Comparing OpenFace to Manual Annotations of Communicative Facial Signals

Emma C. Vriezen

Student number: s1010487

Artificial Intelligence, Radboud University

Bachelor Thesis in Artificial Intelligence (SOW-BKI300)

First supervisor: Dr. J. Holler[1, 2]

Second supervisor: Dr. J. Kwisthout[1]

Daily supervisors: Dr. L. Drijvers[1, 2], Dr. J.P. Trujillo[1, 2]

June 30, 2020

[1] Radboud University; Donders Centre of Cognition; Donders Institute for Brain, Cognition and Behaviour, Nijmegen, the Netherlands
[2] Max Planck Institute of Psycholinguistics, Nijmegen, the Netherlands

Abstract

This thesis assesses the differences between the output of OpenFace and manual annotations of communicative and holistic facial signals. OpenFace is a software program that detects facial signals in videos of human faces as Action Units. These unit of facial movement are not always of interest for research. Human coders might only want to annotate communicative and holistic facial signals, instead of all visible signals. Video annotation is a time-consuming process to do manually, so automation is desired. This thesis explains how the output of OpenFace and annotations of communicative signals differ on conceptual level, goal, and features. These differences should be considered when using OpenFace for annotation of communicative and holistic facial signals. An attempt is made to transform the output of OpenFace into annotations of frowns, blinks, smiles, and gaze aversion by manually finding thresholds and constraints. A minimal agreement is reached between the transformed output and the manual annotations. The conclusion is that OpenFace can be used to automate the annotation of communicative facial signals, but only with the help of machine learning. Unbiased data is required for training, together with objective definitions of communicative facial signals.

Comparing OpenFace to Manual Annotations of Communicative Facial Signals

Face-to-face communication between humans is multimodal. Speech on itself is of the auditory modality. Other modalities are introduced as nonverbal cues, like facial expressions and gaze direction (Hecht & Ambady, 1999). These nonverbal cues add a layer of information on top of what is being said or done, for example evoking the feeling of being addressed (Nagels, Kircher, Steines, & Straube, 2015) or allowing the prediction of the end of a speech turn (Holler, Kendrick, & Levinson, 2018). Facial signals can lead to a quicker and better understanding of speech (Kelly, Özyürek, & Maris, 2010; van Wassenhove, Grant, & Poeppel, 2005). For example, a question might be easier recognised as such when it is preceded or accompanied with raised eyebrows. This faster processing of multimodal signals might be facilitated by the application of Gestalt-like principles. The combinations of different signals are interpreted holistically and can be more easily tied to a specific meaning than unimodal signals (Holler & Levinson, 2019). Statistical regularities between certain communicative actions, like asking a question or giving a response, and facial signals should be discovered to better understand human communication and the perception of facial signals (Ripperda, 2019).

The most straightforward way of discovering which facial expressions are used during conversation, is annotating recordings of face-to-face communication. This must be done by trained raters, to make the results as objective as possible. The expert raters only annotate the communicative facial signals. This means that signals that add extra information on top of speech are annotated, while facial signals that do not provide extra information or are caused by itches, twitches and other external factors are ignored. Both the training and the annotation procedure itself are time-consuming activities (Trujillo, Vaitonyte, Simanova, & Özyürek, 2019). Automatically annotating video data would reduce the costs and eliminate any residual subjectivity. There are several software packages available for automatic facial signal annotation. One of them is OpenFace. This open-source program is freely available for scientific purposes. OpenFace places facial landmarks on a recorded face and deduces the presence and intensity of facial signals by the displacement over time of these landmarks. However, OpenFace cannot simply replace manual coding, because it annotates every visible facial signal instead of only communicative signals. In practice, this leads to an over-abundance of facial signal annotations from OpenFace compared to the manually annotated communicative signals. OpenFace could still be used to automate the annotation of communicative facial signals if a transformation from the output of OpenFace into annotations conceptually similar to the manual coding is possible. To assess this possibility, it is required to understand how the annotations of OpenFace differ from manual coding.

This thesis will clarify these differences and assess how OpenFace and manual coding compare by answering the following research questions (RQs):

RQ1: What are the differences between the automatically generated annotations of facial signals by OpenFace and the manually annotated communicative facial signals?

RQ2: Can the output of OpenFace be transformed into annotations of communicative facial signals by applying manually picked thresholds and constraints?

There is no hypothesis for RQ1 because it is an exploratory question. For RQ2, it is hypothesised that transforming the output of OpenFace into communicative facial signal annotation by merely manually picked rules is not possible. The output of OpenFace is large and person dependent, so to manually pick general thresholds is a too complex task.

In *Background*, OpenFace and its output are described, followed by an overview of the manual communicative coding. *Comparison* contains an elaboration of the differences in working, output, and assumptions between OpenFace and manual coding, mostly relevant for *RQ1*. In *Methods*, the data and general procedure for altering the output of OpenFace to resemble more the manually annotated data are specified, followed by its *Results*. The research questions are answered in the *Discussion*, accompanied with considerations about automating the annotation of communicative and holistic facial signals. The thesis is finalised with the *Conclusion*, with ideas for further research.

