# The Communicative Face

|  |  |
|---|---|
| Name: | Jordy Ripperda |
| E-mail address: | j.ripperda@student.ru.nl |
| Student number: | s4381386 |
| Phone number: | +316 31960040 |
| Specialisation: | Web and Language Interaction |
| Supervision: | Affiliated |
| Course code: | SOW-MKI94-2018-JAAR-V |

**Supervisor 1**

|  |  |
|---|---|
| Name: | dr. Franc Grootjen |
| Telephone: | 024-3612537 |
| E-mail: | f.grootjen@donders.ru.nl |

**Affiliated Supervisor 2**

|  |  |
|---|---|
| Name: | dr. Judith Holler |
| Institute: | Donders Centre for Cognition |
| Telephone: | +31-24-3521268 |
| E-mail address: | judith.holler@mpi.nl |

**Affiliated Supervisor 3**

|  |  |
|---|---|
| Name: | dr. Linda Drijvers |
| Institute: | Donders Centre for Cognition |
| Telephone: | +31-24-3521591 |
| E-mail address: | linda.drijvers@mpi.nl |

Radboud University
Netherlands
December 2019

# Contents

**1 Introduction 1**

1.1 Multimodal communication . . . . . . . . . . . . . . . . . . . . . . 1

1.2 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

1.3 FACS & OpenFace . . . . . . . . . . . . . . . . . . . . . . . . . . 5

1.3.1 Facial landmark detection and tracking . . . . . . . . . . 5

1.3.2 Head pose and eye gaze estimation . . . . . . . . . . . . . 6

1.3.3 Facial signal recognition . . . . . . . . . . . . . . . . . . . 6

1.4 N-grams . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

**2 Finding regularities between facial signals and questions and responses 10**

2.1 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.1.1 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.1.2 Annotations . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.1.3 Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

2.2 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

2.2.1 General counts . . . . . . . . . . . . . . . . . . . . . . . 16

2.2.2 Co-occurrence . . . . . . . . . . . . . . . . . . . . . . . . 18

2.2.3 N-grams . . . . . . . . . . . . . . . . . . . . . . . . . . . 18

2.2.4 Questions-Response gap . . . . . . . . . . . . . . . . . . . 20

2.3 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21

**3 SPUDNIG 23**

3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23

3.2 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25

3.2.1 OpenPose . . . . . . . . . . . . . . . . . . . . . . . . . . 25

3.2.2 Movement/no-movement detection . . . . . . . . . . . . . 26

3.2.3 Post-processing . . . . . . . . . . . . . . . . . . . . . . . 27

3.2.4 Graphical user interface . . . . . . . . . . . . . . . . . . . 27

3.2.5 Validation analyses . . . . . . . . . . . . . . . . . . . . . 28

3.2.6 Does SPUDNIG accelerate the annotation process? . . . . 29

3.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30

3.3.1 Form-based annotations . . . . . . . . . . . . . . . . . . . 30

3.3.2 Meaning-based annotations . . . . . . . . . . . . . . . . . 30

3.3.3 Iconic gestures . . . . . . . . . . . . . . . . . . . . . . . . . 30
3.3.4 Non-iconic gestures . . . . . . . . . . . . . . . . . . . . . . . 30
3.3.5 Movement/gesture detection . . . . . . . . . . . . . . . . . 30
3.3.6 Acceleration . . . . . . . . . . . . . . . . . . . . . . . . . . . 31
3.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31
3.4.1 Performance . . . . . . . . . . . . . . . . . . . . . . . . . . 31
3.4.2 Limitations . . . . . . . . . . . . . . . . . . . . . . . . . . . 32

**4 General Discussion 33**
4.1 Regularities and their timing . . . . . . . . . . . . . . . . . . . . 33
4.2 N-grams . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34
4.3 Gap duration . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35
4.4 SPUDNIG . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35

**5 Conclusion 37**
5.1 Facial signals . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37
5.2 SPUDNIG . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 38

## Abstract

One of the most fundamental human activities consists of communication through human language. The most important aspect of human language is face-to-face interaction, suggesting human language is a multimodal phenomenon. There is an enormous variation in the face's articulators and the potential signals they can produce. This research will investigate whether any regularities regarding those facial signals occur when comparing questions to responses. Data from dyadic conversations were used where participants talked freely for 60 minutes, in an attempt to take a more naturalistic approach compared to most existing literature that uses conversational data from highly controlled environments. Facial signals are recognized with the help of OpenFace, a tool for the automatic detection of facial signals from video data. General (co-)occurrence counts of facial signals, as well as sequences of facial signals, and their timing were analyzed while comparing questions to responses. Significant differences between questions and responses were found both agreeing, as well contradicting existing literature. Therefore this research could provide more insight to what facial signals occur systematically during questions and responses and possibly help to addressee to predict the content or ending of the incoming turn.

Furthermore, SPeeding Up the Detection of Non-iconic and Iconic Gestures (SPUDNIG): a toolkit for the automatic detection of hand movements and gestures in video data was presented. This toolkit was developed since there was no existing toolkit for the automatic detection of hand movements, in contrast to toolkits for the automatic detection of facial signals such as OpenFace. It was demonstrated that SPUDNIG accelerates the process of annotating hand gestures. Therefore SPUDNIG could be used in order to facilitate the time-consuming and labour-intensive task of manually annotating hand gestures.

