Automatic facial expressions analysis during speech communication

Mikhail Baev[a], Alexey Gusev[b]

[a]*LLC Uchet-N,*
Pestelya str. 11-76, St. Petersburg, 191028, Russia
mikhail.baev@gmail.com

[b]*Faculty of Psychology of Lomonosov Moscow State University,*
Mokhovaya st. 11/9, Moscow, 125009, Russia
angusev@mail.ru

***Abstract*— In the article that follows, we are presenting the readers with the results of our research conducted as part of development of a computer system intended for facial expressions analysis. Here we are emphasizing certain problems related to peculiarities of facial expressions categorizing while carrying out automated analysis. The problems stem from, first, articulatory movements and FACS action units (AUs) time-linked overlap; and second, from the necessity to differentiate mimic events varying in meaning, i.e., to separate action units and their combinations as basic emotions indexes, semantic stressing of a speech message, emotion emblems, and mimic dialects.**

We have managed to secure high precision of facial expressions analysis outcomes, applying the so-called FACS-based comprehensive approach, instead of the selective approach, considering core flaws of the latter. We have developed EmoRadar, a computer system specializing in analysing videos; it functions based on direct analysis of facial surface, relying on original protocols of computer vision and is, in fact, an implementation of computer FACS concepts.

The software empirical testing has revealed the necessity to consider specifics of detection and categorization of certain action units and their combinations against the background of articulatory movements during speech communication. Differentiating mimic events varied in meaning, in our opinion, is possible only when based on high-precision analysis of the time of emergence and ending of the action units that are part of mimic events.

***Keywords*— FACS, emotions, affective computing, facial expressions, speech**

I. Introduction

At present, computer systems for automated analysis of a person's emotional state by his/her facial expressions are getting more and more acclaimed in both academic and applied studies of psychology and related subjects. In our opinion, however, there is an obvious lack of software intended for research and practical work that require exact and objective assessment of an emotional state of the person whose face appears in the video.

It should also be stressed that at the moment, there isn't any automated system available that has the capacity to analyse facial expressions' changes within the context of speech communication.

Implementing the Facial Action Coding System (FACS) as a world-renowned comprehensive facial movements description system into any means of automated analysis of facial expressions is complicated by the fact that FACS has not been initially developed as an instrument for the articulatory movements description (personal communications with E. Rosenberg, June 2015). Thus, very little attention is paid to this issue is the FACS Manual [5]. We do agree with P. Ekman's point that "computer-drive precise measurement of facial actions (cFACS) will soon be possible" [4] in both academic and applied research works.

Another important aspect of FACS implementation is the issue of facial expressions' wide variability as indexes of affective states ([1], [3]).

The aim of this publication is to demonstrate the viability of facial expressions' differentiated assessment as part of automated analysis of videos in the context of speech communication. It is our attempt to go beyond the basic emotions expressions framework, set in the FACS Investigator's Guide [5], and other working models used to describe the reflection of affective processes within facial expressions ([3], [10]).

II. Key principles of the software development

We have developed computer FACS (cFACS) software EmoRadar WR for fully automatic analysis of facial expressions in a video. Our algorithms of computer analysis of facial expressions are based on the following principles:

- FACS seen as an instrument of implementing the comprehensive approach (as opposed to the selective approach) to the analysis of facial expressions [8]; the reason for this choice is that it allows to describe all the possible variations of facial expressions;
- Direct measurement of the lighting changes on the facial surface done with the authentic procedures of computer vision specifically focused on detecting AUs as the base units of facial expressions analysis;
- Deliberate rejection of employing neural networks for facial events classification;
- Modelling an expert's perception of facial surface movements' peculiarities while detecting certain AUs.

III. The outcomes

While developing the software, we faced two major problems:

(1) Necessity to distinguish the articulatory movements proper from other types of facial expressions;

(2) Need for differentiated assessment of facial expressions as nonverbal communication instruments that accompany verbal messages, and facial expressions as expressions of emotions per se.

The solution of the first problem is linked to singling out speech fragments in the video and matching them with certain AUs on a time-defined basis.

The problem was being solved in two ways:

(1) Fragments of voiced speech were marked in the audio track;

(2) Around the mouth area, we marked the facial surface movements, specifically unrelated to the speech production function. It was done by choosing the thresholds of the following AUs that are part of basic emotions expression patterns, i.e., numbers 9, 10, 12, 14, 15, 17, 20, 24 ([3], [5]).