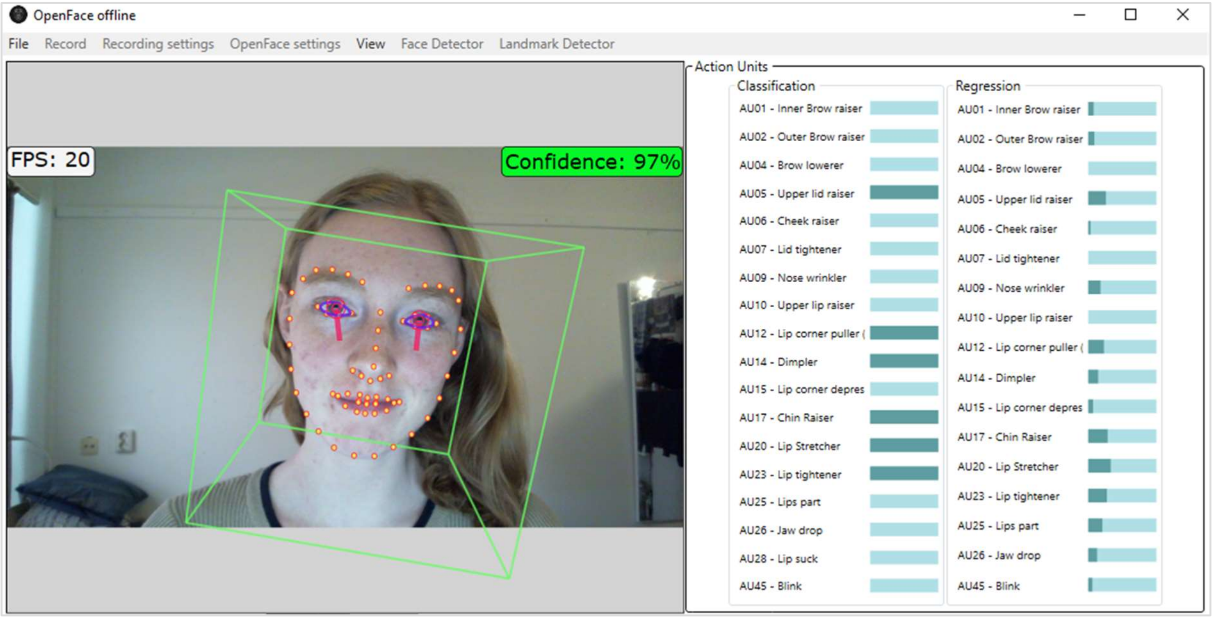
Background

OpenFace 2.0.5

OpenFace is a software toolkit that can be used for annotating information about a recorded face. It is "intended for computer vision and machine learning researchers, affective computing community and people interested in building interactive applications based on facial behavior analysis" (Baltrušaitis, Zadeh, Lim, & Morency, 2018). OpenFace can take a still, video, or webcam stream as input (Figure 1). For every frame, it tries to detect a face. If successful, 68 facial landmarks are superimposed on the face tracking the contours (orange dots) and eyes (purple dots). From these facial landmark estimations, the facial expressions are deduced. Their presence (left column) and intensity (right column) are annotated.
The coordinates of the head with respect to the camera are estimated as well (green box), just as the direction of the gaze (pink rods). These estimations are outputted in an CSV. Earlier

Figure 1

User interface of OpenFace



versions of OpenFace struggled with non-frontal or occluded faces, or with low illumination conditions. Version 2.0 has improved on these situations, by training on datasets with partly occluded faces and using a different neural network architecture.

The facial expressions are described in Action Units (AUs). Each AU is a different visually distinguishable facial movement. Their specifications are given in the manual of the Facial Action Coding System (FACS) (Ekman & Friesen, 1978). FACS was designed with the purpose to be useful in the detection of micro expressions (Ekman & Rosenberg, 2005), hence the smallest visible facial movements are annotated just like large, voluntary facial expressions, namely with the AUs. The intensity of an AU is given on a scale from A to E.

OpenFace predicts 18 different AUs, which is not extensive. These are the AUs annotated in the training data of OpenFace. Overlaps exist between the AU labels of the different training datasets, allowing them to be used for training together (Baltrušaitis, Mahmoud, & Robinson, 2015). The resulting AUs and the muscles that are correspondingly contracted are listed in Table A1, Appendix A. The AUs are derived from the difference in geometry and appearance of the tracked face from the neutral expression. The neutral expression is assumed to be the median of the geometry and appearance of a face.

OpenFace has been trained on two different kinds of datasets: one with AU presence annotations, and the other with AU intensity annotations. OpenFace contains two different models that are trained on these different data. A Support Vector Machine (SVM) annotates presence (0, 1) while a Support Vector Regressor (SVR) predicts the intensity on a scale from 0 (no intensity) to 5 (maximum intensity). The two different models are independent of each other. All 18 AUs are predicted by both models, except AU 28, because there were no data available for training its intensity prediction (Elebash, 2020). Hence, only its presence is annotated.

Manual coding of facial communicative signals

Facial communicative signals consist of facial expressions and eye gaze (Lang, Wachsmuth, Hanheide, & Wersing, 2012). Not all noticeable facial signals contribute to face-to-face dialogue. Movements in the face caused by twitches and scratching have often no communicative value, nor expressions that are evoked by external sources, unrelated to the conversation. For example, moving the gaze in the direction of an unexpected sound is not communicative unlike averting gaze in a thoughtful manner. The facial movements purely resulting from speech do also not bear any communicative value on top of the meaning of the speech itself. Of course, these movements can be minimised or exaggerated to convey additional information, for example during smiling. In such a case, the movement becomes communicative.

The manually coded annotations used in this thesis only describe communicative facial signals. The signals that the human annotators code for are listed in Table A2, Appendix A. To solely annotate communicative signals, more information is used than purely the visible movements, for example the context of the conversation or attributed intentions.

Next to looking at communicativeness, the manual coding also codes holistic facial expressions. Holler and Levinson (2019) define holistic perception of messages as the integrated meaning of different perceived stimuli, possibly from different modalities. The resulting interpreted message has a meaning larger than the sum of its components. For example, a combination of raising the cheeks, squinting the eyes, and stretching the lips can be holistically interpreted as a smile.

Automating the annotation of communicative and holistic signals with OpenFace is a challenge, because OpenFace makes no distinction between non-communicative and communicative signals and only the components of a holistic signal are annotated, instead of their combination.

The presence annotations from OpenFace and the annotations resulting from manual coding for the same time frame and video can be seen in Figure 2 and 3. These images are screenshots from ELAN (“ELAN (Version 5.5) [Computer software],” 2019).