# Chapter 1

# Introduction

## 1.1 Multimodal communication

Nowadays communication comes in all forms and can be found everywhere. Human communication, through spoken language, allows us to share and communicate our knowledge, and sets us apart from all other species on this planet. The underlying cognitive processes of communication and language processing have been researched extensively over the years, yet we are still far away from understanding the full process and all components that are involved. The most important aspect of human language is face-to-face interaction, suggesting human language is a multimodal phenomenon. This means different kinds of cues from different kinds of articulators are involved in the interaction process, which vary in modality (e.g. auditory, visual, haptic, olfactory). Whereas traditionally articulators only represent the organs that produce the sound of language (i.e., the glottis, velum, hard palate, lips, teeth, tongue), in this study articulators also include the forehead, eyebrows, both eyelids, the muscles around the mouth, nose and cheeks. Considering the face contains 43 muscles, there is a huge amount of potential signals produced during speech. As [Holler and Levinson, 2019] describe, this raises two computational challenges. First, not all visual cues are intended to be part of the signal as some are incidental and irrelevant to the content or signal of the speaker (segregation problem). Second, the visual cues that appear to be relevant to the content of the speaker have to be matched with their counterparts, where simply taking into account simultaneity turns out unreliable (binding problem). All of these visual signals are layered onto the vocal signal, resulting in an abundance of signal onsets and offsets. One would think unifying those layers would be quite a complex and demanding task, yet [Holler et al., 2018] showed that questions paired with gestures get faster responses. Such findings suggest that the body plays a significant role in the psycholinguistic processes regarding human communication. This study aims to find whether this is the case for facial signals as well, by testing for regularities in facial signals during speech in order to provide more insight to the psycholinguistic processes of human language processing. While there exists ample literature on spoken language processing, most merely focuses on the auditory linguistic signal. The remainder of this chapter will give an overview of studies that have been done regarding multimodal processing of visual signals during speech. [Benitez-Quiroz et al., 2016] suggested that facial expressions have grammatical function and that some components of human language have evolved from facial expressions of emotion. This study however, like many studies in this field, forced participants to produce the facial expression of negation, and the events did not occur during free speech. [Ekman, 2009] showed that the facial shrug, consisting of an eyebrow flash and one mouth corner being retracted, often is paired with signals being 'I don't know' or 'OK'. These results could serve as evidence of this specific sequence of facial signals occurring more in responses than in questions. Raised eyebrows turned out to serve as question markers, supporting the hypothesis that different visual cues occur in questions relative to responses [Ekman, 2004, Borras-Comes et al., 2014, Sendra et al., 2013, Chovil, 1991]. In [Cavé et al., 2002] they show that, in French, eyebrow raises appeared closer to the start of a turn rather than to the end, and [Cassell et al., 2001] showed that posture shifts happened more towards boundaries of turns rather than within turns. These findings suggest at least some form of regularities between visual signals and speech acts exist, and the aim of this study is to test whether more of those patterns occur during multimodal conversation. In contrast to previous research this study consists of free speech, i.e., participants are not instructed to produce any type of speech or signals. Furthermore this study will not focus on one specific facial signal but will investigate a broader range of facial signals. A subset of the Facial Action Coding System (FACS), which will be described later this chapter, will be investigated. This way a more naturalistic approach whether taken to investigating if facial signals contribute to the understanding processes during communication.

Recent research has demonstrated that prediction plays a fundamental role during the processing of verbal language [Pickering and Garrod, 2013, Van Berkum et al., 2005, DeLong et al., 2005, Federmeier and Kutas, 1999]. This study assumes that facial signals accompanying the conversational turns in human communication also contribute to the addressee's prediction of the content of the incoming turn and its ending. To investigate this assumption, this study aims to find statistical regularities in the production of multimodal turns. The goal is to associate certain speech acts, i.e., questions and responses, with visual signals systematically produced by the speaker, focusing on facial signals since they can be easily recognized by OpenFace [Baltrusaitis et al., 2018], which will be described later. Since regarding this topic almost no research has been done, this study could contribute towards a unified model of communication in speech by providing more insight in which facial signals, if any, are systematically used during questions or responses. This study focuses on question-response sequences since this phenomenon is independent of language and cultures, since it is a universal unit of conversational organization [Stivers et al., 2009].

For most facial signals it is unknown how, if at all, they contribute to the process of understanding. Previous research has shown that certain eyebrow movements serve as emphasizers and question markers [Chovil, 1991]. In [Flecha-García, 2002] it was shown that queries contained more eyebrow raises then replies. Furthermore, similar to [Chovil, 1991] and [Cassell et al., 2001], [Flecha-García, 2010] also found that eyebrow raises occurred most frequently at the start of transactions (i.e. sets of utterances at a dialogue level). At a lower level, it was found that in French eyebrow raises occurred more close to the start of a turn rather than to the end [Cavé et al., 2002]. Although most research focuses on eyebrow movements, the face contains many more muscles that are able to form various facial signals. The goal of this study is to investigate whether more of such regularities occur within the question-response sequences. The following questions are at the center of this research:

> *"Are there any regularities between speech acts, i.e., questions and responses, and facial signals (e.g. eyebrow raises might occur more in questions whereas eyebrow lowerers might occur more in responses)?"*

If so:

> *"Are there any particular sequences of facial signals within questions/responses that occur systematically?"*

And:

> *"What is the timing of the occurrences/sequences of facial signals with regards to questions and responses?"*

Furthermore:

> *Will certain facial signals lead to shorter gap durations between questions and its corresponding responses?*

The hypothesis is that there are certain regularities between speech acts and facial signals (e.g. eye brow raises tend to occur more during questions). Since an exploratory approach is taken there are no specifically expected regularities regarding facial signals. Another hypothesis is that questions tend to contain more eyebrow raises than responses, based on the findings in [Flecha-García, 2002]. The final hypothesis is that (sequences of) facial signals, occur more at the start of each question or response, like it appeared to be for eyebrow raises [Flecha-García, 2010]. This would support the idea that signals appear early in a question or response, such that the addressee has more time to predict information about the content and expected ending of the speaker's turn.

