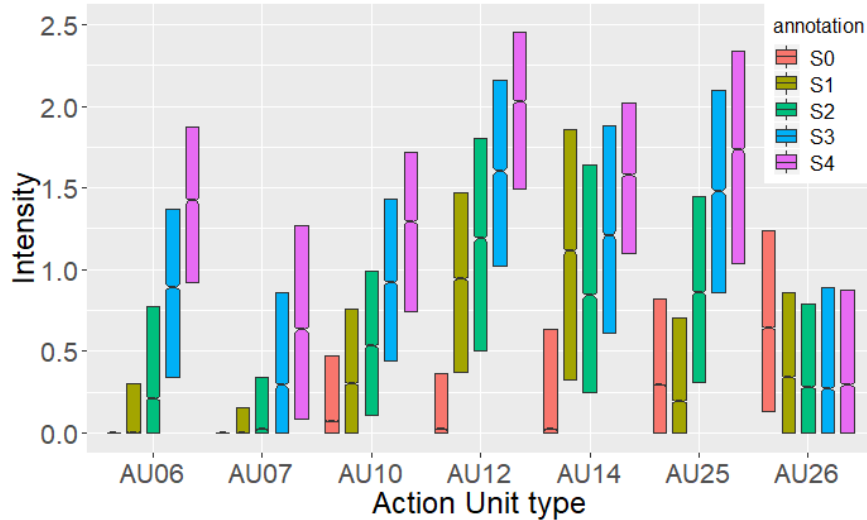


Figure 10. The boxplot distributions of the intensity of each AU for each class of smiles (S0 corresponds to no smile and the smiles scale runs for S1 to S4). The inside band mark shows the median of the intensities distribution. Lower and upper hinges of the boxplot correspond respectively to the first and third quartile of the distribution (i.e. the 25th and 75th percentiles).

Machine, Random Forest, Neural Network, ...) would perform better is out of the scope of the present study. The method we propose is driven by the two following remarks. Firstly we know that the SIS scale combines differently the Action Unit intensities at each of the 5 levels of the smiling scale. The model will then be built up step by step in order to mimic such a mechanism. Secondly, our manually annotated corpus, if regarded as training data, is modest in size. The corpus contains indeed around 1250 labeled time intervals from S0 to S4 that need to be explained by the measurements of the AU intensities at each video frame. In practice, the number of parameters entering the model has to be drastically controlled in order to avoid potential overfitting problems.

Herein, we make the choice to decompose the overall task in 4 iterative steps that are illustrated table 5. The first step consists in splitting dichotomously the time line between no smile areas S0 and smiling areas labeled S1 S2 S3 S4 (i.e. the compound class grouping the S1, S2, S3 and S4 labels together). Intervals which are predicted S1 S2 S3 S4 will be sliced during step 2 in time intervals labeled whether S1 or whether the compound class S2 S3 S4. Step 3 creates S2 and S3 S4 intervals from the parent S2 S3 S4 intervals and step 4 terminates the job by separating S3 from S4 time intervals.

At each step of the process the task is similar. An interval of time (herein a sequence of video frames) has to be splitted in smaller time intervals receiving respectively whether label 1 or label 2 accounting for the specific AU measurements along the time line. The dynamic of such a process can be described by a 2 states automaton (see figure 11) which evolves frame after frame driven by the AU intensity measurements and in function of the history of the system (i.e. the sequence of states previously occupied by the automaton). The evolution of the system is considered as a stochastic process and is characterized by

Table 5

The schematic structure of the 4 steps entering the smiles automatic annotation engine.

	S0	S1	S2	S3	S4
Step 1	S0	S1 S2 S3 S4			
Step 2	S0	S1	S2 S3 S4		
Step 3	S0	S1	S2	S3 S4	
Step 4	S0	S1	S2	S3	S4

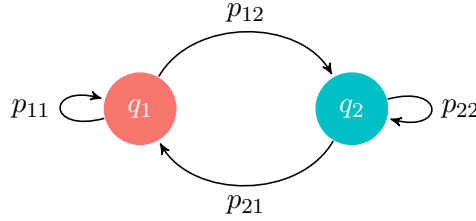
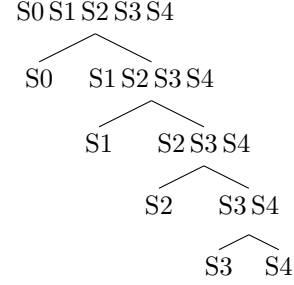


Figure 11. The 2 states automaton describing the smiles machine for each step.

the transition’s probabilities from state q_k at time $t - 1$ towards state q_l at time t , i.e.

$$p_{kl} = p_{kl}(t) \equiv p(q_{t-1} = q_k \rightarrow q_t = q_l) \tag{1}$$

with the indexes k and l being equal to 1 or 2. The random process is therefore fully specified by 4 probabilities of transition: the probabilities p_{11} and p_{22} to remain respectively in state q_1 and state q_2 and the probability of transition p_{12} to evolve from state q_1 toward state q_2 (and reciprocally p_{21} the probability of transition from q_2 toward q_1). These probabilities will be approximated hereafter by a right hand side two terms equation³, i.e.

$$p(q_t|f_t, q_{t-1}) \approx \alpha p(q_t|q_{t-1}) + (1 - \alpha) p(q_t|f_t) \tag{2}$$

where the first term accounts for the dynamic of the system whereas the second term describes the instantaneous relation between the current state and the measurements of the AU intensities. The α coefficient will allow to tune the balance between the two effects.

The expression $p(q_t|q_{t-1})$ stipulates that the probability of transition toward the current state q_t depends only upon the previous state q_{t-1} and not upon higher order term in the state history such as q_{t-2} for example. We do not pretend herein that the real underlying process verify this assumption (i.e. the Markov property). It is rather a convenient mathematical hypothesis which allows to make use of standard algorithms for computing the sequence of maximal probability. In particular, the Viterbi algorithm (Forney, 1973; Viterbi, 1967) will be used for recovering the best solution which is proposed as the output of our automatic annotation tool. The expression of the probability mentioned equation 2 differs however from the one describing the general class of Hidden Markov Model (HMM,

³The variables and parameters entering the equation are defined step by step as the text progresses.

see for example (Rabiner, 1989)). Some methods such like the EM algorithm for parameters estimation (Dempster, Laird, & Rubin, 1977) will therefore not be applicable to the present case.

The second term of the right hand side of equation 2, $p(q_t|f_t) \equiv p(q|f)$, models the instantaneous relation between the states of the model and the values of the AU intensity measurements represented herein by the function $f \equiv f(I_{AU})$. For a given frame of the video and its associated AU intensity measurements, the term $p(q|f)$ specifies the probability of the system to be whether in state q_1 or in state q_2 . For each of the 4 steps indexed by j , the 4 composite functions $f_j(t)$ will be chosen in practice as a linear combination of the AU intensity $I_i(t)$, i indexing the AU type:

$$f_j(t) = \sum_i \beta_{ij} I_i(t) \quad (3)$$

For each step of the overall process, the coefficients are normalized (i.e. $\sum_i \beta_{ij} = 1$) and allow to account for the relative contribution of the specific AU type to the considered step. The composite function $f_j(t)$ are therefore expressed on a continuous scale varying from 0 to 5 as the individual AU intensities $I_i(t)$ provided by the OpenFace software.

4.3 The training stage

The training stage consists in extracting the parameters of the model from the manually annotated corpus (presented section 3.3). The training is performed for each of the 4 steps of the analysis. Starting from step 1, a q_1 or q_2 state is allocated to each frame of the video according to the value of the SIS manual annotation (i.e. q_1 for S0 intervals, q_2 for the remaining group of intervals annotated S1, S2, S3 and S4). Step 2 will consider only time intervals belonging to this group, letting aside the S0 annotations. The new states q_1 and q_2 for step 2 are allocated (i.e. q_1 to S1 annotations and q_2 for S2, S3 and S4 group) and the procedure is iterated until step 4. Table 6 summarized the number of video frames at each step of the analysis and the number of frames respectively in state q_1 or q_2 .

Table 6

For each step of the model specification, the number of frames of the subsamples and the respective proportions of annotated data associated with the q_1 and q_2 states.

	q_1	q_2	# q_1	# q_2	#frame
Step 1	S0	S1 S2 S3 S4	351 505	322 777	674 282
Step 2	S1	S2 S3 S4	134 715	188 062	322 777
Step 3	S2	S3 S4	80 885	107 177	188 062
Step 4	S3	S4	47 250	59 927	107 177

4.3.1 Definition of the AUs composite function f_j . At each step of the analysis the composite function f_j will be defined as a linear combination of the AU intensities (see equation 3). The coefficients β_{ij} are obtained by a multiple linear regression analysis where the dependent variable is the state of the system (the binary variable $q_t = q_1$ or $q_t = q_2$) and the explanatory variables are the OpenFace AUs intensity measurements $I_i(t)$ varying continuously on a scale from 0 to 5.