Figure 1 shows the screenshot of our software interface displaying the outcomes of facial activity analysis done in the course of speech production. The intervals corresponding to the voiced speech are marked as AU 50 on one of the horizontal lines. Facial activity events, automatically categorized as FACS AUs are marked on the other lines of the timeline: (A) AU 12 and 20 emerging against the background of articulation; (B) AU 10 is amplifying the articulation by the way of emphasizing the meaningful speech fragment in order to strengthen the communicative effect.

The solution to the second problem lay within the implementation of the above-mentioned comprehensive approach in the differentiated categorization of facial expressions as various time-defined combinations of FACS AUs. On the one hand, according to the Investigator's Guide, the basic emotion of Happiness (or the so-called "Duchenne smile" as the marker of a genuine emotion expression) is defined as combination of AUs 6 and 12. On the other hand, an outwardly similar expression of the "social smile" is categorized as bilateral emergence of AU 12 in the form of a strict combination of both AU L12 and AU R12 in time, with AU 6 absent. The social smile may be linked in time with a voiced statement, or have an independent meaning as a communicative sign. Using various combinations of AUs 9, 10, 12, 14 in time, we also can single out yet another variant of outer expression of Happiness in the form of the so-called "coy smile". Figure 2 is also showing a screenshot displaying the outcomes of facial activity analysis done in the process of speech production. The intervals corresponding to the voiced speech are marked as AU 50 on one of the horizontal lines. Facial activity events, automatically categorized as FACS AUs, as well as basic emotions are marked on the other lines: (A) Social smile as a co-speech gesture; (B) Happiness expressed as genuine emotion.

Thus, the second problem was being solved with the help of a system of rules, distinguishing the facial activity patterns, specific for basic emotions expressions, from outwardly similar patterns that make a complex with the voiced statement. In these rules, the peculiarities of overlapping of certain AUs in time were taken into account. We would like to once again emphasize that:

(a) The complexity of the task lies in the fact that facial expressions of basic emotions are made up by the same set of AUs as communication and speech related facial activity;

(b) the latter are facial activity expressive instruments that accentuate and modify the voiced message (for review, see: [2], [6], [7]).

IV. Conclusion

We managed to develop the computer FACS software with the capacity to carry out automated analysis of facial expressions in videos, based on the principles of singling out the combinations of FACS AUs within the comprehensive approach implementation. The created rules of detection of the time-space facial activity patterns allow to conduct differentiated assessment of facial expressions in the course of speech production. The rules are determined by sophisticated analysis of time-based overlap of varied action units, while considering their individual meanings within a facial event formation, including applying the fuzzy logic principles. The angle of the approach we have designed, is related to the choice of the adequate facial expressions analysis unit, namely, FACS action units.

References

[1] L. Feldman Barrett et al., "Emotional Expressions Reconsidered: Challenges to Inferring Emotion from Human Facial Movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, 2019, pp. 1–68, https://doi.org/10.1177/1529100619832930.

[2] J. Bavelas, N. Chovil, "Some pragmatic functions of conversational facial gestures," *Gesture*, 17(1), 98–127, 2018.

[3] D. T. Cordaro et al., "Universals and Cultural Variations in 22 Emotional Expressions Across Five Cultures," *Emotion*, Advance online publication, June 12, 2017, http://dx.doi.org/10.1037/emo0000302.

[4] P. Ekman, "FACS: Yesterday and Today," in E. L. Rosenberg, P. Ekman, Eds., *What the Face Reveals Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, New York, USA: Oxford University Press, 2020, pp. 614-618.

[5] P. Ekman, W.V. Friesen, J.C. Hager, *Facial Action Coding System (FACS): The Manual & the Investigator's Guide*, Salt Lake City U, USA: A Human Face, 2002.

[6] A. Kendon, *Gesture: Visible action as utterance*, Cambridge, UK: Cambridge Univ. Press, 2010.

[7] D. McNeill, *Gesture and thought*, Chicago, USA: University of Chicago Press, 2005.

[8] E. L. Rosenberg, "FACS in the 21st Century," in E. L. Rosenberg, P. Ekman, Eds., *What the Face Reveals Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, New York, USA: Oxford University Press, 2020, pp. 1-22.

[9] E. L. Rosenberg, P. Ekman, Eds., *What the Face Reveals Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, New York, USA: Oxford University Press, 2020.