Another phenomenon that suggests that there indeed should be some sort of regularities is one of the most fascinating properties of human communication, namely the turn-taking system. The turn-taking system represents the rapid exchange of short turns at talking in conversations[Levinson, 2016, Holler et al., 2016]. The turn-taking system depends on rules to minimize the number of turns [Sacks et al., 1978]. The first person to respond obtains the rights to the turn, and releases them on upon turn-completion. Turns do not

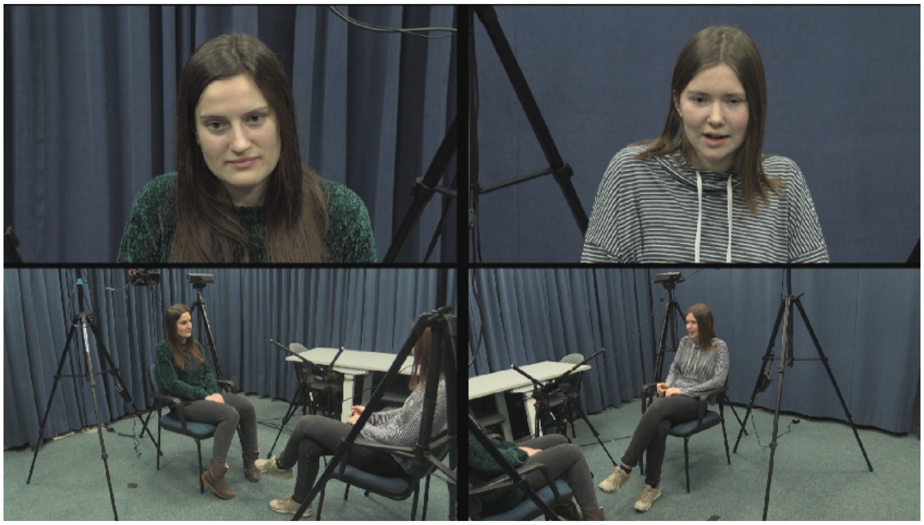


Figure 1.1: Upper panels: one face camera for each participant. Lower panels: one body camera for each participant.

have a fixed size, and they tend to be very short, about 2 seconds on average [Levinson and Torreira, 2015]. The turn-taking systems minimizes overlap between turns and does so highly efficiently. Even in conversations with more than 2 people, speech streams only tend to involve more than 1 speaker (i.e. overlap between different speakers) in less than 5%. The modal gap between turns tends to be only 200 ms [Stivers et al., 2009], which is of equal length as a single syllable. In theory it seems impossible to produce meaningful responses during conversation, taking into account the short response times above, together with the fact the production of single word already takes 600 ms to produce [Indefrey and Levelt, 2004]. Yet people manage to succeed in doing so. This implies that responses must be planned during the incoming turns in order to produce a response in time. Therefore prediction of turn-ending and anticipation of incoming turn content is required. In this thesis it is assumed that facial signals help to constitute these predictions. Certain facial signals could for example indicate that the current turn represents a question and that an answer is soon expected from the addressee.

## 1.2 Data

To investigate the research questions this thesis addresses, data from a multimodal communication corpus (CoAct corpus, ERC project #773079) will be used. The data was collected during an experiment conducted by researchers at and from the Max Planck Institute for Psycholinguistics (MPI) in Nijmegen. It consists of 18 pairs of Dutch native speakers that engaged in dyadic, casual conversation. The conversation was one hour in total and divided into three parts of twenty minutes each: free speech, discussion about a given topic, and planning a vacation respectively. Participants were both audio- and video-recorded and for each participant a face and a body camera was used (see Figure 1.1).

To find any regularities both counts of question and responses together with the occurring facial signals are investigated as well as the co-occurrences of those facial signals. The annotations of the questions and responses is done by trained researchers from the MPI and is done in ELAN [Wittenburg et al., 2006], a professional tool for the creation of complex annotations on video and audio resources. The annotations of the facial signals are later added with the help of OpenFace [Baltrusaitis et al., 2018], a facial behavioural analysis toolkit that is capable of recognizing facial signals from video data.

Furthermore this study will not focus on one specific facial signal but will investigate a broader range of facial signals. Due to the performance op OpenFace, only a subset of the Facial Action Coding System (FACS), which will be described in the next chapter, will be investigated.

## 1.3 FACS & OpenFace

This study aims to find systematic facial signals, if any, within the question-response sequences, and focuses on the Facial Action Coding System (FACS) [Ekman and Friesen, 1978]. The FACS refers to a set of facial muscle movements that often correspond to a displayed emotion. Researchers have for a long time manually annotated such facial signals, which is a very labour-intensive and time-consuming process. Therefore, automating the recognition process of facial signals is an important area of machine learning and computer vision research because of its relevance for behavioral analysis and for the further development of human-computer interaction (HCI). The FACS was able to decompose facial expressions in such a way that it lends itself to automatic recognition of those signals. Since manual facial behaviour analysis is such a labour-intensive task, there has been an increasing interest in an automatic version of this. OpenFace 2.0 [Baltrusaitis et al., 2018] is a tool that aims to solve that problem and will be used during this study to recognize facial signals. This tool is developed for machine learning, computer vision and facial behaviour analysis researchers. It can be used to obtain the facial landmark detection, head pose estimation, facial action unit recognition and eye gaze estimation from video data. Figure 1.2 shows an image of OpenFace's output from a video from the corpus used in this study. Next the pipeline of OpenFace will be discussed briefly.