Figure 2

Action Unit presence annotation from OpenFace

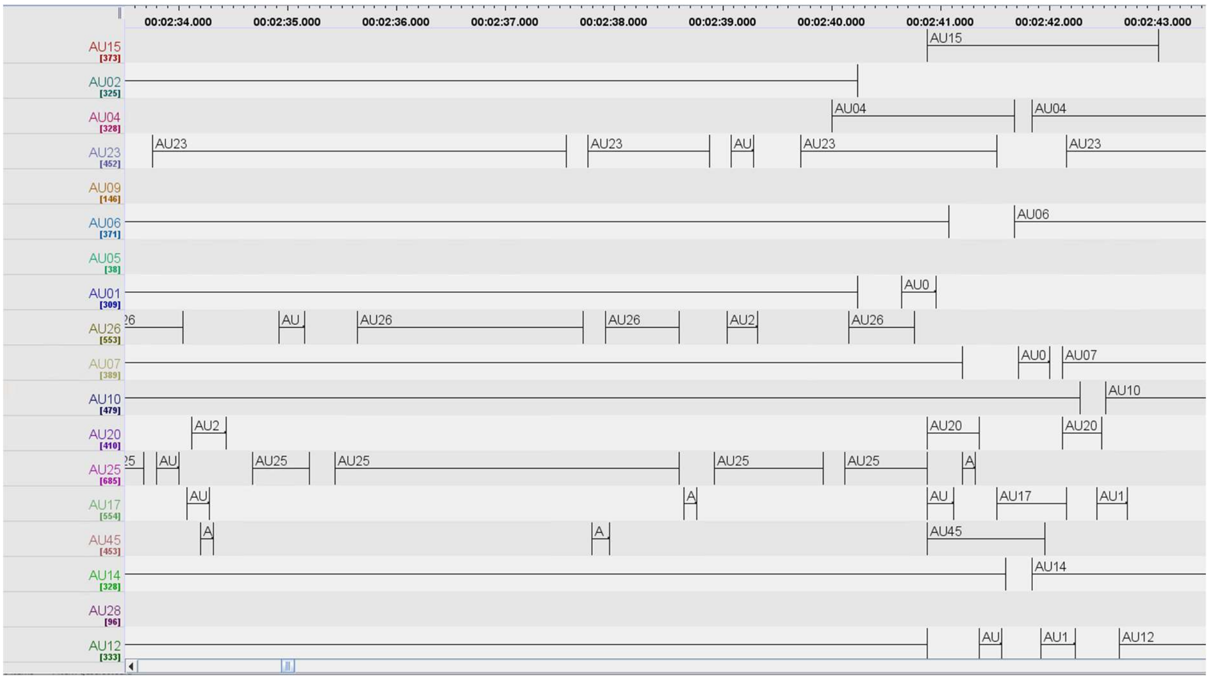
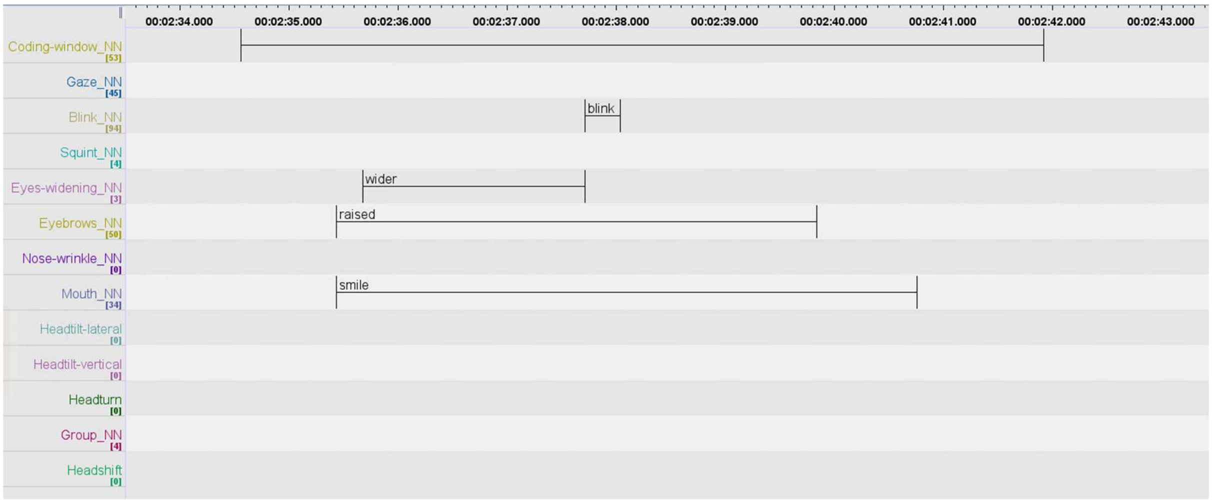


Figure 3

Manual annotations of communicative and holistic facial signals



Comparison

OpenFace has been designed with a different goal in mind than the human manual coders have, so a simple transformation from the output of OpenFace to annotations comparable to the manual annotations is not possible. In this section, the differences between OpenFace and manual coding are listed. Knowing what the differences are is required for assessing whether, to which extent, and how using OpenFace for automating the coding of communicative signals is possible.

Conceptual Differences

OpenFace annotates facial movements with Action Units (AUs), which are part of the Facial Action Coding System (FACS) (Table A1, Appendix A). Every AU stands for a specific visually discernible movement on the face. Hence, the intention of OpenFace is to annotate every movement that it detects. Which movements it detects or fails to detect will be discussed later.

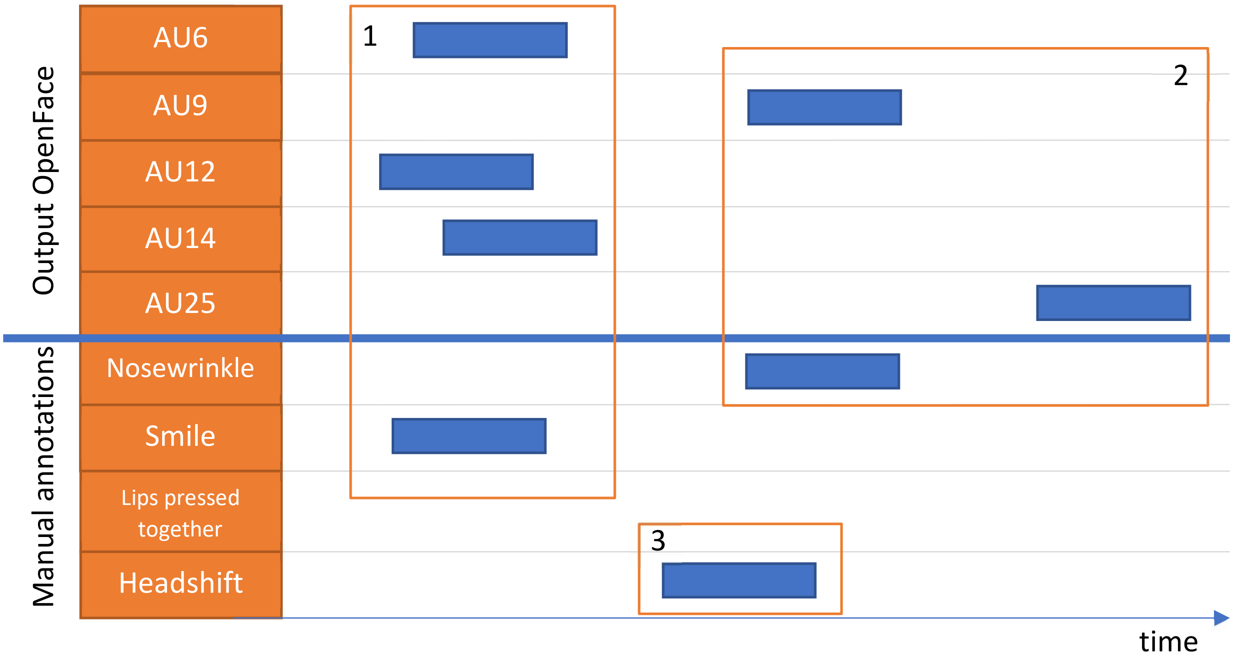
Manual coding takes place on another conceptual level. Namely, not all atomic movements that are defined in the FACS are coded separately. Instead, the more general, higher level movements are considered (see Table A2, Appendix A). The goal of manual coding is also different from the goal of OpenFace. Namely, only the movements that add meaning to what the speaker is communicating to an addressee are annotated, opposed to annotating every perceivable movement. The annotation is done from the perspective of an addressee. Only what this addressee is assumed to perceive is annotated.