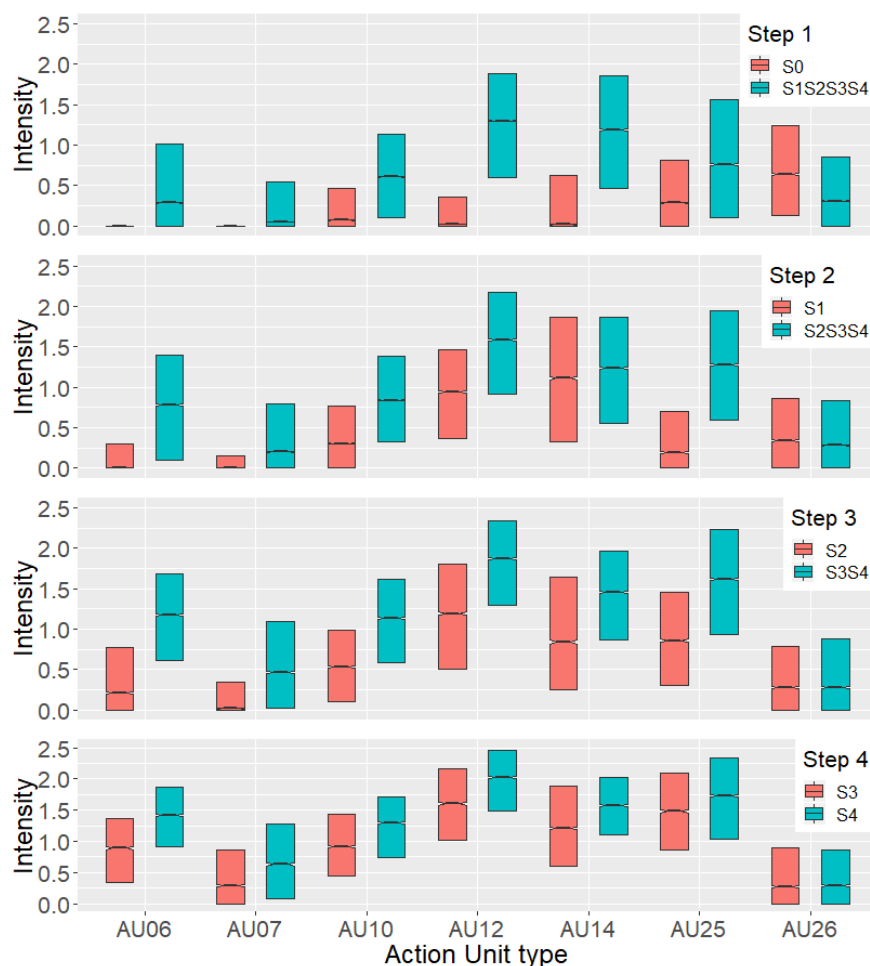


Figure 12. The boxplot distributions of the intensity of each AU for the 4 steps of the analysis.

For each of the 4 steps, figure 12 shows the difference between the boxplot distributions of each AUs intensity for the two states. The median of the intensities distribution is drawn as an inside band mark and the dispersion of the distribution is proportional to the size of the boxplot. For a given AU, the discriminant power between two states is higher when the two boxplots do not overlap. The AU12, which reflects the contraction of the zygomaticus major muscle (lip corner puller), is for example the most discriminant action unit for separating during step 1 no-smiling areas (S0 label) from smiling areas labeled from S1 to S4. During step 2, the AU25, which is activated when showing upper teeth, reveals discriminant boxplot differences for distinguishing between S1 label (closed mouth smiles, see table 2) from the group S2, S3 and S4 encoding open mouth smiles, wide open mouth smiles and laughter.

Table 7 summarizes the final values of the coefficients β_{ij} entering the definition of the f_j composite functions for the 4 steps. They have been computed by performing a multiple linear regression analysis for each step. The AU contributions less than 5% of the

Table 7

For each step of the automatic annotation procedure, the coefficients defining the 4 composite linear functions of the AU intensity.

Name	Step 1	Step 2	Step 3	Step 4
AU06	-	0.45	0.67	0.62
AU07	0.26	-	0.15	0.21
AU10	-	-	-	-
AU12	0.64	0.22	-	0.17
AU14	-	-	-	-
AU25	0.10	0.33	0.18	-
AU26	-	-	-	-

total have been turned off (i.e. $\beta_{ij} = 0$) and the remaining coefficients are normalized in such a way that they sum to 1. During the step 1 for example, the main contribution is brought by AU12 (0.64) which confirm the results suggested above by the boxplots analysis. Some action units do not take part in the definition of the f functions because they are not discriminant for the smile detection task (this is the case for AU26 for example) or because an action unit strongly correlated with them already contributes to the f definition (this is the case for AU10 and AU14 which exhibit a high correlation with AU12). By construction low values of f_j are associated with the q_1 state whereas high values populate preferentially the q_2 state.

The introduction of the composite function f_j for each step of the analysis allows to reduce the dependency between the binary space of states (q_1 and q_2 values) and the multidimensional space of the AU intensities to a one dimensional relationship. This reduction is illustrated figure 13 where the boxplot distribution of the composite functions f_j are contrasted in function of the state (i.e. q_1 versus q_2) for each of the 4 steps. The coefficients defining the composite functions f_j have been chosen for maximizing the discriminant power between the two states.

4.3.2 Estimation of the conditional probability $p(q_i|f_i)$. Once the variables f_j have been defined, it is straightforward to estimate the conditional probabilities that the system belongs to state q_1 or q_2 , given the values of the AU intensities. One introduces for the f functions a discrete scale by defining 12 bins starting from 0 with a bin width of 0.2. The first bin span the interval $[0, 0.2[$, the second the interval $[0.2, 0.4[$ and the last bin contains values greater than 2.2 (i.e. the interval $[2.2, 5]^4$). For each step, the associated subsample mentioned in table 6 is selected, sliced by bins according to the values of the f_j function and the proportion of frames with state q_1 and respectively q_2 is computed for each bin.

The panels of figure 14 show these proportions for the 4 steps. By construction, the probability of state q_2 follows an S-shaped curve, the probability is smaller for low values of f_j and increases as f_j grows. For step 1 for example, the probability to be in the smiling state q_2 (i.e. labels S1, S2, S3 or S4) is around 10% for a measured f_1 in the bin $[0, 0.2[$ and

⁴By definition the maximal value for the f variables is 5 (see section 4.2). However in practice high value of f are rare and the last bin boundaries have been chosen in order to contain a sufficient number of values.

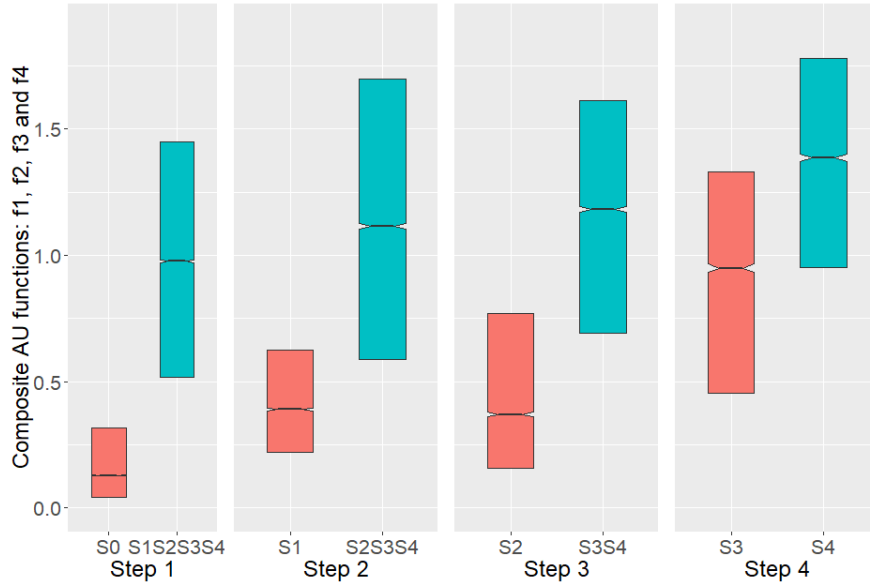


Figure 13. The boxplot distributions of the 4 composite AU functions proper to the 4 steps of the automatic annotation.

Table 8

The matrix of transition probabilities for step 1 ($q_1 = S0$ and $q_2 = S1S2S3S4$). The first row of the matrix is the probability distribution of the current state given that the previous state is q_1 , i.e. $p(q|q_1)$.

	q_1	q_2
q_1	0.9958	0.0042
q_2	0.0046	0.9954

increases to 99% for f_1 in the interval $[1.6, 1.8[$. This difference becomes less marked as one advances through the steps (for the bin $[1.6, 1.8[$, the probability $p(q_2)$ equals 97% for step 2, 92% for step 3 and falls to 73% for step 4).

4.3.3 Estimation of the transition probability $p(q_t|q_{t-1})$. The 2×2 matrix representing the transition probability $p(q_t|q_{t-1})$ can be obtained for each subsample associated to each of the 4 steps by considering the pairs of adjacent states (q_{t-1}, q_t) and by partitioning them in 4 subsets following their values (i.e. (q_1, q_1) , (q_1, q_2) , (q_2, q_1) and (q_2, q_2)). The two probability distributions $p(q|q_1)$ and $p(q|q_2)$ can afterwards be extracted. Table 8 gives these probabilities for step 1. The probability that the system remains in state q_1 is equals to 0.9958 and that conversely it transits toward state q_2 is 0.0042. Note that the values of these probabilities of transition depend closely on the frame rate of the video. The results presented table 8 have been obtained with a video frame rate of 25 frames per second.