### 1.3.1 Facial landmark detection and tracking

Facial landmarking and tracking is represented by the process of detecting and tracking points of interest in the face (see Figure 1.2). To solve this problem, OpenFace uses a Convolutional Experts Constrained Local Model (CE-CLM) proposed by [Zadeh et al., 2017]. To initialize to CE-CLM model

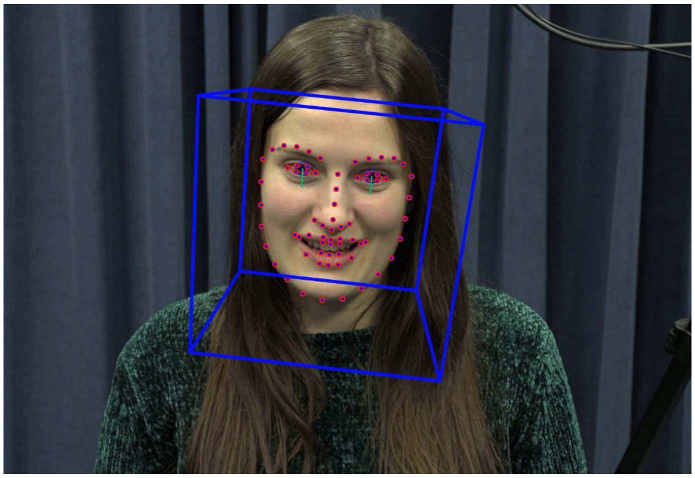


Figure 1.2: OpenFace applied to a video from the corpus used in this study. Blue box represents head pose estimation; red dots represent facial landmark detection; green beams represent eye-gaze estimation.

a Multi-task Convolutional Neural Network (MTCNN) face detector is used [Zhang et al., 2016]. To prevent tracking drift a four layer CNN is used to report tracking failure. If the validation CNN fails the model is reinitialized by the MTCNN face detector.

### 1.3.2 Head pose and eye gaze estimation

As the CE-CLM uses a 3D representation of the facial landmarks, OpenFace additionally is able to estimate head pose through solving the Perspective-n-Point problem [Hesch and Roumeliotis, 2011]. This is represented by the blue box in Figure 1.2.

To estimate the eye gaze, OpenFace uses a Constrained Local Neural Field (CLNF) landmark detector [Baltrusaitis et al., 2013] that was trained on the SynthesEyes dataset [Wood et al., 2015]. This dataset contains around 11000 close-up images of eyes, each annotated with landmark and gaze information. Eye gaze is represented by the green beams in Figure 1.2. This study however will not use head pose and eye gaze information because the focus of this study lies on facial signals.

### 1.3.3 Facial signal recognition

For recognizing facial signals OpenFace detects facial action unit (AU) presence and intensity from the FACS discussed in chapter 1.3. A method similar to [Baltrusaitis et al., 2015], that uses linear kernel Support Vector Machines (SVM), to recognize the AUs. Although deep learning models often outperform

SVMs, [Baltrusaitis et al., 2018] have demonstrated its competitiveness compared to recent deep learning models. The AUs that OpenFace is able to detect are shown in Table 1.1. Those are also the AUs this research will focus on. Some AUs might be dropped during analysis based on how well OpenFace performs in recognizing them.

## 1.4 N-grams

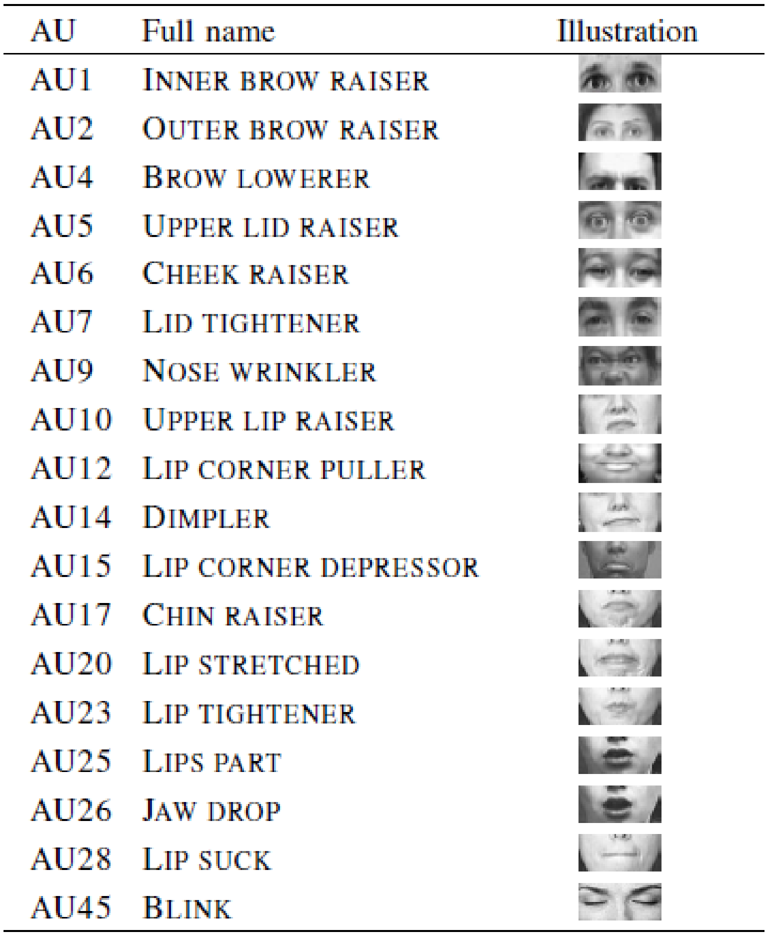
Finding regularities between facial signals and the questions and responses, also entails finding whether certain sequences of facial signals occur systematically (e.g. brow lowerer followed by the upper lip being raised during a response). In order to find such sequences n-gram models are used.

In computational linguistics, n-grams are contiguous sequences of $n$ items from a given sample of speech or text. Items vary from words, letters, syllables and phonemes. In this study the items represent occurrences of facial signals (also called action units) at a certain time point. The benefits of n-gram models are its simplicity and scalability. N-gram models are widely used in various fields. In natural language processing (NLP) n-grams have been used as machine learning features [Sidorov et al., 2014, Pagliardini et al., 2017]. [Pak and Paroubek, 2010] and [Kouloumpis et al., 2011] used n-grams among other in sentiment analysis in tweets. Furthermore n-grams form the basis of many speech recognition systems [Katz, 1987, Hirsimaki et al., 2009, Young, 1996, Bellegarda, 2000].