Besides the level of movement and the goal of the annotations, OpenFace and manual coding also differ in the features they can annotate. OpenFace does not distinct between unilateral or bilateral movements, while the manual coding does. Unilateral facial signals can bear different meaning form their bilateral counterparts, for example raising one eyebrow can express disapprovement or judging something questionable, while raising both brows can show surprise. Also, while OpenFace does track the position of the head in three-dimensional space and estimates the direction of the eye gaze of the tracked face, it does not annotate any head movements or gaze aversions. Manual coding does however allow for their annotation, of course only when communicative.

The differences between OpenFace and manual coding in level, goal and features are visualised in Figure 4.

Figure 4

Schematic visualisation of the differences in coding between OpenFace and manual coding



Note. A blue bar represents an annotation. 1. Difference in level: OpenFace codes a larger facial expression (i.e. smile) as separate Action Units, while manual coding makes one holistic annotation for it. 2. Difference in goal: If a facial signal is communicative, both OpenFace and the manual coding annotate it (AU9 and nose wrinkle). When a signal is not communicative, only OpenFace annotates it (AU25 and no annotation for ‘Lips pressed together’). 3. Difference in features: next to facial expressions, manual coding annotates other communicative signals, for example head shifts. OpenFace does not have AUs that describe this feature.

Overcoming the Differences

To assess whether the annotations of OpenFace can be used to automate coding of communicative and holistic signals, their differences in coding level, goal and features should be overcome.

Coding Level. The low-level atomic movements annotated by OpenFace can be combined to higher level, as-perceived-by-human annotations. This should be done based on the relations between the Action Units and holistic codes. These relations can be found in several ways:

Obvious Relations. Action Units describing the movement of a specific region of the face have a relation to a more holistic code of the same region. For example, AU1, 2 and 4 each describe eyebrow movements. Hence, they have a relation with the general holistic tier 'Eyebrow'. Meanwhile, AU1 and 2 code for respectively inner and outer brow raises, and hence bear a relationship with the higher-level codes that stand for eyebrow raises ('raised', 'unilateral raised', etc.).

Anatomic Relations. The FACS describe which facial muscles are contracted during which Action Unit (Farnsworth, 2019). Meanwhile, the criteria of the holistic codes allow for determining which facial muscles are involved by which holistic movement. A relation between an AU and a holistic code exist when they involve the same muscle contractions. For example, the manual code 'Squint' has contraction of one or both eyelids as a criterium. AU44 also codes for a squint and uses the muscle orbicularis oculi pars palpebralis, but OpenFace does not annotate this action unit. It does however annotate AU6 and AU7 which use the same muscle as AU44. Hence, it there might be a relation between the squint and AU6 and AU7, respectively the cheek raiser and lid tightner.

Statistical Relation. OpenFace uses two models to generate the annotation of Action Units; an SVR regression model (AU_r) for intensity and an SVM classification model

(AU_c) for presence, as explained in *Background*. Both the models are imperfect and make mistakes, as will be discussed further below. Simply combining the by AU_c annotated presence of AUs and renaming them to the holistic tiers is bound to result in many false positives and false negatives. More complex criteria might however allow for a transformation from the automatic OpenFace output to holistic annotations. It is possible that a tier of holistic annotation corresponds to a specific pattern in the output of OpenFace, for example a combination of minimal activations of AU_r (thresholds) together with the presence or absence of certain AUs according to AU_c.

Coding Goal. While OpenFace is trained to detect all discernible facial movements, manual coding merely annotates the perceivable facial movements that have a communicative value, which means that the movement adds value to what the speaker is saying. To transform the output of OpenFace to communicative annotations, detected Action Units that are not part of a communicative facial signal should thus be discarded. Two things that could be done to at least remove a part of the non-communicative annotations are:

Remove Speech Artefacts. Predominantly AU25 and AU26, but also other Action Units, are present during vocal speech. From the output of OpenFace, the frames in which the recorded participant is speaking can potentially be filtered out by a trained algorithm, by using AU_r or the facial landmarks around the mouth. These frames should then not be completely disregarded; during speech, communicative signals can still be produced, also with the mouth (think of imitating another person or smiling while talking). Special criteria could be set for the frames in which the participant is speaking, for example higher thresholds for the output of AU_r related to Action Units around the mouth.

Threshold Minimum AU Duration. The length of Action Units can be given a minimum duration threshold. This leads to that very briefly detected Action Units can be disregarded and not be transformed into communicative annotations. Such thresholds could be based on minimum length of communicative facial signals. The pitfalls here are that OpenFace might fail to annotate the complete duration of an Action Unit, or that a minimum duration for a communicative facial signal does simply not exist.

Coded Features. Some of the manual annotations, like Eyebrow movements, can be unilateral (i.e. raising one eyebrow as opposed to both). OpenFace does not distinct between unilateral or bilateral occurrence of AUs. This could be overcome by using the facial landmark position output. For every supposed movement that could be unilateral, the movement of the relevant landmarks can be compared between the left and right side of the face. If the displacement of one side is small enough compared to the other side that it is neglectable, then the movement can be considered unilateral. Otherwise, it is bilateral.

Head movements and gaze aversion are also annotated with manual coding. The FACS contains action units for these movements, but the models of OpenFace are not trained to detect them. Both head and gaze movements should be hidden in the positional part of the output of OpenFace, however. In *Methods*, it is described how gaze aversion annotations have been extracted from the output of OpenFace.

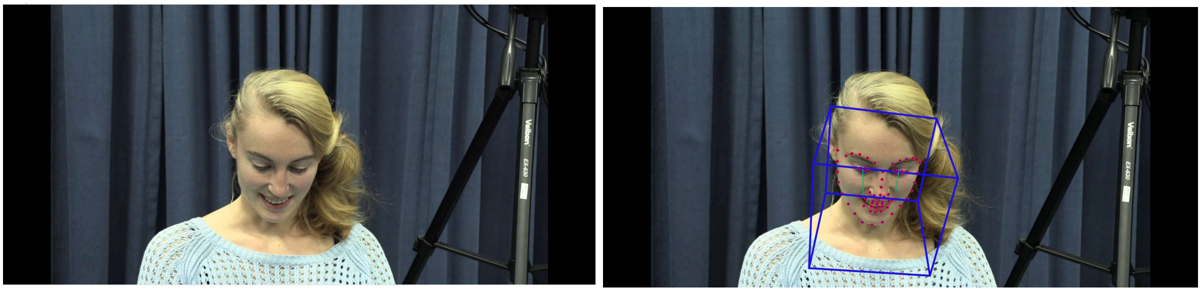
Flaws of OpenFace

So far, the ideas proposed to transform the output of OpenFace into annotations of holistic and communicative facial signals assumed that the output is correct. Its output is however far from perfect. Baltrušaitis et al (2018) report that the Action Unit intensity prediction of OpenFace often outperforms other AU detection methods. The correlation coefficients reported are from a comparison with the DISFA dataset. This dataset features recordings of participants watching a 4-minute long video with fragments specially chosen to evoke spontaneous facial expressions (Mavadati, Mahoor, Bartlett, Trinh, & Cohn, 2013). This is a more stable situation than a free face-to-face conversation, because the participants fixate on the screen and do not speak. Even though OpenFace has been trained on data featuring speech as well, the question remains how well the reported correlations hold for conversational videos.