4.3.4 Estimation of the α parameter. For each step of the analysis, the α parameter balances the contribution of the two terms of the model described in equation 2. For α equals 0, the model will behave as if the frames of the video were independent. It will therefore decide based solely on the current frame value of the composite function

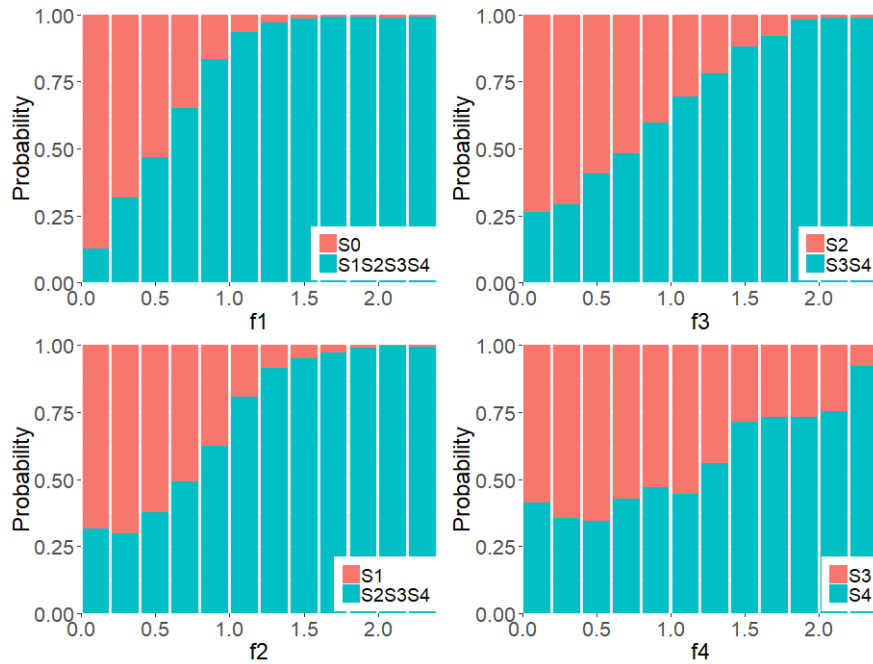


Figure 14. The states probability in function of the 4 composite AU functions proper to the 4 steps of the automatic annotation.

Table 9

The α parameters which maximize the f -measure for each step of the automatic annotation procedure

	α_{\max}	f -measure	recall	precision
Step 1	0.68	0.809	0.748	0.882
Step 2	0.65	0.790	0.727	0.866
Step 3	0.65	0.800	0.792	0.809
Step 4	0.68	0.727	0.760	0.696

f (i.e. the combination of the AU intensities) if the frame belongs to state q_1 or q_2 . On the other hand, as α increases, the model will refrain the system to change of state due to fluctuations of the f value since the probability of transition to change of state is far smaller than the probability to remain in the same state (see the example of the state transition matrix given table 8). The first term $p(q_t|q_{t-1})$ of equation 2 acts therefore as an inertial force which prevents the system to change of state too quickly.

The α parameter for each step has been chosen in such a way that it minimizes the discrepancy between the model prediction and the manual annotations of the gold standard. The results of the minimization are summarized table 9. The α values are similar for the 4 steps.

Figure 15 illustrates the impact of the *inertia* parameter α on the predicted output of the model. The figure is drawn for step 1, which consists in slicing the time line in intervals with no-smile (labeled S0, q_1 state) versus intervals where smile occurs (S1, S2, S3 or S4 labels, q_2 state). The top left panel shows the histogram of the duration in seconds for the

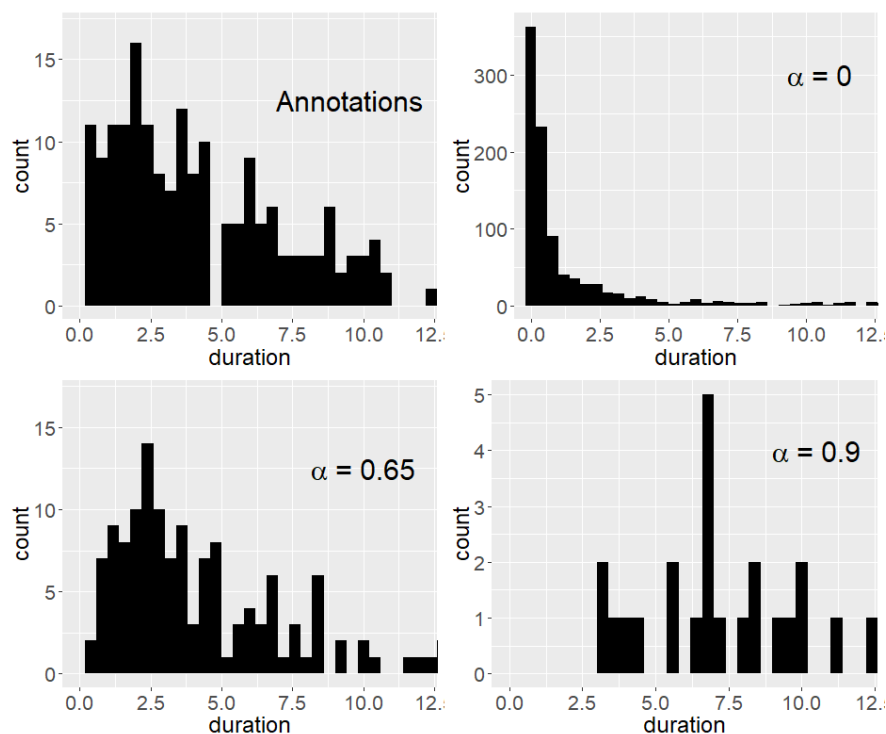


Figure 15. The distribution of the duration (in seconds) of the S0 intervals for the manual annotations (top left panel) and for the automatic annotations provided by the model for 3 values of the α parameter during step 1.

S0 intervals manually annotated on the training corpus. The three other panels show the same histogram for the model predictions for 3 values of the α parameter.

The bottom left panel is for $\alpha = 0.65$, the value which maximizes the f-measure between the annotations and the predictions. The distribution of S0 intervals looks similar to the one manually annotated in term of numbers and duration. The top right panel shows the predicted distribution for $\alpha = 0$. The predicted intervals are much more numerous and too short in duration when compared with the annotated ones. On one hand, it appears that the model with an inertial parameter turned off predicts far too much S0 intervals of only one frame long (40 milliseconds in duration). In that case, the model is too sensitive to the short time fluctuations of the composite function f due to other phenomenon than smile action (for example the contraction of zygomaticus muscles during speech production). On the other hand, a large α parameter (e.g. $\alpha = 0.9$, bottom right panel) does not predict enough S0 intervals of mild duration (i.e. around 1 second) and rather proposes too large intervals without accounting for the variation of the composite function f .

4.3.5 The states probability of transition. The model is now complete. For each of the 4 steps of the analysis, the 4 ingredients have been obtained:

- The optimal coefficients entering the definition of the composite functions f_j (see subsection 4.3.1)
- The conditional probabilities of the q_1 and q_2 states by bins of f_j (see subsection 4.3.2)

- The probabilities to transit from the previous state to the current state (see subsection 4.3.3)
- The inertia parameter α (see subsection 4.3.4)

This allows to compute the quantity $p(q_t|f_t, q_{t-1})$ of equation 2 which is the probability of transition from the previous state q_{t-1} to the current state q_t accounting for the current value of the composite AU intensities function f_t . Figure 15 presents these probabilities of transition for each of the 4 steps of the analysis. For each step, the top panel shows the probability of transition from state q_1 towards states q_1 or q_2 depending on the values of the composite function f and the bottom panel the probability of transition from state q_2 towards states q_1 or q_2 . These $4 \times 2 \times 12$ bin probabilities fully specify the stochastic process underlying the model.

4.4 The best solution

The automatic annotation task proceeds step by step. During step 1, the f_1 function combination of the AU intensities is computed at each frame of the video. Accounting for these f_1 values, the probabilistic model allows to compute the probability associated with any sequence of q_1 and q_2 states, each video frame receiving whether the q_1 or the q_2 state values. Among all the possible sequences, we will select the best solution, i.e. the sequence of states which has the maximal probability. This is done by applying the Viterbi algorithm ((Forney, 1973; Viterbi, 1967)), a dynamical programming algorithm which allows to retrieve the maximal probability solution without exploring the entire space of possible sequences.

The sequence corresponding to the best solution is afterwards splitted into q_1 intervals which receive the S0 label and q_2 intervals corresponding to the compound group of S1, S2, S3 and S4 labels. The S0 intervals will thereafter constitute the S0 outputs of the automatic annotation tool. Step 2 will apply on each of the remaining compound S1 S2 S3 S4 intervals taken individually. For each of them, the f_2 values are computed and the maximal sequence of states $q_1 \equiv S1$ and $q_2 \equiv S2 S3 S4$ is selected. The process is iterated and ends up with step 4 which proposes the best solution for slicing the compound S3 S4 intervals in time areas annotated S3 or S4. The 4 steps of the iterative process and the resulting final output are illustrated figure 17.