However, n-grams are not limited to natural language. [Santos et al., 2009] and [Wang et al., 2007] showed that n-grams can be used to detect unknown malware, solving the problem that not merely registered viruses can be detected. In the field of computational biology n-grams are used for protein- and DNA-sequencing, which is explained as determining the order of nucleotides in DNA or the amino acid sequence of a protein.

This study will use n-grams to find whether specific sequences of facial signals occur systematically in the scope of question-response sequences. N-grams are well-suited for this problem because they find the most occurring sequences of facial signals that exist in the data. N-grams are preferred over sequential pattern mining, because they are better at handling data sparsity (i.e. some signals might occur more than others), since it searches for existing patterns rather than examining a combinatorially explosive number of possible sequence patterns.

The remainder of this thesis is organized as follows. The next chapter will discuss the methods, results, and discussion of the results of the analysis for finding regularities regarding facial signals in question-response sequences. Chapter 3 will discuss the development of SPeeding Up the Detection of Non-iconic and Iconic Gestures (SPUDNIG): a toolkit for the automatic detection of hand movements and gestures in video data [Ripperda et al., 2019]. This toolkit was developed during this thesis since there was no existing toolkit for the automatic detection of hand gestures, in contrast to OpenFace which is able to solve this



| AU | Full name | Illustration |
|---|---|---|
| AU1 | Inner brow raiser | |
| AU2 | Outer brow raiser | |
| AU4 | Brow lowerer | |
| AU5 | Upper lid raiser | |
| AU6 | Cheek raiser | |
| AU7 | Lid tightener | |
| AU9 | Nose wrinkler | |
| AU10 | Upper lip raiser | |
| AU12 | Lip corner puller | |
| AU14 | Dimpler | |
| AU15 | Lip corner depressor | |
| AU17 | Chin raiser | |
| AU20 | Lip stretched | |
| AU23 | Lip tightener | |
| AU25 | Lips part | |
| AU26 | Jaw drop | |
| AU28 | Lip suck | |
| AU45 | Blink | |

Table 1.1: List of AUs that OpenFace is able to detect. Copyright © 2018, IEEE

problem for facial signals. This chapter contains its own introduction, methods and results since this can be seen as a side project of this thesis. Chapter 4 provides a general discussion followed by a conclusion including suggestions for future work.

# Chapter 2

# Finding regularities between facial signals and questions and responses

## 2.1 Methods

This chapter provides an overview of the methods and tools used in this thesis in order to find any regularities between speech acts, i.e., questions and responses, and facial signals.

### 2.1.1 Data

The data that was used is described in chapter 1 and Figure 1.1. The data includes 18 pairs of Dutch native speakers who engaged in dyadic casual conversation. The conversation was one hour in total and divided into three parts of twenty minutes each: free speech, discussion about a given topic, and planning a vacation, respectively. This resulted in 60 minutes of data for 36 speakers (2160 hours).

The exact timing of the questions and responses were annotated in ELAN [Wittenburg et al., 2006] by trained researchers from the Max Planck Institute (MPI) Nijmegen. In total 3191 questions and 2209 responses were found.

### 2.1.2 Annotations

Instead of manually annotating each facial signal/action unit made by the participants during the questions and responses, this research uses OpenFace, described in chapter 1.3, for the automatic recognition of these facial signals. Each video was analyzed using OpenFace. Since this needed to be performed for each video individually, and OpenFace does not provide any tools to process multiple files at the same time, or even sequentially, I created a user-friendly command-line interface in Python (`https://github.com/jorrip/OpenFaceAnalyzer`). Here users can specify a root folder, an output folder, and an optional parameter -suffix (i.e. if suffix is specified as 'face.mpg' it will only analyze all files in the root folder that end with 'face.mpg'). An example command to run the program would look as follows:

```
python C:\Downloads\run_open_face.py -root C:\workspaces\videos
-open C:\OpenFace\OpenFace_2.0.5_win_x64\FeatureExtraction.exe
-out C:\OpenFace\Output -suf face.mpg
```

The OpenFace output was then imported into ELAN using Exploface (`https://github.com/emrecdem/exploface`), and merged with the question and response annotations. Figure 2.1 demonstrates an example file in ELAN with the question and response annotation together with OpenFace's output. Each facial signal (see Table 1.1 for which signals OpenFace is able to recognize) has its own tier (i.e. row of annotations), as well do the questions and responses. As can be seen in Figure 2.1, lots of facial signals occur according to OpenFace.