[10] K. R. Scherer et al., Dynamic Facial Expression of Emotion and Observer Inference, *Frontiers in Psychology*, vol. 10, 2019, https://doi.org/10.3389/fpsyg.2019.00508.

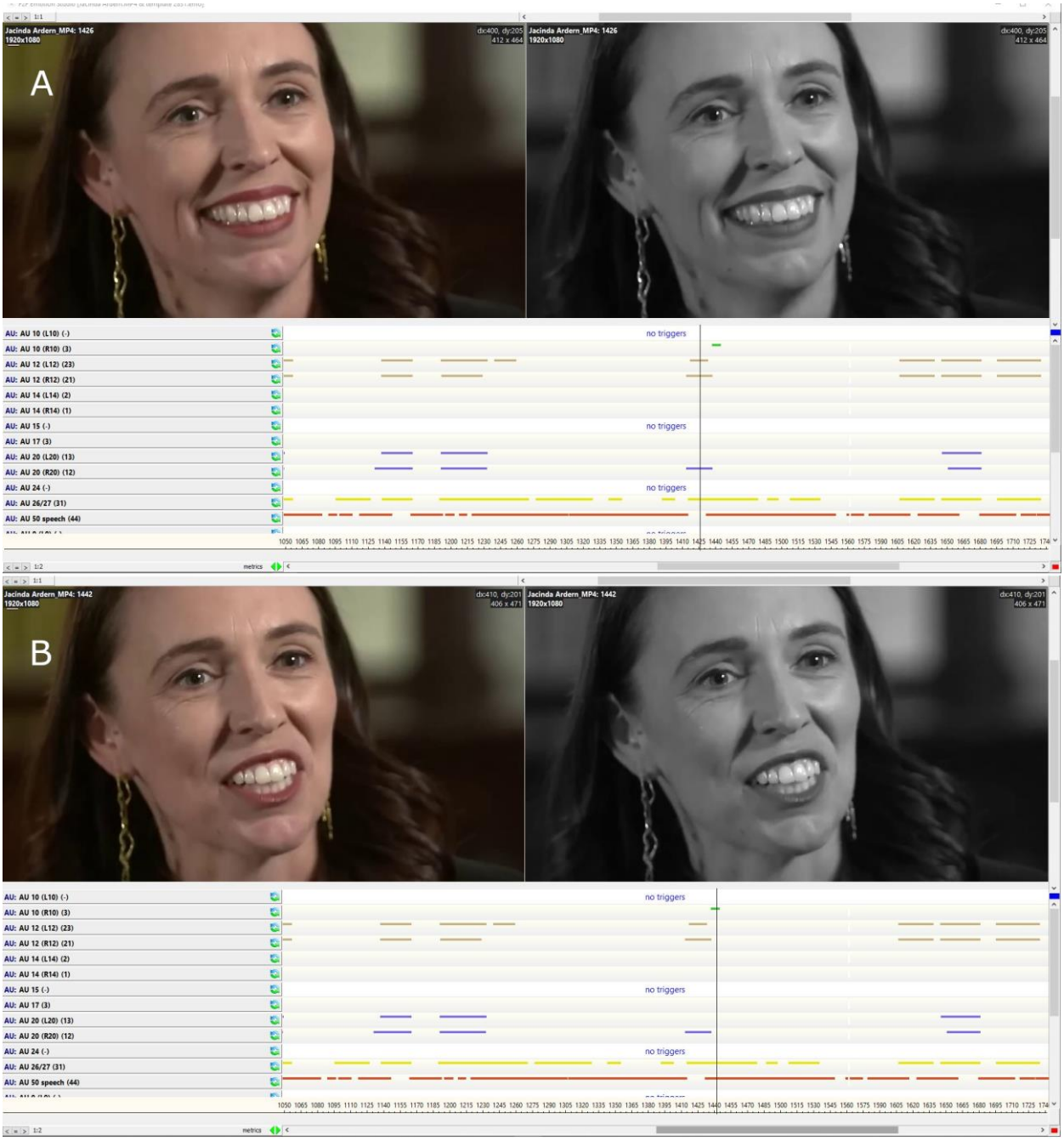

Fig. 1 EmoRadar software screenshot displaying the outcomes of the Jacinda Ardern's interview analysis. A: AU 12 and 20; B: AU 10. The line at the bottom shows frame numbers. The vertical line marks the corresponding frame.
https://www.youtube.com/watch?v=Kz2m7O3tfWo&t=37s