4.5 The reliability of the automatic annotation

We have seen previously (section 2.2) that the OpenFace output is sometimes missing or erroneous due to face occlusions, head positions too far from the frontal pose, rapid movements and so on. We combined the information given by OpenFace about the confidence level of detection and our own rejection criteria to propose a measure of reliability. We introduce a filter function $F(t)$ associating to each frame a boolean value which becomes *true* if at least one of the following criteria is met:

- The confidence level given by the OpenFace output is less than 0.8
- The head angle in pitch, yaw or roll angle is too large (greater than 45 degrees)

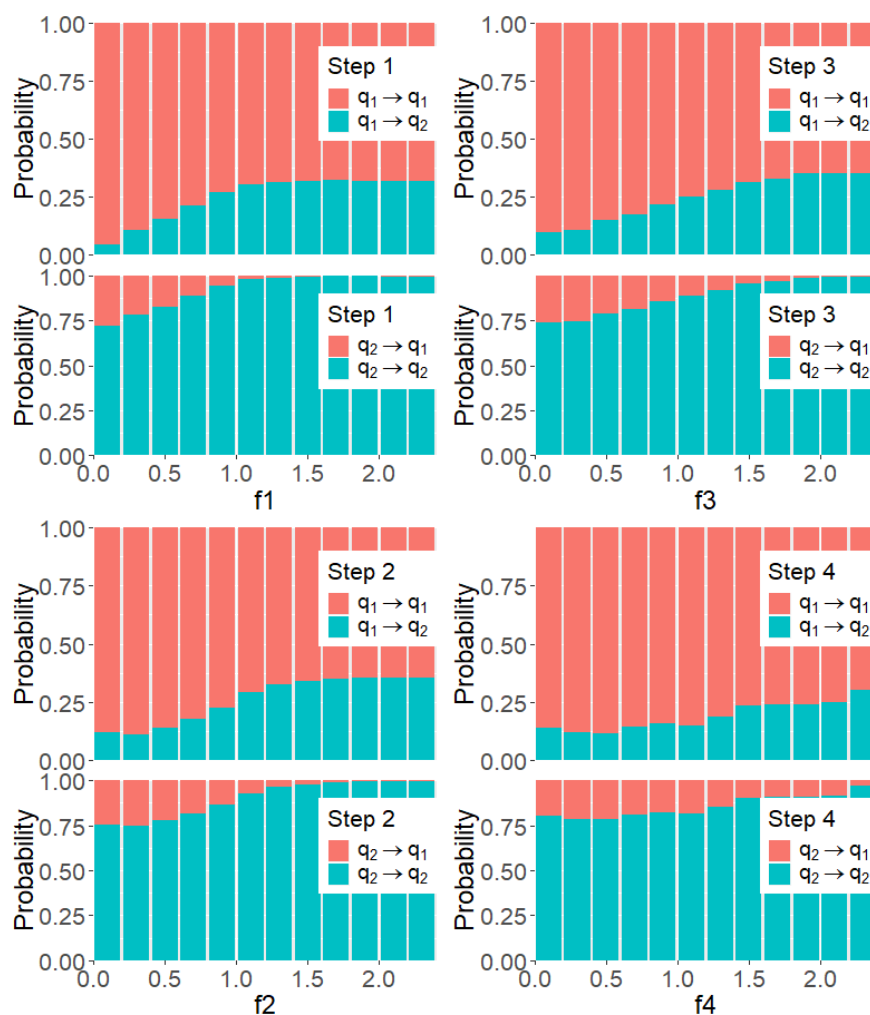


Figure 16. For each of the 4 steps (top left and right and bottom left and right panels), the state probabilities of transition in function of the bin values of the composite functions f . Each panel is subdivided in a top panel which presents the probability of transition from state q_1 (e.g. S0 for step 1) towards q_1 or q_2 and a bottom panel which presents the probability of transition from state q_2 (e.g. the compound group S1 S2 S3 S4 for step 1) towards q_1 or q_2 .

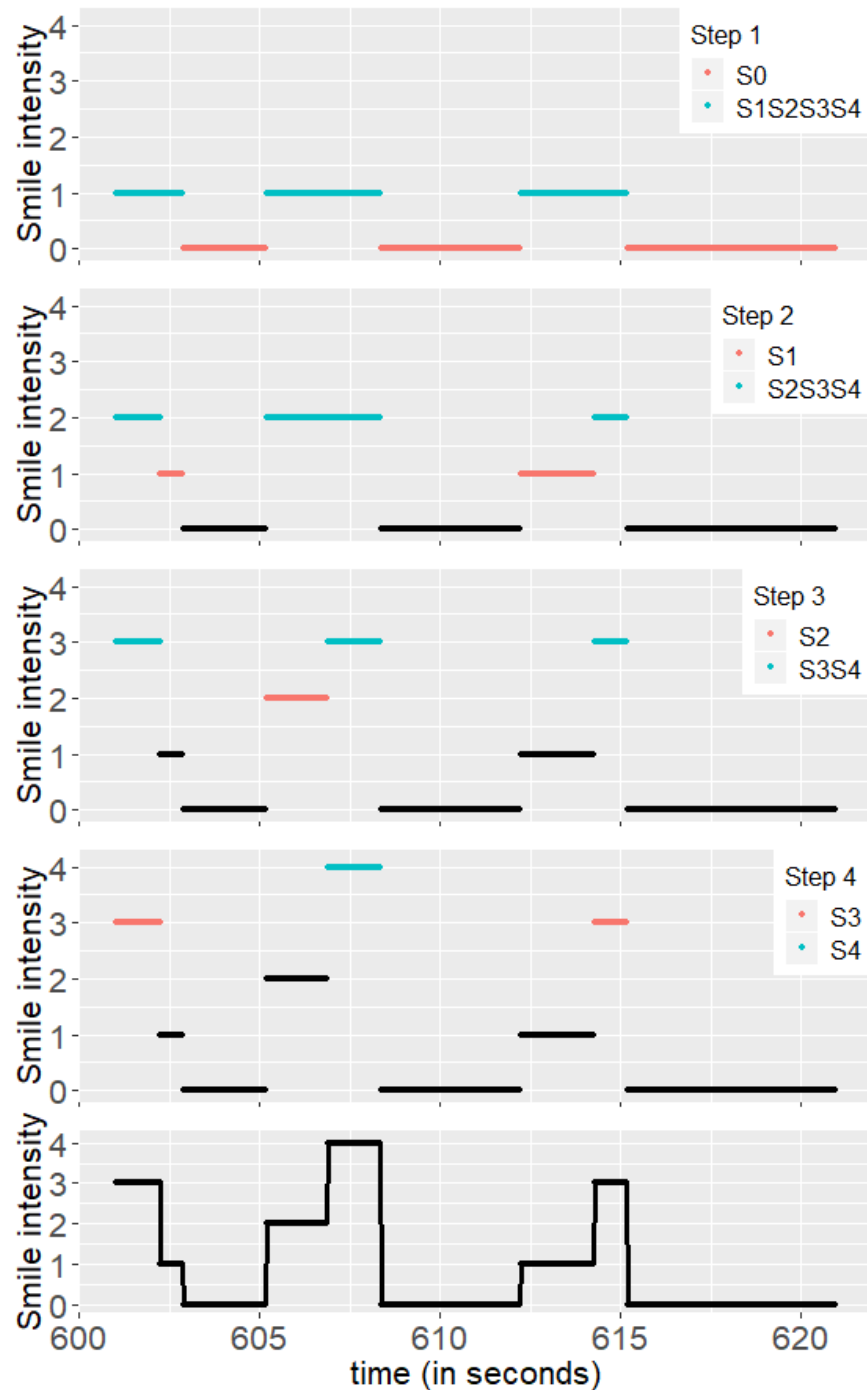


Figure 17. Illustration of the step by step processing applied to the data. During the first step, the time line is sliced in intervals belonging whether to group 1 (S0 annotations) or group 2 formed by the remaining annotations (S1, S2, S3 and S4 all-in-one). During the second step, the intervals of this group are sliced in intervals S1 and group S2, S3 and S4. The process is repeated until step 4 which consists in slicing the remaining intervals in area annotated S3 or S4. At the end of the process, the time line is sliced in contiguous intervals labeled with intensity from 0 to 4 (bottom panel).

- The effective size of the head (i.e. the scaling parameter) is too high or too small compared to the mean value (greater than 3σ)

The filter function signals the frames which are likely to be problematic to some extent. By smoothing the $F(t)$ function with a smoothing window of size 0.4 seconds and by applying a given threshold, we obtain finally a set of time intervals with a questionable reliability. These areas receive a special mark (i.e. X intervals) indicating that the automatic annotation is not available.

5 Evaluation of the tool

We propose two ways for evaluating the performance of the automatic smile annotation tool SMAD⁵. The first will consist in comparing the outputs of the automatic annotation engine with the manually annotated gold standard. The classical measures of precision, recall, f-measure and Cohen’s κ coefficient will give an idea of the performance of the tool. A second and more concrete evaluation will concern the time spent for annotating a video, starting from the labeled time intervals furnished by the output of the tool. In that case the task will consist in manually correcting the labels and the frontiers of the time intervals given by the automatic tool. The gain in annotation time will then be compared with the time required for manually annotating a video without any pretreatment.

Hereafter, we will not compare the performance of our system with other smiling annotation tools. The reasons are twofold. Firstly, as the evaluation procedure relies on a specific smiling scale, the performance measures will therefore strongly depend on this smiling scale. Indeed, the different numbers measuring the observed agreement will dramatically differ depending on whether a binary scale such as the presence or absence of smile is considered (An et al., 2015; Chen et al., 2017; Guo et al., 2018; Shan, 2012; Zhang et al., 2015) or whether a multi-level smiling scale such as the SIS scale is adopted. Among the studies proposing a multi-level smiling intensities scale (Bartlett et al., 2003; Bartlett et al., 2006; Girard et al., 2015; Jiang et al., 2019; Shimada et al., 2010; Vinola & Vimala Devi, 2019), the number of levels characterizing each scale is also a crucial parameter severely impacting the result of the evaluation. Secondly, another source of problem hindering a direct comparison lies in the heterogeneous experimental conditions proper to each study. For example our tool aims at proposing a dynamical SIS scale automatic annotation for subjects involved in a face-to-face conversation. A large part of the video record is composed in that case of scenes inherently containing facial movements caused by speech production. These “noisy” facial movements will make to some extent the smile detection more tricky. A comparison with results obtained on smile annotated corpus not affected by speech production such as for example the GENKI-4K database⁶ would therefore prove to be irrelevant.

⁵SMAD stands for Smile Motion Automatic Detection.