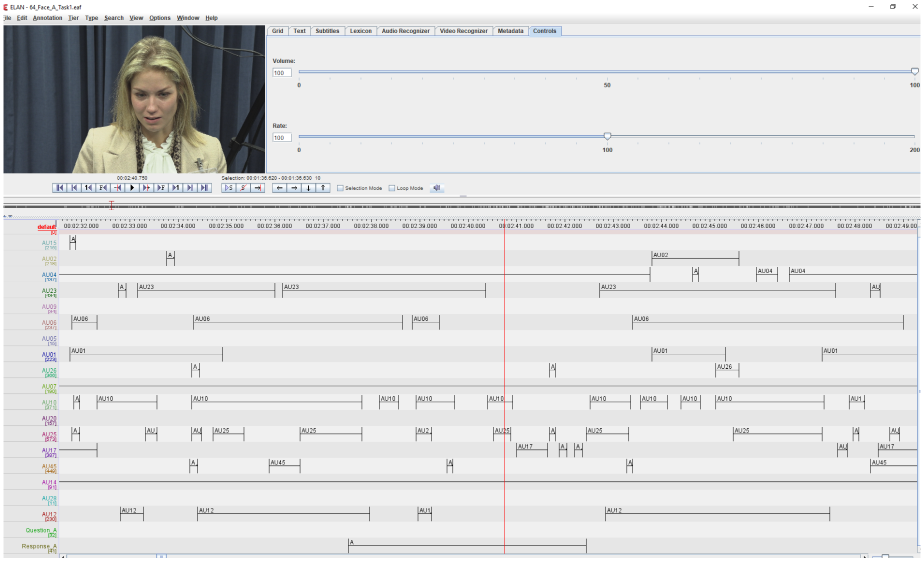


Figure 2.1: ELAN file with question and response annotation together with OpenFace's annotations. Top left presents a still of the video at a certain point in time, which is indicated by the vertical red line. In the bottom half, each action unit contains its own row of annotations generated by OpenFace, and the bottom two rows indicate the occurrence of a question or response recognized by human coders. The current time point is located during a response of the participant in the video.

Unfortunately OpenFace's output contains lots of false positives, resulting in an overabundance of annotations. Possible explanations for this are that participants facing their head away from the camera, or even looking down which often results in incorrectly recognized eye blinks.

However, since the main goal of this thesis is to provide a pipeline for investigating whether there are regularities between the questions and responses and the facial signals, OpenFace's output will not be cleaned during this research. Although this might highly affect the results, cleaning the OpenFace output lays beyond the scope of this thesis.

### 2.1.3 Analysis

All analysis steps are done in Python 3.7.0 with Jupyter Notebook [Kluyver et al., 2016]. For each video, the tiers for the facials signals and the questions and responses are exported from ELAN, which results in a tab-separated file with on each line the begin and end time (hh:mm:ss.ms) of the annotation, followed by a token indicating which facial signal or question/response occurred during that annotation (e.g. AU25 or Question). After the hh:mm:ss.ms format was converted to the corresponding frame number, these tab-delimited files were converted into a Pandas DataFrame[1], a two-dimensional tabular data structure often used for analyzing large structured data sets [McKinney, 2011]. The DataFrame served as a timeline of the corresponding video, where each row represented a frame, and each column represented the facial signal/action unit or question or response happening in that frame. The DataFrame contained binary values where '1' would indicate the occurrence, and '0' the absence of the corresponding facial signal or question/response happening in the corresponding frame.

Next, from this timeline the on- and off-sets of the questions and responses were determined, and extracted from 500 ms before onset til 500 ms after offset of the corresponding question or response. This was done to capture signals that would occur before onset of the speech signal. If taking into account 500 ms before onset would result in overlap with a preceding question or response from the same speaker, this step was omitted to eliminate within-speaker effects from overlap between questions and responses. This resulted in 3191 questions and 2209 responses of unequal lengths, produced by 36 different speakers that talked for 1 hour. Of those, 143 questions and 30 responses contained overlap from preceding questions and responses hence the 500 ms before onset was omitted. Each question or response is stored as a list, where each element in the list is represented by a string containing the action units occurring in the corresponding frame (e.g. ['AU1AU2', 'AU3AU4', 'AU5AU6'] where AU1 and AU2 occur in the first frame, AU3 and AU4 in the second frame etc.).

From here, descriptive statistics can be easily calculated, such as general occurrence, and co-occurrence counts for example, where questions were compared to responses. Furthermore it was investigated how many facial signals occurred

[1] `https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html`

per question or response instance (henceforth: QRI), to exclude effects from the fact that there are more questions than responses. In more detail, simply counting the occurrences would create a bias towards questions since there are more questions in the data than responses. However, this would still not be a fair comparison, since overall questions might be have a longer duration than responses (or vice versa), which would cause increased occurrence counts for the questions. Therefore it was also investigated how many facial signals occurred per QRI, when only taking into account the interval between 500 ms before onset and a fixed amount of time after onset, to ensure the durations of the QRIs are equal. This was also performed for intervals subsequently to the previous interval. The following intervals were investigated: 500 ms before onset - onset, onset - 500 ms after onset, and 500 ms before offset - offset. This way it was investigated which facial signals occur more towards the start, or towards the end of a QRI. In each interval, the occurrences of each facial signal were counted during questions and responses. Whenever the second interval would overlap with the third interval, i.e., in QRIs with a duration shorter than 1000 ms, the corresponding QRI would be excluded from the analysis. For each facial signal, and in each interval, a Welch two-sample t-test [Welch, 1947] was used to test for a statistical difference in occurrence of the facial signals in questions versus responses. This test is used since it does not assume the data to be normally distributed, and is capable of handling large, unequal sample sizes, which is necessary since the number of question is not equal to the number of responses.

Next the co-occurrence of facial signals was investigated by creating a co-occurrence matrix, where again was differentiated between questions and responses. Two types of co-occurrences were investigated: one where the facial signals occurred in the same frame (so completely simultaneously), and one where facial signals simply occurred in the same question or response.

Next, to investigate whether certain sequences of facial signals occurred frequently, n-grams were used, where a unit of the n-gram sequence is represented by the signals occurring in a frame. Note that a QRIs are represented as lists where each element indicates which action units occur in the corresponding frame. Here follows an example of a question to illustrate this:

```
[’AU15AU06AU05AU10AU17AU14AU12’,
’AU15AU06AU05AU10AU17AU14AU12’,
’AU15AU06AU05AU10AU17AU14AU12’,
’AU06AU05AU10AU17AU14AU12’,
’AU06AU05AU10AU17AU14AU12’,
’AU06AU05AU10AU14AU12’,
’AU06AU05AU10AU25AU14AU12’,
’AU06AU05AU10AU25AU14AU12’,
’AU06AU05AU10AU25AU14AU12’,
’AU06AU05AU10AU25AU14AU12’,
’AU06AU05AU10AU25AU14AU12’,
’AU06AU05AU10AU25AU14AU12’,
’AU06AU05AU10AU25AU14AU12’,
’AU06AU05AU10AU25AU14AU12’,
’AU06AU05AU10AU25AU17AU14AU12’,
’AU06AU05AU17AU14AU12’,
’AU04AU06AU05AU17AU14AU12’,
’AU04AU06AU05AU17AU14AU12’]
```