When OpenFace annotates a signal wrongly, it is either a false positive (FP) or a false negative (FN). For FPs, an Action Units presence is detected while the respective movement does not occur in the input. For FNs, OpenFace fails to detect an Action Unit visible in the input. A direct cause for either FPs or FNs is difficult to pin down from the output alone. Both AU presence and intensity models take the landmark positions, normalised for the current person, as input. These landmarks are the output of a neural network preceding in the OpenFace pipeline and hence not the objective truth. If the landmarks are placed wrongly, then the AU annotations are wrong as well. This is one possible cause of both FPs and FNs. Another cause linked to the facial landmarks is not their wrong placement, but a warped view of the face. For example, when somebody looks downward, the eyes might not be visible anymore and the landmarks above and below the eye might be practically overlapping (Figure 5). This is then interpreted as a blink, while in fact the participant does not blink, or it is impossible to decide from the footage whether there is a blink or not. So, occlusions by moving the head or blocking vision with a limb or an object can result in wrong annotations. Theoretically, these mistakes can be filtered out by detecting when the head is turned away from the camera, by the coordinates of the facial landmarks. In *Methods*, an attempt to reduce the amount of falsely annotated blinks by using the output from both models is described.

Figure 5

Frame from a video which OpenFace annotates as blink, but is actually gaze aversion



To transform the output of OpenFace to holistic and communicative signal annotations, both the classification and regression model for AUs should be used. Sometimes one model fails to detect an AU while the other picks up on it. False positives can partially be filtered out as well, by ignoring all positive presence annotations when the intensity is too low. What 'low' is, has to be determined for each AU separately, because the intensity outputted by OpenFace does differ between the AUs. Some AUs can have an intensity above 0, even when the AU is not present on the face. OpenFace normalises the AU intensity for each specific video by subtracting the n_{th} percentile (which is nowhere defined specifically) of intensity from all output. This correction might not be effective if the dynamic range of an AU is small (Baltrušaitis, 2019). Other AUs are rarely ascribed an intensity other than 0. Any sort of thresholding should be person dependent as well, because the extent to which some AUs are annotated can vary a lot between persons. For example, for some faces, AU4 (brow lowerer) is detected in a great part of the recording. For others, AU4 is rarely annotated. This inter-person difference makes simple thresholding of intensity to filter out false positives difficult.

It is also possible that both the presence and intensity model fail to detect a visible facial movement. False negatives like these cannot be accounted for, because they cannot be distinguished from a neutral face from the output alone.

When OpenFace detects a visible movement, it can still make mistakes in determining the onset and offset of the respective AU. These mistakes are still undesirable, because a human would still have to check every resulting annotation for the correct placement in time. Too long annotations have a too early onset or too late offset. Both the classification and regression model can annotate too long annotations. Both cases could be improved by looking at the rate of change of the intensity if AU_r has outputted anything different than 0. The onset of an AU is marked by an increase of the intensity. Then the AU can last for an arbitrary amount of time, with the ending marked with a decrease in intensity back to the level before the onset. The rise and fall in intensity can be detected by thresholding its rate of change. This can be used to automatically adjust the begin and end points of the annotations. An application of this idea is described in *Methods*.

Short annotations can be considered messy in comparison to the manual annotations. However, it should be considered that the goal of OpenFace is to detect every single facial movement, even the ones that a human would not consciously notice in face-to-face communication. This is why the abundance of annotations is not inherently bad. From the perspective of the purpose of OpenFace, it might be more useful to overannotate than to risk having false negatives.

Methods

To see whether a transformation from the output of OpenFace to annotations of communicative and holistic signals was possible, it was tried to clean up AU4 (brow lowerer) and AU45 (blink), next to detecting smiles and gaze aversion.

Dataset

The manual data used for assessing the transformation were 63 videos of 24 participants (16 females, 8 males) involved in 3 different tasks. The videos were around 20 minutes long each and recorded at 25 frames per second. They were manually annotated on the occurrence of a question or a response, which yielded 2434 annotation windows with an average length of 6.80 seconds (standard deviation of 2.99 seconds). Total duration of the windows together is 35 minutes and 58.6 seconds. The manual annotations and the output of OpenFace had been cropped to these coding windows, to prevent that values outside of the windows influence the outcome of any comparison. The data are not publicly available.

Before altering the data, the AU intensity and presence for all frames with a confidence lower than 0.97 were removed.

Approach

Cleaning up AU4 was a demonstration of a simple threshold. AU4 was chosen for this demonstration, because manual inspection showed that AU4 was annotated as ‘present’ too frequently, while its intensity was usually above 0. A threshold for 2.5 had been chosen: AU4 is annotated only when its intensity (regression model of OpenFace AU_r) was 2.5 or higher. The output of the presence model (classification model of OpenFace AU_c) was ignored. The resulting annotations were merged into one if they were 1 or 2 frames between them. The remaining annotations of 1 or 2 frames long were dropped. The final AU4 annotations were compared to the manual annotations ‘frown’, ‘frown-raised’ and ‘unilateral-frown-raised’.

Changing the annotations of AU45 was done by using the rate of change of the intensity output. AU45 was often annotated too long by OpenFace, for example when a participant was looking downward, causing the eyes to be hidden from the camera by the eyelid. At those instances, it was often still possible to visually discern blinks from the video recording, because movement of the eyelid and around the eye is still recorded. This might be reflected in the outputted intensity of AU45 by AU_r. The value of the rate of change for a frame *t* is the slope of the regression line of the points *t-2*, *t-1* and *t*, calculated with the function *linregress* from (“scipy.stats.linregress,” 2019). A blink would finally be annotated if frame had a rate of change above 0.3, followed within 10 frames by a rate of change smaller than -0.3 with two frames later higher than -0.2. The resulting AU45 annotations were compared with the manual annotation ‘blink’.