⁶The MPLab GENKI-4K Database (<http://mplab.ucsd.edu/>) is a 4000 face images database with a wide range of subjects of different ages and races and with variable pose, illumination and imaging conditions. Among the GENKI-4K database 2162 images are labeled as smile and 1838 as non-smile. The GENKI-4K database is often used as a benchmark to evaluate automatic smile detection tools (e.g. (An et al., 2015; Chen et al., 2017; Guo et al., 2018; Shan, 2012; Zhang et al., 2015)).

5.1 Precision, recall, f-measure and κ coefficient

We first evaluated the performance of our automatic smile annotation tool by computing the classical measures of precision, recall, f-measure and Cohen’s κ coefficient. For the sake of clarity the detailed examination of the evaluation has been moved to appendix A and we report herein the main results of the evaluation step. The performances are satisfying for class S0 (neutral facial expression) and class S4 (laughter): 90.6% of the frames manually annotated S0 have been successfully predicted and 79.9% of the frames which were predicted S0 are also manually annotated S0. These ratios are respectively equal to 77% and 57.1% for the laughter class S4. The confusion matrix reveals nevertheless that intermediate classes S1, S2 and S3 are more difficult to disentangle, even if those predictions remain useful.

5.2 The gain in annotation time

A second evaluation of the tool performance has been conducted in term of time saving and annotation quality. In order to perform the evaluation of time saving, the first stage was to manually correct the labels (from S0 to S4) and the interval boundaries of the automatic outputs. The second stage is to compare the time required to correct these outputs with the time spent for the manual annotation without pretreatment.

In practice, the evaluation was performed on 4 participants of the CHEESE! corpus (AG-ER and AC-MZ, approximately 15 minutes each). Those 2 interactions are different from the 2 gold standard interactions. The automatic annotations were manually corrected by an expert judge of the SIS. Correcting the 4 automatic outputs required only 6 hours. It means that the judge spent 1 hour on average to correct 10 minutes of video record. By contrast, an expert annotator spends on average 1 hour to manually annotate 1 minute of video featuring one participant (60 hours for the overall 1 hour videos of the gold standard, see section 3.3). The adopted procedure reduces therefore the annotation time by a factor 10. This powerful result allows to consider the annotation task on a larger scale since it provide the opportunity to investigate broader corpus.

The evaluation task deserves a closer investigation. The correction of the automatic outputs consists in modifying the smiling intensity labels and adjusting the time boundaries of the predicted intervals. Concerning the labels, the results show that 76% of the video frames automatically annotated are properly predicted. This global accuracy could appear surprising at first glance since this value is greater than the micro-averaged F-Measure of 68% (see A). One possible explanation could be that the exposure to the predicted labels provided by the automatic tool influences the expert’s judgement.

A detailed inspection reveals that 89% of the video frames automatically labeled S0 have been let unchanged (so 11% have received a correction) and that this correction rate grows respectively to 32% for S1, 43% for S2, 45% for S3 and finally decreases to 33% for laughter class S4. These results are consistent with those obtained in section A: the frequent classes S0 and S4 require less correction than the intermediate classes.

Concerning the annotations of the 4 videos evaluated, 842 intervals have been automatically detected. Among those, 12% corresponds to questionable X intervals (see section 4.5). Analyzing the manual correction reveals that the automatic tool produces more intervals than needed. Indeed, at the end of the correction step 25% of the interval boundaries

have been removed. Among the boundaries which were left, 64% correspond to the initial location given by the automatic tool which means that 36% have been modified with a time shift.

6 The HMAD and SMAD softwares

The scripts and source codes of our automatic smile annotation machine can be downloaded at the following github url address:

`https://github.com/srauzy/HMAD`

The main objective of the HMAD project (the acronym stands for *Head Movement Automatic Detection*) is to provide scripts allowing to detect automatically head and facial movements from a video record. The project makes use of existing solution such the state-of-the-art OpenFace toolkit (Baltrušaitis et al., 2015; Baltrušaitis et al., 2016; Baltrušaitis et al., 2018) in order to analyze the video signal, track the face, provide head pose estimation, perform facial landmark detection, facial action unit recognition, and eye-gaze estimation. The HMAD scripts are therefore an additional layer which propose automatic tools for specific head or facial movements detection task. The EBMAD subproject (*EyeBrows Movement Automatic Detection*) is specialized in the detection of eyebrows raising and frowning actions (see (Rauzy & Goujon, 2018) for details) whereas the SMAD subproject (*Smile Movement Automatic Detection*) is devoted to smiles automatic annotation.

The HMAD and SMAD programs are scripts written in R (R Core Team, 2016) but no specific knowledge about the R language is required for running the HMAD and SMAD commands. The wiki pages of the project detail how to proceed to third-party installations (i.e. the OpenFace toolkit (Baltrušaitis et al., 2018) and the R or RStudio softwares (R Core Team, 2016; RStudio Team, 2015)).

The HMAD commands enable to create a project associated to each video to be treated, to launch OpenFace on that video and to transform the processed OpenFace output in a suitable format compatible with HMAD inputs. A postprocessing treatment is also performed at this stage which allows for example to identify sequences of frames where the measurements are likely to be problematic as mentioned section 4.5. A complete description of the data files created and their associated formats can be found in the wiki pages of the HMAD project.

The SMAD commands launch the automatic smile detection engine on the HMAD outputs and generate the automatic annotation of smile intervals following the SIS system. The SMAD output consists therefore in a sequence of adjacent time intervals which are labeled with a smile intensity varying from S0 to S4 or which have received the special mark X indicating that the scene corresponding to this time interval has to be checked manually.

The automatic outputs may be edited through a multimodal annotation tool, for example the ELAN software (Sloetjes & Wittenburg, 2008). In that aim, SMAD includes a command which creates an output compatible with the ELAN Annotation Format (i.e. the *eaf* file extension). An example of SMAD output edited through the ELAN interface is illustrated figure 18.

HMAD is an open source and collaborative project, so please feel free to contribute in bringing fresh thinking, pieces of code, new challenges, etc.

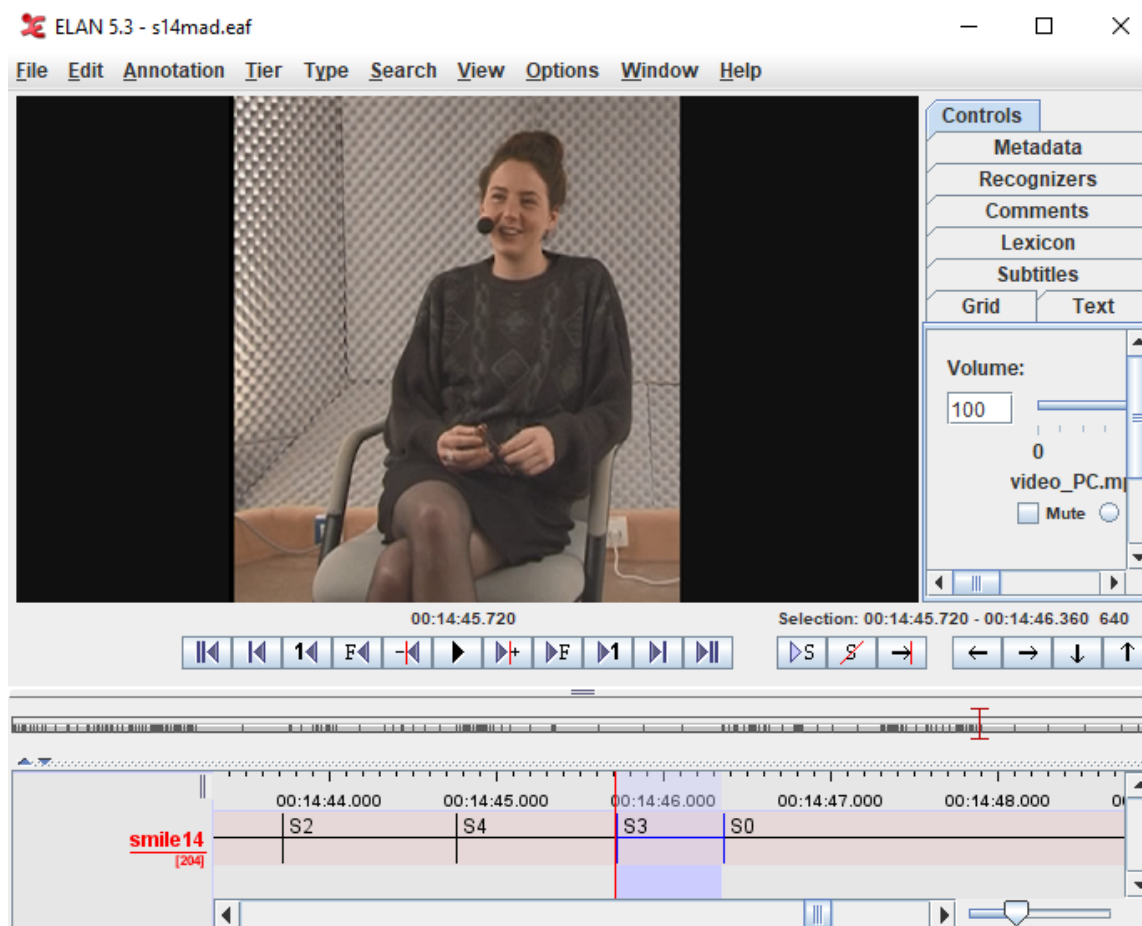


Figure 18. Example of SMAD output formatted for the ELAN software.