In this example, the first line indicates which action units occur in the first frame (AU15AU06AU05AU10AU17AU14AU12), and the second line which action units occur in the second frame etc. As already mentioned, OpenFace’s output is still messy which results in an overabundance of action units. When searching for n-grams that occur frequently, the ideal case would be to check all possible combinations where from each frame one action unit is picked. An attempt to produce all possible combinations was made by calculating the Cartesian product, however, with an average QRI duration of 30 frames (= 1.2 sec) and an average of 5 action units per frame, this would result in $30^5 = 24300000$ possible combinations per question or response, which would take too long to analyze.

To overcome this problem, instead of generating all combinations, the entire frames were used as units for the n-grams. So in the example above ’AU15AU06AU05AU10AU17AU14AU12’ (first frame) would count as one unit. This would mean that when checking for n-grams of size 3, a sequence of facial signals would be checked of 3 frames. As can be seen in the example question, subsequent frames are often identical, since one frame only represent 0.04 sec when using 25 fps. This would mean that to find meaningful n-grams, which cover more than 0.5 seconds, n-grams with a least a size of 12 would need to be checked. Therefore the questions and responses are post-processed, meaning that consecutive identical frames are removed. For the example displayed above, the post-processed result would look as follows:

```
['AU15AU06AU05AU10AU17AU14AU12',
'AU06AU05AU10AU17AU14AU12',
'AU06AU05AU10AU14AU12',
'AU06AU05AU10AU25AU14AU12',
'AU06AU05AU10AU25AU17AU14AU12',
'AU06AU05AU17AU14AU12',
'AU04AU06AU05AU17AU14AU12']
```

This way information about the duration of the facial signals is lost, but information about the order in which the signals occur is preserved. The advantage is that the size of the n-grams that need to be checked is much smaller. Therefore, when representing the data like this, n-grams are a well-suited method in order to find sequences of facial signals that occur most frequently.

Finally, it was investigated whether certain facial signals lead to shorter gap duration between the question and its corresponding answer. If this is the case, this would support the idea that certain facial signals help the addressee to predict the content and expected ending of the turn, allowing them to reply faster. For each question, its corresponding response was found by checking for a response within 250 ms from before the end of the question until 250 ms subsequent to the end of the question. This time window is used since gaps between turns tend to be around 200 ms [Stivers et al., 2009], and because turn transitions sometimes have some overlap [Levinson and Torreira, 2015], meaning that responses would start before the question has ended.

Together with the duration of the gap, it was listed how often each facial signal occurred during the corresponding question. Based on other findings from the analysis, the Spearman's correlation coefficient between the occurrence count of some facial signals and the gap duration was computed.

## 2.2 Results

### 2.2.1 General counts

From the data a total of 3191 questions and 2209 responses were extracted. Table 2.1 presents the total occurrence counts of the action units for questions and responses.

| | AU01 | AU25 | AU20 | AU09 | AU06 | AU02 | AU15 | AU14 | AU28 |
|---|---|---|---|---|---|---|---|---|---|
| **Question count** | 2317 | 6474 | 2336 | 1091 | 3076 | 2212 | 3055 | 4243 | 209 |
| **Response count** | 2038 | 4550 | 1754 | 653 | 2114 | 1854 | 2312 | 3106 | 163 |
| **Total** | 4355 | 11024 | 4090 | 1744 | 5190 | 4066 | 5367 | 7349 | 372 |

| | AU04 | AU23 | AU07 | AU10 | AU12 | AU17 | AU26 | AU45 | AU05 |
|---|---|---|---|---|---|---|---|---|---|
| **Question count** | 3184 | 4202 | 3560 | 4993 | 3227 | 4116 | 4869 | 3888 | 1153 |
| **Response count** | 2343 | 3018 | 2447 | 3322 | 2069 | 2824 | 3644 | 2999 | 938 |
| **Total** | 5527 | 7220 | 6007 | 8315 | 5296 | 6940 | 8513 | 6887 | 2091 |

Table 2.1: Counts of the action units recognized by OpenFace during questions and responses.

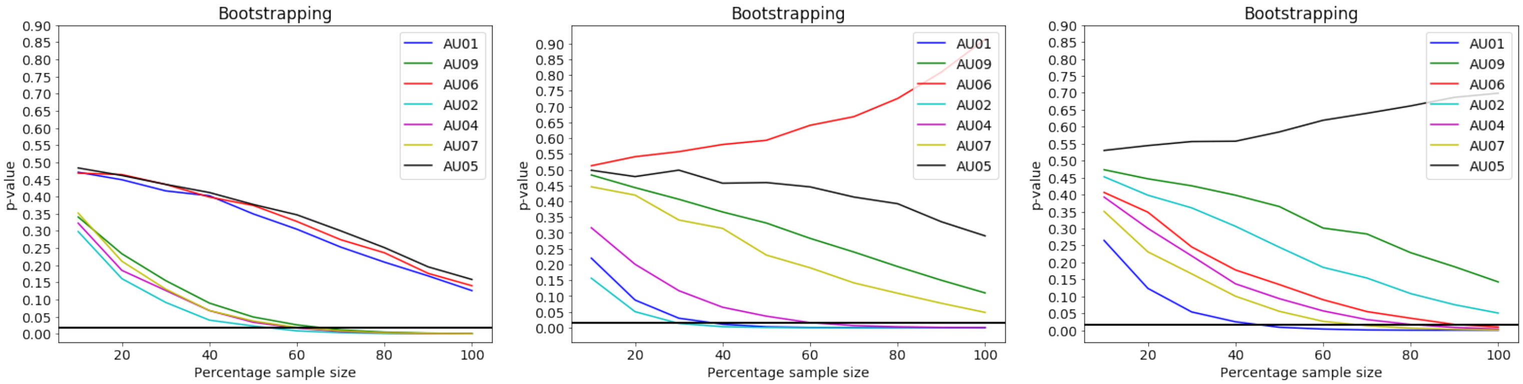
As can be seen the numbers are quite excessive, especially the actions units regarding mouth movements (23, 25, 20, 15, 14, 12, 17, 26, 10, 45). These high occurrence counts are explained by the fact that participants are moving their mouths while talking, causing OpenFace to recognize an overabundance of facial signals regarding mouth movements. Therefore these action units were discarded from further analysis.