Smiles were not explicitly annotated by OpenFace, while it is part of the manual annotations. For the extraction of smiles, the intensity outputted for AU6 (cheek raiser), AU12 (lip corner puller) and AU14 (dimpler) were used, because a rise in intensity for these three AUs was noticed on several manually checked smiles. The mean and standard deviation for all three AU intensities within one recording were calculated. The start of a smile annotation was when the intensity of all three AUs is more than their respective standard deviation higher than their respective mean. An annotation ended when one of the intensities of one of the three AUs got below its mean instead of below the initial threshold, because the intensity of a smile could wear off towards the end and should not be ended too early. To smooth out the start of the annotations, the onset was moved to the first preceding frame where all three AUs are above their mean intensity.

Just like smiles, gaze aversion was not annotated by OpenFace, though an estimate of the direction of gaze was part of the output. The mean and standard deviation of the *x* and *y* gaze angle were calculated for the output of OpenFace for one recording. A frame was annotated as ‘gaze aversion’ when the *x* or *y* gaze angle differed respectively 1.3 and 1.5 of their standard deviation from their mean. Resulting annotations of less than 4 frames (< 0.16 s) were dropped. The remaining annotations were concatenated when separated by only a blink (AU45) or when less than 11 frames apart from another, because that was the maximum length for a blink in the altered blink annotations.

Any resulting annotation shorter than 3 frames (< 0.12 s) was dropped.

Reliability evaluation

Cohen’s kappa κ measures interrater reliability corrected for chance agreement, i.e. κ “is the proportion of agreement *after* chance agreement is removed from consideration” (Cohen, 1960). It was obtained for each of the four annotation types and their manual counterparts with the use of ELAN 5.5 (Hellwig et al., 2020). This statistic was also calculated for the original presence output of OpenFace for AU4 and AU45. ELAN 5.5 implemented easyDIAg (Holle & Rein, 2015) for the calculation. Annotations were marked as overlapping when at least 60% of the longest from the pair coincided with the other annotation. The resulting values of κ were interpreted as a degree of agreement (McHugh, 2012).

Results

AU4 and Frowning

The global agreement matrices created by ELAN are shown in Table B1, Appendix B.

There is no global agreement between the presence annotations of OpenFace AU4_c and the manual frowning annotations ($\kappa = 0.0175$, $\kappa_{max} = 0.3122$, raw agreement (RA) = 0.1345). For frown presence, $\kappa = 0$ ($\kappa_{max} = 0.0995$, RA = 0.5659), while for frown absence $\kappa = 0$ ($\kappa_{max} = 0.7123$, RA = 0.4730).

After thresholding of the AU4 intensity in the OpenFace output, it has a minimal global agreement with the manual annotations ($\kappa = 0.2730$, $\kappa_{max} = 0.8300$, RA = 0.6840). For frown presence, $\kappa = 0.0737$ ($\kappa_{max} = 0.7835$, RA = 0.8562), while for frown absence $\kappa = 0.5149$ ($\kappa_{max} = 0.9092$, RA = 0.8228).

AU45 and Blinking

The global agreement matrices created by ELAN are shown in Table B2, Appendix B.

There is initially also no global agreement between the presence annotations of OpenFace AU45_c and the manual blinking annotations ($\kappa = 0.1531$, $\kappa_{max} = 0.8365$, RA = 0.3037). For blink presence, $\kappa = 0.0452$ ($\kappa_{max} = 0.7608$, RA = 0.6228), while for blink absence $\kappa = 0.2608$ ($\kappa_{max} = 0.9710$, RA = 0.6438).

After using the rate of change of the AU45 intensity in the OpenFace output, it has a minimal global agreement with the manual annotations ($\kappa = 0.3329$, $\kappa_{max} = 0.9421$, RA = 0.5141). For blink presence, $\kappa = 0.4518$ ($\kappa_{max} = 0.9640$, RA = 0.7742), while for blink absence $\kappa = 0.4783$ ($\kappa_{max} = 0.9616$, RA = 0.7402).

AU6, 12, 14 and Smiles

The global agreement matrices created by ELAN are shown in Table B3, Appendix B.

The smiles extracted from the intensity of AU6, AU12 and AU14 in the OpenFace output and the manual smile annotations have a global interrater reliability of $\kappa = 0.2058$ (κ_{max}

= 0.5487, RA = 0.4101), which stands for a minimal agreement. For smile presence, κ = 0.1778 (κ_{max} = 0.4503, RA = 0.6996), while for smile absence κ = 0.3532 (κ_{max} = 0.8104, RA = 0.6737).

Eye Gaze Estimation and Gaze Aversion

The global agreement matrices created by ELAN are shown in Table B4, Appendix B.

The gaze aversion annotations extracted from the gaze angle in the OpenFace output and the manual gaze aversion annotations have a global interrater reliability of κ = 0.1750 (κ_{max} = 0.8525, RA = 0.3685), which stand for no agreement. For gaze aversion, κ = 0.0616 (κ_{max} = 0.7688, RA = 0.6587), while for no gaze aversion κ = 0.3114 (κ_{max} = 0.9796, RA = 0.6565).

Discussion

RQ1, "What are the differences between the output of OpenFace and the manual codes", is answered by the contents of *Comparison*. The difference between the annotations outputted by OpenFace and the manual annotations are the conceptual level of the used codes, the coding goal, and the coded features. There are no straight-forward similarities between OpenFace and manual coding, except the annotation of blinks. More complex relations between the output of OpenFace and the manual annotations could be discovered with enough ground truth manual data.

The results from the attempt to transform the output of OpenFace show that the automatic annotations can be improved with respect to the manual codes, positively answering *RQ2, "Can the output of OpenFace be transformed into manual annotations by manually applying thresholds and constraints?"*. Both blink and frown annotations are improved from no agreement between the output of OpenFace and manual annotations to a minimal agreement. For only the positive annotations of frowns, the agreement after the transformation is still negligible with κ = 0.0737. However, the maximum theoretical value of κ has increased from $\kappa_{max} = 0.0995$ to $\kappa_{max} = 0.7835$. This means that an agreement on positive frown annotations between the original OpenFace output and the manual annotations was not possible, due to the differences in their marginal distributions. In this specific case, OpenFace made 4915 positive annotations of frowns, while the manual annotators had only detected 405 frowns. After the transformation of the automatic output, only 648 automatic positive frown annotations remained. The distribution of positive frown annotations over the videos had become more similar to manual coding.

Smiles were extracted with a minimal agreement as well. The extracted gaze aversion annotations have no agreement to the manual annotations.