7 Conclusions and perspectives

We presented herein the building up of an automatic tool for sequencing a video record in a series of adjusted time intervals labeled following the 5 levels *Smiling Intensity Scale* of Gironzetti et al. (2016). The dynamic of the smile activity was modeled by an ad-hoc stochastic process driven by the measured intensities of the facial Action Units associated with smile actions. The statistical model was trained on a manually annotated gold standard corpus consisting of 2 pairs subjects in face-to-face conversations of approximately 15 minutes.

The performance of our automatic smile annotation tool was first evaluated by computing the classical measures of precision, recall, f-measure and Cohen's κ coefficient on the gold standard itself. The results are satisfying for class S0 (neutral facial expression) and class S4 (laughter): 90.6% of the frames manually annotated S0 have been successfully predicted and 79.9% of the frames which were predicted S0 are also manually annotated S0. These ratios are respectively equal to 77% and 57.1% for the laughter class S4. The confusion matrix reveals nevertheless that intermediate classes S1, S2 and S3 are more difficult to disentangle, even if those predictions remain useful.

A second and maybe more concrete evaluation reveals that the tool can be used with benefits for annotation purpose. An experiment conducted on video records from the CHEESE! corpus (different from the gold standard) reveals that manually correcting the labels and interval boundaries of the automatic outputs reduces by a factor 10 the annotation time. Indeed, 6 hours were necessary to correct 1 hour of video whereas 60 hours were necessary to manually annotate 1 hour of video without pretreatment. This major result answers to our main concern which was to solve the time consuming issue related to the manual annotation task. SMAD is thus an important contribution and is expected to bring a clear-cut answer in the landscape of multimodal annotation and facial recognition.

Our annotation engine is pipelined with the output of the state-of-the-art toolbox OpenFace which allows to track the face and to measure the intensities of the facial Action Units of interest all along the video. The source code of the SMAD software and the documentation are available to download at the HMAD open source project url <https://github.com/srauzy/HMAD>.

Thanks to SMAD, two face-to-face conversational audio-video corpora⁷ composed of 26 dyadic interactions (8h of conversations which represents 16h of individual video track) will be automatically annotated and manually corrected. We estimate the time cost for this manual correction at around 96 hours (16×6 hours), whereas this operation would have taken 960 hours (16×60 hours), with our manual procedure without pretreatment. This massive amount of corrected data will contain a larger diversity of face configurations and smiling phenomena. This new reliable sample will feed the training corpus which will lead to improve the robustness of our smile model.

As said earlier, the presence of speech creates noise in the automatic detection of smiling. One perspective of this study is to take into account the speech activity in smile annotation. This speech activity can be detected independently by automatically analyzing the audio track. It will give the opportunity to split the gold standard in intervals of speech and silence (Inter Pausal Unit : speech units separated by at least 200ms silence). It will be then possible to create two smile detectors, one specialized in silence areas and the other in speech intervals, which can be merged afterwards in a single system. Such a system will improve the performance of the automatic smile detection tool.

It is also worth to mention that the methodology underlying our automatic smile detector can be transposed to other types of gesture. Thanks to the recent advances in the field of Machine Learning and Pattern Analysis, new tools have been developed offering the opportunity to automatically capture arms, hands and body motions (see for example the software OpenPose (Cao, Hidalgo Martinez, Simon, Wei, & Sheikh, 2019)). As SMAD generates smile annotations from OpenFace outputs, one can imagine to apply the same approach to automatically annotate arms, hands and body gestures from OpenPose outputs. Several works exploring this track have been already proposed (e.g. Seger, Wanderley, and Koerich, 2014, Schneider, Memmesheimer, Kramer, and Paulus, 2019, Kowdiki and Khaparde, 2021, ...).

At last, the SMAD tool and its potential improvements will enable to automatically annotate smile intensities of several multimodal corpus. This large amount of data will lead

⁷This data set composed by CHEESE! and PACO, will be used to analyse the role of smiling in the organization of the conversation

to a deeper understanding of the role of the smile in conversational interactions.

References

- Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and cognitive processes*, 15(6), 593–613.
- Amoyal, M., & Priego-Valverde, B. (2019). Smiling for negotiating topic transitions in French conversation. In *Gesture and Speech in Interaction, GespIn 2019*, Paderborn, Germany.
- An, L., Yang, S., & Bhanu, B. (2015). Efficient smile detection by extreme learning machine. *Neurocomputation*, 149(PA), 354–363. doi:10.1016/j.neucom.2014.04.072
- Argyle, M. (1975). *Bodily communication*. Methuen : London.
- Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4), 555–596. doi:10.1162/coli.07-034-R2
- Baltrušaitis, T., Mahmoud, M., & Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *11th IEEE International Conference on Automatic Face and Gesture Recognition*.
- Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2013a). 3D constrained local model for rigid and non-rigid facial tracking. In *Cvpr, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.
- Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2013b). Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops* (pp. 354–361). doi:10.1109/ICCVW.2013.54
- Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). OpenFace: An open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*.
- Baltrušaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)* (pp. 59–66). doi:10.1109/FG.2018.00019
- Bartlett, M. S., Littlewort, G. C., Braathen, B., Sejnowski, T. J., & Movellan, J. R. (2003). A prototype for automatic recognition of spontaneous facial actions. *Advances in Neural Information Processing Systems*, 15, 1271–1278.
- Bartlett, M. S., Littlewort, G. C., Franck, M. G., Lainscsek, C., Fasel, I. R., & Movellan, J. R. (2006). Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6), 22–35.
- Bateson, G., Winkin, Y., Bansard, D., Cardoen, A., & Birdwhistell, R. (1981). *La nouvelle communication*. Ed. du Seuil Paris.
- Bavelas, J. B., & Gerwing, J. (2007). Conversational hand gestures and facial displays in face-to-face dialogue. *Social communication*, 283–308.
- Brugman, H., Russel, A., & Nijmegen, X. (2004). Annotating multi-media/multi-modal resources with elan. In *Lrec*.
- Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., & Sheikh, Y. A. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

- Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, *22*(2), 249–254. Retrieved from <http://dl.acm.org/citation.cfm?id=230386.230390>
- Chen, J., Ou, Q., Chi, Z., & Fu, H. (2017). Smile detection in the wild with deep convolutional neural networks. *Machine Vision and Applications*, *28*(1), 173–183. doi:10.1007/s00138-016-0817-z
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.
- Cohn, J. F., & De la Torre, F. (2014). Automated face analysis for affective computing. In R. Calvo, S. D’Mello, J. Gratch, & A. Kappas (Eds.), *The oxford handbook of affective computing*.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal Of The Royal Statistical Society, Series B*, *39*(1), 1–38.
- Dhall, A., Goecke, R., Gedeon, T., & Sebe, N. (2016). Emotion recognition in the wild. *Journal on Multimodal User Interfaces*, *10*(2), 95–97. doi:10.1007/s12193-016-0213-z
- Ekman, P. [P.], Friesen, W., & Hager, J. (2002). *Facial action coding system: Research nexus*. Salt Lake City, UT: Network Research Information.
- Ekman, P. [P.], & Friesen, W. (1975). *Unmasking the face : A guide to recognizing emotions from facial clues*. Englewood Cliffs : Prentice-hall.
- Ekman, P. [Paul]. (1984). Expression and the nature of emotion. *Approaches to emotion*, *3*, 19–344.
- Ekman, P. [Paul], Davidson, R. J., & Friesen, W. V. (1990). The Duchenne smile: Emotional expression and brain physiology: II. *Journal of personality and social psychology*, *58*(2), 342.
- Ekman, P. [Paul], & Friesen, W. V. (1978). *Facial action coding system: Manual*. Consulting Psychologists Press.
- El Haddad, K., Chakravarthula, S. N., & Kennedy, J. (2019). Smile and laugh dynamics in naturalistic dyadic interactions: Intensity levels, sequences and roles. In *2019 international conference on multimodal interaction* (pp. 259–263). ACM.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.
- Forney, G. D. (1973). The Viterbi algorithm. *Proceedings of IEEE*, *61*(3), 268–278. doi:<http://dx.doi.org/10.1109/PROC.1973.9030>
- Freire-Obregón, D., & Castrillón-Santana, M. (2015). An evolutive approach for smile recognition in video sequences. *International Journal of Pattern Recognition and Artificial Intelligence*, *29*(01). doi:10.1142/S0218001415500068. eprint: <https://doi.org/10.1142/S0218001415500068>
- Girard, J. M., Cohn, J. F., & De la Torre, F. (2015). Estimating smile intensity: A better way. *Pattern Recognition Letters*, *66*, 13–21.
- Gironzetti, E., Attardo, S., & Pickering, L. (2016). Smiling, gaze, and humor in conversation: A pilot study. In L. Ruiz-Gurillo (Ed.), *Metapragmatics of humor: Current research trends* (pp. 235–254). doi:10.1075/ivitra.14.12gir