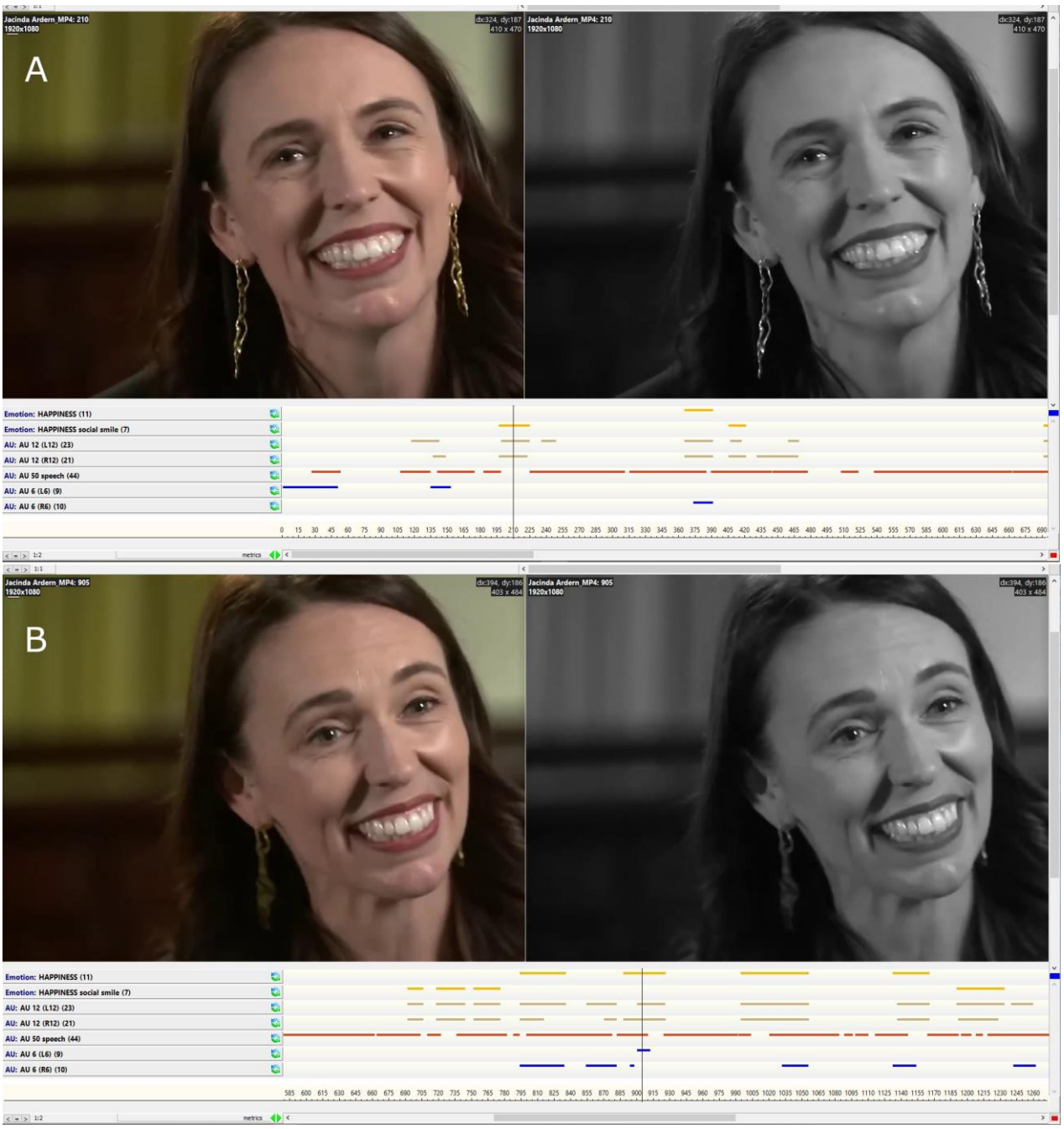

Fig. 2 EmoRadar software screenshot displaying the outcomes of the Jacinda Ardern's interview analysis. A: Social smile as a co-speech gesture; B: Happiness expressed as genuine emotion. The line at the bottom shows frame numbers. The vertical line marks the corresponding frame. https://www.youtube.com/watch?v=Kz2m7O3tfWo&t=37s