The shown improvements refute the hypothesis that transforming the output of OpenFace into communicative and holistic annotations is impossible. However, the improvements are not great enough to automate their coding with OpenFace. These results were obtained by manually inspecting the corpus of data and picking thresholds that yielded promising results on single, handpicked instances. This is not the way to go when automating the annotation of communicative signals. Machine learning should be used to find patterns between the output of OpenFace and the communicative, holistic annotations.

An artificial neural network (ANN) could look at the output of OpenFace for a window of frames simultaneously and output the likeliness for any of the manual tiers. Recurrent neural networks would especially be suitable. These ANNs can take time-series as input and store relevant information in their memory, to use in its analysis of future input. Input can also be of arbitrary length. This kind of learning is called supervised learning and requires ground truth target data (i.e. manual annotations) to learn the weights of the ANN. Without ground truth data, unsupervised learning algorithms are an option. These algorithms find patterns in the data they are trained on, and cluster the data. The resulting clusters should then be assessed manually to see if any of the clusters approximate any of the manual tiers.

Machine learning is not used in this thesis, because of the lack of unbiased ground truth data. The manual annotations used to obtain the results presented in this thesis only described questions and responses, which might contain biases about which facial expressions occur together with other expressions (i.e. during questions and responses). Using these for training an algorithm that has to recognise communicative signals from any conversational video input could lead to a biased algorithm.

With machine learning, one has to be careful to prevent over- or underfitting. The different manual tiers might be unbalanced, with some tiers occurring more frequently than others. Sparsity of (certain tiers of) annotations might make naïve supervised learning difficult. Since the pipeline of OpenFace also contains neural networks, its output is not ensured to be correct. In the case that OpenFace fails to detect Action Units, no relation will be found between its output and the manual annotations. However, even when OpenFace is imperfect, a transformation from its output to holistic tiers is theoretically possible. Training could also be performed on the positions of the facial landmarks detected by OpenFace. This would eliminate the problems caused by possible mistakes by OpenFace in its Action Unit recognition.

Another point of discussion is the degree to which the manual annotations can be regarded as 'truth'. The smile extraction presented in this thesis resulted in an overabundance of positive smile annotations, compared to the holistically annotated smiles. It could be that these extra annotations are simply false, but it could also be possible that faint smiles are really visible. The human annotator could have not considered it a smile enough, or not communicative enough, to make an annotation of it. It is important to set clear rules about what to annotate as a smile and what not. If somebody is happily telling a story and showing signs of smiling all the time, should the total duration be annotated as smiling? Or should only the moments on which the smile gets stronger than average be annotated? These kinds of considerations should be made for any communicative signal to eliminate subjectivity as much as possible. Several videos should all be coded by different human coders. These results should be compared to assess the interrater reliability between the human coders, and to discover if subjectivity is playing a part. Even with training data which have been accepted as true, it might be impossible to extract communicative facial signal annotations from the output of OpenFace without an objective definition of communicative signals. Furthermore, OpenFace generates its output based only on video data. To determine communicativeness, maybe more than only the visual domain is required. Possibly, the context of the ongoing dialogue is essential information. If this is the case, then OpenFace can never be used to fully automate the annotation of communicative facial signals with solely its current input.

The methods in this thesis did not address the problems caused by occlusions, either by averting the face away from the camera, or by blocking it with a limb or object, for any facial signal other than blinking. Problems caused by moving the face could be overcome by detecting the pitch, yaw and roll of the head. It should be possible to do this from the output of OpenFace, especially after normalising the 3D facial landmark coordinates with respect to the distance to the camera. The proportions between the landmarks on opposite sides of the face are indicative of the pitch, yaw and roll of the head. Apart from directly being applicable for the annotation of head movements, knowing how the head is moved can be used to filter out mistakes in Action Unit or other signal recognition.

If the recorded participant has the freedom to reposition themselves in their chair or to walk around, it should be taken into consideration that not all changes is head location are necessarily head movements. This might not be discernible anymore from merely the output of OpenFace. The 3D output of OpenFace is in respect to the camera. So, if somebody moves their body closer to the camera, and thereby their head too, without shifting their head forward, it can be seen in the estimated coordinates that the face has moved closer to the camera. When the participant does not move their body, but shifts their head forward, their face also appears closer to the camera. From merely the output, a distinction between these two situations cannot be made. It should however be possible to extract the movement up to a certain extent.

Automating the recognition of communicative signals is worthwhile. It would overcome the subjectivity between different raters, is faster and eliminates the time and financial means required to train the raters. Research after communicative signals is required in order to understand the underlying cognitive mechanisms of holistic perception and to discover the relations between different speech acts, facial signals, and emotions. Increased understanding of these topics is useful in learning about social understanding of people with autism or other disorders. Also, in the field of human-computer interaction, digital faces can be improved when supporting their synthesised speech with communicative facial signals, while improving the understanding of human input to digital systems.

Conclusion

The differences between the output of OpenFace and the manual codes (*RQ1*) are the conceptual level of the used codes, the coding goal, and the coded features. *RQ2*, *"Can the output of OpenFace be transformed into manual annotations by manually applying thresholds and constraints?",* does not receive a binary answer in this thesis. It is demonstrated that the output of OpenFace can be transformed into a format more similar to manual codes of communicative and holistic facial signals. However, the improvement is minimal. The output of OpenFace contains so many variables, that it is impossible for a human to find relations between the output and manual coding that generalise to all videos in the dataset. Since OpenFace does not take context into account for its result, it remains the question whether it is every possible to separate the communicative annotations from the non-communicative output. More precise understanding of what makes a facial signal communicative or not is essential in the process of automating the annotation of communicative signals.

It is recommended to continue the automation with machine learning techniques (e.g. recurrent neural networks). The data used in this thesis were not used for training a model that can be used for videos of conversations in general, because the data were specific to windows of questions and responses in the video and are hence biased. In future research, a model could be made from this data particularly for question and responses video excerpts. If a more general model is desired, less biased training data should be collected first.

References

Baltrušaitis, T. (2019). OpenFace Wiki: Action Units. Retrieved June 27, 2020, from https://github.com/TadasBaltrusaitis/OpenFace/wiki/Action-Units

Baltrušaitis, T., Mahmoud, M., & Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (Vol. 06, pp. 1–6). IEEE. https://doi.org/10.1109/FG.2015.7284869

Baltrušaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (pp. 59–66). IEEE. https://doi.org/10.1109/FG.2018.00019

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System*. Palo Alto: Consulting Psychologists Press.

Ekman, P., & Rosenberg, E. L. (2005). *What the Face RevealsBasic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)* (2nd ed.). New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195179644.001.0001

ELAN (Version 5.5) [Computer software]. (2019). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Elebash, D. (2020). AU28_r is missing in csv output file. Retrieved June 24, 2020, from https://github.com/TadasBaltrusaitis/OpenFace/issues/885

Farnsworth, B. (2019). Facial Action Coding System (FACS) - A Visual Guidebook.