- Gorisch, J., & Prévot, L. (2014). Aix-DVD, LPL. <https://www.ortolang.fr/market/corpora/sldr000891>. Retrieved from <https://hdl.handle.net/11403/sldr000891/v1>
- Goujon, A., Bertrand, R., & Tellier, M. (2015). Eyebrows in French talk-in-interaction. In *Gesture and speech in interaction - 4th edition (gespin 4)* (pp. 125–130). Nantes, France.
- Guo, X., Polania, L., & Barner, K. (2018). Smile detection in the wild based on transfer learning. In *13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)* (pp. 679–686). doi:10.1109/FG.2018.00107
- Guy, B. (2013). *La communication non verbale: Comprendre les gestes: Perception et signification*. ESF Sciences Humaines.
- Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57*(4), 596–615.
- Harker, L., & Keltner, D. (2001). Expressions of positive emotion in women's college yearbook pictures and their relationship to personality and life outcomes across adulthood. *Journal of Personality and Social Psychology*, *80*(1), 112–124. doi:10.1037/0022-3514.80.1.112
- Heerey, E. A., & Crossley, H. M. (2013). Predictive and reactive mechanisms in smile reciprocity. *Psychological Science*, *24*(8), 1446–1455. doi:10.1177/0956797612472203. eprint: <https://doi.org/10.1177/0956797612472203>
- Holler, J., Schubotz, L., Kelly, S., Hagoort, P., Schuetze, M., & Özyürek, A. (2014). Social eye gaze modulates processing of speech and co-speech gesture. *Cognition*, *133*(3), 692–697.
- Jensen, M. H. (2015). Smile as feedback expressions in interpersonal interaction. *International Journal of Psychological Studies*, *7*(4).
- Jiang, H., Coskun, M., Badokhon, A., Liu, M., & Huang, M.-C. (2019). Hidden smile correlation discovery across subjects using random walk with restart. *IEEE Transactions on Affective Computing*, *10*(1), 76–84. doi:10.1109/TAFFC.2017.2774278
- Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta psychologica*, *26*, 22–63.
- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.
- Kent, A., Berry, M. M., Luehrs Jr., F. U., & Perry, J. W. (1955). Machine literature searching VIII. Operational criteria for designing information retrieval systems. *American Documentation*, *6*(2), 93–101. doi:10.1002/asi.5090060209. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.5090060209>
- Kerbrat-Orecchioni, C., & Cosnier, J. (1987). *Décrire la conversation*. Lyon: Presses universitaires de Lyon.
- Kowdiki, M., & Khaparde, A. (2021). Automatic hand gesture recognition using hybrid meta-heuristic-based feature selection and classification with dynamic time warping. *Computer Science Review*, *39*, 100320. doi:<https://doi.org/10.1016/j.cosrev.2020.100320>
- Krippendorff, K. (2008). Systematic and random disagreement and the reliability of nominal data. *Communication Methods and Measures*, *2*(4), 323–338. doi:10.1080/19312450802467134. eprint: <https://doi.org/10.1080/19312450802467134>

- Krumhuber, E. G., Likowski, K. U., & Weyers, P. (2014). Facial mimicry of spontaneous and deliberate Duchenne and Non-Duchenne smiles. *Journal of Nonverbal Behavior*, *38*, 1–11.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33* 1, 159–74.
- Martinez, B., Valstar, M., Jiang, B., & Pantic, M. (2019). Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, *10*(3), 325–347. doi:10.1109/TAFFC.2017.2731763
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.
- McNeill, D. (2012). *How language began: Gesture and speech in human evolution*. Cambridge University Press.
- Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.
- Powers, D. M. W. (2012). The problem with kappa. In *Proceedings of the 13th conference of the european chapter of the association for computational linguistics* (pp. 345–355). Avignon, France: Association for Computational Linguistics. Retrieved from <http://dl.acm.org/citation.cfm?id=2380816.2380859>
- Priego-Valverde, B., Bigi, B., Attardo, S., Pickering, L., & Gironzetti, E. (2018). Is smiling during humor so obvious? A cross-cultural comparison of smiling behavior in humorous sequences in American English and French interactions. *Intercultural Pragmatics*. Retrieved from <https://hal.archives-ouvertes.fr/hal-01923442>
- R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. doi:10.1109/5.18626
- Rauzy, S., & Goujon, A. (2018). Automatic annotation of facial actions from a video record: The case of eyebrows raising and frowning. In *Workshop on "Affects, Compagnons Artificiels et Interactions", WACAI 2018* (7 pages). Magalie Ochs. Porquerolles, France. Retrieved from <https://hal.archives-ouvertes.fr/hal-01769684>
- RStudio Team. (2015). *Rstudio: Integrated development environment for r*. RStudio, Inc. Boston, MA. Retrieved from <http://www.rstudio.com/>
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction* (pp. 7–55). Elsevier.
- Sanders, A. F. (2013). *Elements of human performance: Reaction processes and attention in human skill*. Psychology Press.
- Schneider, P., Memmesheimer, R., Kramer, I., & Paulus, D. (2019). Gesture recognition in rgb videos using human body keypoints and dynamic time warping. In S. Chalup, T. Niemueller, J. Suthakorn, & M.-A. Williams (Eds.), *Robocup 2019: Robot world cup xviii* (pp. 281–293). Cham: Springer International Publishing.

- Seder, J. P., & Oishi, S. (2012). Intensity of smiling in facebook photos predicts future life satisfaction. *Social Psychological and Personality Science*, 3(4), 407–413. doi:10.1177/1948550611424968. eprint: <https://doi.org/10.1177/1948550611424968>
- Seger, R. A., Wanderley, M. M., & Koerich, A. L. (2014). Automatic detection of musicians' ancillary gestures based on video analysis. *Expert Systems with Applications*, 41(4, Part 2), 2098–2106. doi:<https://doi.org/10.1016/j.eswa.2013.09.009>
- Shan, C. (2012). Smile detection by boosting pixel differences. *IEEE Trans. Image Processing*, 21(1), 431–436. doi:10.1109/TIP.2011.2161587
- Shimada, K., Matsukawa, T., Noguchi, Y., & Kurita, T. (2010). Appearance-based smile intensity estimation by cascaded support vector machines. In *Asian conference on computer vision workshops* (pp. 277–286).
- Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, 85(3), 257–268. doi:10.1093/ptj/85.3.257. eprint: <http://oup.prod.sis.lan/ptj/article-pdf/85/3/257/9411262/ptj0257.pdf>
- Sloetjes, H., & Wittenburg, P. (2008). Annotation by category - ELAN and ISO DCR. In *Proceedings of the 6th international conference on language resources and evaluation (LREC 2008)*. Retrieved from <https://tla.mpi.nl/tools/tla-tools/elan/>
- Vettin, J., & Todt, D. (2004). Laughter in conversation: Features of occurrence and acoustic structure. *Journal of Nonverbal Behavior*, 28(2), 93–115.
- Vinola, C., & Vimala Devi, K. (2019). Smile intensity recognition in real time videos: Fuzzy system approach. *Multimedia Tools and Applications*, 78(11). doi:10.1007/s11042-018-6890-8
- Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2), 260–269. doi:10.1109/TIT.1967.1054010
- Walecki, R., Rudovic, O., Pavlovic, V., & Pantic, M. (2019). Copula ordinal regression framework for joint estimation of facial action unit intensity. *IEEE Transactions on Affective Computing*, 10(3), 297–312. doi:10.1109/TAFFC.2017.2728534
- Whitehill, J., Littlewort, G., Fasel, I. R., Bartlett, M. S., & Movellan, J. R. (2009). Toward practical smile detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31, 2106–2111.
- Zhang, K., Huang, Y., Wu, H., & Wang, L. (2015). Facial smile detection based on deep learning features. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 534–538.

Authors' addresses

Mary Amoyal

Laboratoire Parole et Langage

5 avenue Pasteur

BP 80975, 13604 Aix-en-Provence

FRANCE

mary.amoyal@univ-amu.fr

Stéphane Rauzy

Laboratoire Parole et Langage

5 avenue Pasteur

BP 80975, 13604 Aix-en-Provence

FRANCE

stephane.rauzy@univ-amu.fr

About the authors

Mary Amoyal is currently research assistant at the Laboratoire Parole et Langage in Aix Marseille University (France). She received in 2022 a PhD in Linguistics from Aix Marseille University. Her researchs span the field of Interactional Linguistics and investigate the role of facial gestures such as smile during the conversational interaction.

Stéphane Rauzy is currently research engineer employed by the CNRS at the Laboratoire Parole et Langage in Aix Marseille University (France). He received in 1993 a PhD in Theoretical Physics and Astronomy from Aix Marseille University. His main activity is to conceive and develop automatic tools in the field of Natural Language Processing. His area of expertise includes the modeling of complex and multimodal phenomena that can be found in Linguistics.

Appendix

Evaluation by means of confusion, recall and precision matrices

One way to evaluate the performance of the SMAD tool is to make use of the classical measures of precision, recall and f-measure introduced in the field of information retrieval (Kent, Berry, Luehrs Jr., & Perry, 1955). It requires during a first step to compute the confusion matrix between the observed values (herein the manually annotated smile intensities) and the values predicted by the automatic tool. Table A1 presents the confusion matrix for our gold standard corpus. The rows of the matrix stand for the predicted labels (the 5 levels in the SIS system plus the X mark signaling problematic frames) and the columns stand for the observed, actual, manually annotated values. For each video frame, the pair (predicted value, observed value) is formed and one increments the count of the corresponding cell in the confusion matrix. The global count over all the cells is equals to the total number of frames of the corpus.

Table A1

The confusion matrix for the gold standard corpus. Rows stand for predicted (i.e. herein automatic output) values and columns for observed/actual (i.e. herein manually annotated) values. The counts are in number of video frames.

	S0	S1	S2	S3	S4
S0	46446	7645	3187	638	193
S1	4368	8678	2464	681	318
S2	224	1847	2911	1024	523
S3	112	732	1476	1757	1136
S4	108	545	1846	2924	7270
X	1032	545	217	319	480

The intervals of time marked X have to be checked manually. For our gold standard,

Table A2

The confusion matrix of table A1 with the X areas removed and figuring the marginal counts $n_{i.}$ and $n_{.j}$.