Retrieved June 3, 2020, from https://imotions.com/blog/facial-action-coding-system/

Hecht, M. A., & Ambady, N. (1999). Nonverbal communication and psychology: Past and future. *New Jersey Journal of Communication*, *7*(2), 156–170. https://doi.org/10.1080/15456879909367364

Hellwig, B., Van Uytvanck, D., Hulsbosch, M., Somasundaram, A., Tacchetti, M., & Geerts, J. (2020). 1.9.2. Editing multiple files and analysis of multiple files. Retrieved June 3, 2020, from https://www.mpi.nl/corpus/html/elan/ch01s09s02.html

Holle, H., & Rein, R. (2015). EasyDIAg: A tool for easy determination of interrater agreement. *Behavior Research Methods*, *47*(3), 837–847. https://doi.org/10.3758/s13428-014-0506-7

Holler, J., Kendrick, K. H., & Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic Bulletin & Review*, *25*(5), 1900–1908. https://doi.org/10.3758/s13423-017-1363-z

Holler, J., & Levinson, S. C. (2019). Multimodal Language Processing in Human Communication. *Trends in Cognitive Sciences*, *23*(8), 639–652. https://doi.org/10.1016/j.tics.2019.05.006

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two Sides of the Same Coin. *Psychological Science*, *21*(2), 260–267. https://doi.org/10.1177/0956797609357327

Lang, C., Wachsmuth, S., Hanheide, M., & Wersing, H. (2012). Facial Communicative Signals. *International Journal of Social Robotics*, *4*(3), 249–262. https://doi.org/10.1007/s12369-012-0145-z

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., & Cohn, J. F. (2013). DISFA: A Spontaneous Facial Action Intensity Database. *IEEE Transactions on Affective Computing*, *4*(2), 151–160. https://doi.org/10.1109/T-AFFC.2013.4

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Nagels, A., Kircher, T., Steines, M., & Straube, B. (2015). Feeling addressed! The role of body orientation and co-speech gesture in social communication. *Human Brain Mapping*, *36*(5), 1925–1936. https://doi.org/10.1002/hbm.22746

Ripperda, J. (2019). *The Communicative Face*. Radboud University.

scipy.stats.linregress. (2019). Retrieved June 3, 2020, from http://docs.scipy.org/doc/scipy-0.16.0/reference/generated/scipy.optimize.fminbound.html

Trujillo, J. P., Vaitonyte, J., Simanova, I., & Özyürek, A. (2019). Toward the markerless and automatic analysis of kinematic features: A toolkit for gesture and movement research. *Behavior Research Methods*, *51*(2), 769–777. https://doi.org/10.3758/s13428-018-1086-8

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, *102*(4), 1181–1186. https://doi.org/10.1073/pnas.0408949102

Appendix A

Table A1

AUs predicted by OpenFace and the corresponding muscles

AU	Description	Facial muscle
1	Inner Brow Raiser	Frontalis pars medialis
2	Outer Brow Raiser	Frontalis pars lateralis
4	Brow Lowerer	Corrugator supercilii, Depressor supercilii
5	Upper Lid Raiser	Levator palpebrae superioris
6	Cheek Raiser	Orbicularis oculi pars palpebralis
7	Lid Tightner	Orbicularis oculi pars palpebralis
9	Nose Wrinkler	Levator labii superioris alaquae nasi
10	Upper Lip Raiser	Levator labii superioris
12	Lip Corner Puller	Zygomaticus major
14	Dimpler	Buccinator
15	Lip Corner Depressor	Depressor anguli oris
17	Chin Raiser	Mentalis
20	Lip Stretcher	Risorius platysma
23	Lip Tightener	Orbicularis oris
25	Lips Part	Labii inferioris, mentalis, orbicularis oris
26	Jaw Drop	Masseter, temporalis, internal pterygoid
28	Lip Suck	Orbicularis oris
45	Blink	Levator palpebrae superioris, orbicularis oculi, pars palpebralis

Table A2

Tiers and annotations used for manual communicative signal coding

Tiers	Entry values	Criteria
Gaze	-	Speaker averts gaze
Blink	-	Brief closure of the eye
Squint	-	Contraction lower (and upper) eyelid(s)
Eyes-widening	-	Upper lid movement that opens eye wider than usual
Eyebrows	Raised	Upward movement
	Unilateral-raised	Upward movement of one eyebrow
	Lowered	Downward movement without contraction in the middle
	Frown	Contraction in the middle with or without downward movement
	Frown-raised	Contraction in the middle with upward movement
	Unilateral-frown-raised	Contraction in the middle with upward movement of one eyebrow
Nose-wrinkle	-	Contraction on top of the nose
Mouth	Lips pressed together	Thinning
	One/both corners pulled back	Stretch
	One/both corners pulled down	Upside-down smile

Tiers	Entry values	Criteria
	Lips pursed	Formed like for a kiss
	Smile	Combined with speech, hence not pure laughter
	Unilateral smile	Smile on one side of the mouth

Note. Gaze estimation is no AU, but an estimation by OpenFace of the direction of the gaze.

Appendix B

Table B1

Global agreement matrices for AU4_c and frown-related manual annotations, and for the altered AU4 and frown related manual annotations

OF output	Manual annotation ‘frowns’		
	frown	no frown	Unmatched
	Original AU4_c		
frown	150	1262	4209
no frown	18	1650	4441
Unmatched	321	1333	-
	Altered AU4		
frown	86	4	603
no frown	30	4596	690
Unmatched	347	489	-

Table B2

Global agreement matrices for AU45_c and manually annotated blinks, and for the altered AU45 and manually annotated blinks

OF output	Manual annotation 'blink'		
	blink	no blink	Unmatched
	Original AU45_c		
blink	2072	1006	5295
no blink	18	6048	4893
Unmatched	3761	3568	-
	Altered AU45		
blink	4080	10	2417
no blink	14	7770	2726
Unmatched	2759	3205	-

Table B3

Global agreement matrix for smile annotations derived from the output of OpenFace, and manually annotated smiles

OF output	Manual annotation 'smile'		
	smile	no smile	Unmatched
	Smiles extracted from OF output		
smile	840	392	2548
no smile	40	3970	2435
Unmatched	544	961	-

Table B4

Global agreement matrix for gaze annotations derived from the output of OpenFace, and manually annotated gaze aversion

OF output	Manual annotation 'gaze'		
	gaze away	gaze not away	Unmatched
	Gaze aversion extracted from OF output		
gaze away	972	578	2606
gaze not away	220	4544	2427
Unmatched	1705	1917	-