	S0	S1	S2	S3	S4	Total
S0	46446	7645	3187	638	193	58109
S1	4368	8678	2464	681	318	16509
S2	224	1847	2911	1024	523	6539
S3	112	732	1476	1757	1136	5213
S4	108	545	1846	2924	7270	12741
Total	51258	19495	11884	7034	9440	99111

Table A3

The recall matrix for our gold standard. Within a given observed class (each column), the proportions of predicted classes sums to 1 by definition.

	S0	S1	S2	S3	S4
S0	0.906	0.392	0.268	0.090	0.021
S1	0.085	0.445	0.207	0.097	0.034
S2	0.005	0.095	0.245	0.147	0.055
S3	0.002	0.038	0.124	0.250	0.120
S4	0.002	0.030	0.156	0.416	0.770
Total	1.000	1.000	1.000	1.000	1.000

they represent a small amount of data (i.e. about 2.55% of the 4 videos duration). The X areas are discarded from the confusion matrix and we are left with a 5×5 predicted versus observed square matrix. Let n_{ij} be the count of frames corresponding to the cell at row index i and column index j , we define hereafter the marginal count $n_{i.} = \sum_j n_{ij}$ as the total count of frames with a predicted label corresponding to row index i and the marginal count $n_{.j} = \sum_i n_{ij}$ as the total count of frames with an observed label corresponding to column index j . These marginal counts are illustrated table A2 for our gold standard data.

The coefficients of the recall matrix \mathbf{R} are defined as $R_{ij} = n_{ij}/n_{.j}$ and trace within each observed class indexed by j the distribution of the predicted classes indexed by i . The recall matrix for the gold standard is given table A3. The results are satisfying. 90.6% of the frames manually annotated S0 have been successfully predicted as S0 (see table A3, column 1 first row). The remaining have been erroneously predicted as S1 at 8.5%, as S2 at 0.5% and 0.2% fall in smile classes S3 and S4. The class of laughter S4 has also a good recall score (i.e. 77%, see column 4) whereas the intermediate classes S1, S2 and S3 are more difficult to predict (i.e. a recall of 44.5%, 24.5% and 25% respectively).

The precision matrix furnishes an orthogonal view of this description. The precision matrix \mathbf{P} is defined as $P_{ij} = n_{ij}/n_{i.}$ and gives the distribution of the observed classes indexed by j within each predicted class indexed by i . The results applied on the gold standard are presented table A4. They read as follows: among the frames which were predicted S0, 79.9% of them belong really to the non-smiling class S0, 13.2% are indeed of the S1 class, 5.5% of the S2 class, 1.1% of the S3 class and 0.3% belong to the remaining class S4. The

Table A4

The precision matrix for our gold standard. Within a given predicted class (each row), the proportions of observed classes sums to 1 by definition.

	S0	S1	S2	S3	S4	Total
S0	0.799	0.132	0.055	0.011	0.003	1.000
S1	0.265	0.526	0.149	0.041	0.019	1.000
S2	0.034	0.283	0.445	0.158	0.080	1.000
S3	0.022	0.140	0.283	0.337	0.218	1.000
S4	0.009	0.046	0.145	0.229	0.571	1.000

table A4 diagonal indicates that 53% of predicted S1 are real S1, 44% of predicted S2 are real S2, 34% of predicted S3 are real S3 and 57% of predicted S4 are real S4.

Precision, recall, f-measure and κ coefficient

The knowledge of the confusion matrix or alternatively of the couple formed by the recall and precision matrices fully specify the results obtained during the evaluation process. However it is sometimes convenient to summarize the whole exercise by a single measurement. To this extent some averaged quantities are proposed in the literature.

For example the precision and recall for each classes (i.e. the diagonal coefficients P_{kk} and R_{kk} populating the precision and recall matrices) can be averaged in order to define the macro average precision P_{macro} and recall R_{macro} :

$$P_{\text{macro}} = \frac{1}{K} \sum_k P_{kk}; \quad R_{\text{macro}} = \frac{1}{K} \sum_k R_{kk} \quad (4)$$

where K is the total number of classes. The score of f-measure is classically defined as the harmonic mean of the precision and recall measurements, i.e.

$$F_{\text{macro}} = 2 \frac{P_{\text{macro}} R_{\text{macro}}}{P_{\text{macro}} + R_{\text{macro}}} \quad (5)$$

The macro coefficients attribute an equal weight to each class whatever their effective frequencies. If one wants to take into account this frequency effect (i.e. frequent classes will contribute more to the average than rare classes), the micro-averaged quantity can be used:

$$F_{\text{micro}} = P_{\text{micro}} = R_{\text{micro}} = \frac{1}{n} \sum_k n_{kk} \quad (6)$$

where n_{kk} are the diagonal counts of the confusion matrix and $n = \sum_i n_i = \sum_j n_j$ is the total size of the evaluation sample. This quantity is also known as the measure of *accuracy* or *observed agreement* and is simply the number of correctly predicted observations over the total number of observations.

These quantities illustrate averaged proportions but do not inform about the performance (i.e. the predictive power) of the classifier tool. Cohen proposed the κ statistic

Table A5

The various agreement coefficients computed from the predictions and the manual annotations for individual participant and for the whole gold standard corpus (last line of the table).

Participant	Cohen’s κ	F_{micro}	P_{macro}	R_{macro}	F_{macro}
MAPC_MA	0.450	0.722	0.494	0.519	0.506
MAPC_PC	0.343	0.643	0.492	0.416	0.451
JSCL_JS	0.513	0.640	0.546	0.544	0.545
JSCL_CL	0.525	0.703	0.548	0.513	0.530
ALL	0.495	0.676	0.536	0.523	0.529

(Artstein & Poesio, 2008; Carletta, 1996; Cohen, 1960; Fleiss, 1971) which compares the agreement value mentioned above with the one which is expected by chance:

$$\kappa = \frac{n_o - n_e}{n - n_e}; \quad n_o = \sum_k n_{kk}; \quad n_e = \sum_k \frac{n_{.k} \times n_k}{n} \quad (7)$$

where n_o is the count of correctly predicted observations and n_e the count of expected agreements computed assuming that the distributions of the predicted and observed values are independent. The Cohen’s κ statistic has a zero mean value if the prediction is made at random and approaches unity if the agreement is maximal. Discussions concerning the limitations of the method and the biases it suffers can be found in (Krippendorff, 2008; Powers, 2011, 2012; Sim & Wright, 2005) for example. Note that herein the ordinal aspect of the smile scale (i.e. S0 class is closer to S1 class than to S4 class) is not taken into account. The five levels of the smiling scale are therefore considered as nominal data.

The results of the evaluation on the gold standard are presented table A5 participant by participant and for the whole corpus as well. The overall accuracy F_{micro} indicates that 67.6% of the labels are correctly predicted with some variability between participants (a minimum of 64% for participant JS and a maximum of 72.2% for participant MA). The macro averaged scores P_{macro} , R_{macro} and F_{macro} which attribute an equal weight among classes are lower (i.e. around 53%). It reveals that the predictive power of the tool is not homogeneous from class to class. That feature was mentioned in the previous subsection, non-smiling S0 and laughter S4 classes which contribute more heavily to F_{micro} because they are frequent show higher scores than the S1, S2 and S3 intermediate classes.

The global Cohen’s κ coefficient is equal to 0.495 which corresponds to a *moderate agreement* according to the Landis&Koch agreement scale (Landis & Koch, 1977). However as pointed out in Artstein and Poesio (2008), the interpretation of agreement coefficient values is “*still little more than a black art*” and depends on various factors such the number of classes considered, their internal relationship and at the end the specific purposes to achieve. For this reason a more concrete evaluation criteria based on the gain in annotation time is herein preferred and is proposed next subsection.

The scores mentioned above have been computed by comparing the predicted automatic output with the manual annotations of the gold standard. This procedure is not recommended in general because of the risk of *overfitting*. Overfitting happens when specific properties of the training data (but not representative of the whole dataset) are learnt

by the model. In that case, the evaluation process applied on the training data themselves leads to overestimate the performance of the system. In order to cope with this potential problem, we performed a cross validation procedure. The gold standard was splitted in two subsamples: a training dataset containing 90% of the gold standard used to compute the fitting parameters of the model and a test dataset containing the remaining 10% of the corpus which was let aside for the evaluation task. We generated 100 random partitions of the gold standard. For each partition, the test dataset consists in 4 time intervals (one per gold standard participant) of 10% duration and with a random starting time boundary. The 8 remaining intervals (2 intervals per participant, preceding and following the test segment) form the training dataset. For each partition, the model parameters are estimated on the training data and a confusion matrix is computed for the test sample. The confusion matrices of the 100 random partitions are merged at the end.

Table A6

The agreement coefficients obtained by applying the cross validation procedure (the training dataset represents 90% of the gold standard and the test dataset the remaining 10%). Results previously presented on the whole gold standard are echoed on the second line.

TRAINING	TEST	Cohen's κ	F_{micro}	P_{macro}	R_{macro}	F_{macro}
90% GOLD	10% GOLD	0.483	0.679	0.517	0.500	0.508
100% GOLD	100% GOLD	0.495	0.676	0.536	0.523	0.529

The results are summarized by the agreement coefficients presented table A6. There is no significant difference when comparing these numbers with the results presented above for the whole gold standard. It means that all along the training stage described section 4.3, we have controlled the risk of overfitting in particular by limiting the number of parameters entering the statistical model.

One can not exclude however the case of some special smiling configurations absent from the gold standard and therefore not captured by the engine. The cross validation procedure will be unfortunately of no help at detecting that case. For example our gold standard is rich of only 4 participants which may fail to represent the diversity of smiling behaviours. This last point will be discussed later on in our conclusive section 7.