Table 2.2 presents the mean occurrence counts of the corresponding action units per QRI. A distinction is made between the interval from 500 ms before onset until onset, the interval from onset until 500 ms after onset, and the interval from 500 ms before offset until offset. The first represents the facial signals preceding the question or response, the second the facial signals at the start of the question or response, and the last the facial signals towards the end of the question or response. The mean and standard deviation for the occurrence count of each facial signal and each interval are presented during questions and responses. Furthermore for each facial signal, the outcomes of the statistical test (Welch's two-sample t-test) are given that tested the difference in occurrence of the corresponding facial signal between the questions and responses, including p-values and effect sizes displayed as Cohen's D. Since three different intervals are tested, a Bonferroni correction was performed and a significance level of $0.05/3 = 0.0167$ was used.

| Before | AU01 | | AU02 | | AU04 | | AU05 | | AU06 | | AU07 | | AU09 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Questions** | **Responses** | **Questions** | **Responses** | **Questions** | **Responses** | **Questions** | **Responses** | **Questions** | **Responses** | **Questions** | **Responses** | **Questions** | **Responses** |
| Mean | 0.3090 | 0.3300 | 0.3002 | 0.3558 | 0.5867 | 0.5324 | 0.1413 | 0.1562 | 0.4798 | 0.4581 | 0.6186 | 0.5650 | 0.1131 | 0.0837 |
| Std | 0.4924 | 0.4984 | 0.4863 | 0.5091 | 0.5281 | 0.5350 | 0.3711 | 0.3861 | 0.5319 | 0.5292 | 0.5379 | 0.5342 | 0.3246 | 0.2976 |
| p-value | 0.1258 | | 0.0006* | | 0.0002* | | 0.1582 | | 0.1401 | | 0.0003* | | 0.0006* | |
| Cohens D | -0.0424 | | -0.1117 | | 0.1021 | | -0.0412 | | 0.0408 | | -0.1000 | | 0.0944 | |
| **Start** | **AU01** | | **AU02** | | **AU04** | | **AU05** | | **AU06** | | **AU07** | | **AU09** | |
| | **Questions** | **Responses** | **Questions** | **Responses** | **Questions** | **Responses** | **Questions** | **Responses** | **Questions** | **Responses** | **Questions** | **Responses** | **Questions** | **Responses** |
| Mean | 0.2579 | 0.3232 | 0.2360 | 0.3101 | 0.5114 | 0.4604 | 0.0940 | 0.1028 | 0.4005 | 0.4020 | 0.5425 | 0.5152 | 0.0765 | 0.0652 |
| Std | 0.4376 | 0.4678 | 0.4247 | 0.4626 | 0.5000 | 0.4985 | 0.2919 | 0.3037 | 0.4901 | 0.4904 | 0.4983 | 0.4999 | 0.2658 | 0.2469 |
| p-value | 0.0001* | | 0.0001* | | 0.0002* | | 0.2905 | | 0.9125 | | 0.0401 | | 0.1099 | |
| Cohens D | -0.15 | | -0.167 | | 0.1023 | | -0.0294 | | -0.003 | | 0.0547 | | 0.04 | |
| **End** | **AU01** | | **AU02** | | **AU04** | | **AU05** | | **AU06** | | **AU07** | | **AU09** | |
| | **Questions** | **Responses** | **Questions** | **Responses** | **Questions** | **Responses** | **Questions** | **Responses** | **Questions** | **Responses** | **Questions** | **Responses** | **Questions** | **Responses** |
| Mean | 0.2795 | 0.3404 | 0.3000 | 0.3259 | 0.5560 | 0.5134 | 0.1250 | 0.1213 | 0.5064 | 0.4690 | 0.6327 | 0.5840 | 0.1053 | 0.0928 |
| Std | 0.4700 | 0.4964 | 0.4778 | 0.4831 | 0.5228 | 0.5290 | 0.3510 | 0.3441 | 0.5215 | 0.5179 | 0.5227 | 0.5259 | 0.3190 | 0.2994 |
| p-value | <0.0001* | | 0.0505 | | 0.0035* | | 0.6987 | | 0.0092* | | 0.0008* | | 0.1423 | |
| Cohens D | -0.126 | | -0.05 | | 0.0809 | | 0.0107 | | 0.072 | | 0.0107 | | 0.0404 | |

*significant at p <0.0167

Table 2.2: Results from the Welch two-sample t-tests including p-values and effect sizes (Cohen's D). The difference in means of the occurrence counts of the facial signals were tested between questions and responses. A Bonferroni corrected significance level of 0.0167 was used. Before = 500 ms before onset until onset; start = onset until 500 ms after onset; end = 500 before offset until offset.



(a) Before: 500 ms before onset until onset

(b) Start: onset until 500 ms after onset

(c) End: 500 ms before offset until offset

Figure 2.2: Bootstrapping results. The x-axis presents the percentage of the sample that is used (10% of the sample - 100% of the sample). The y-axis presents the average p-value for the corresponding sample size. The horizontal line represents the Bonferroni corrected significance level.

**Bootstrapping**

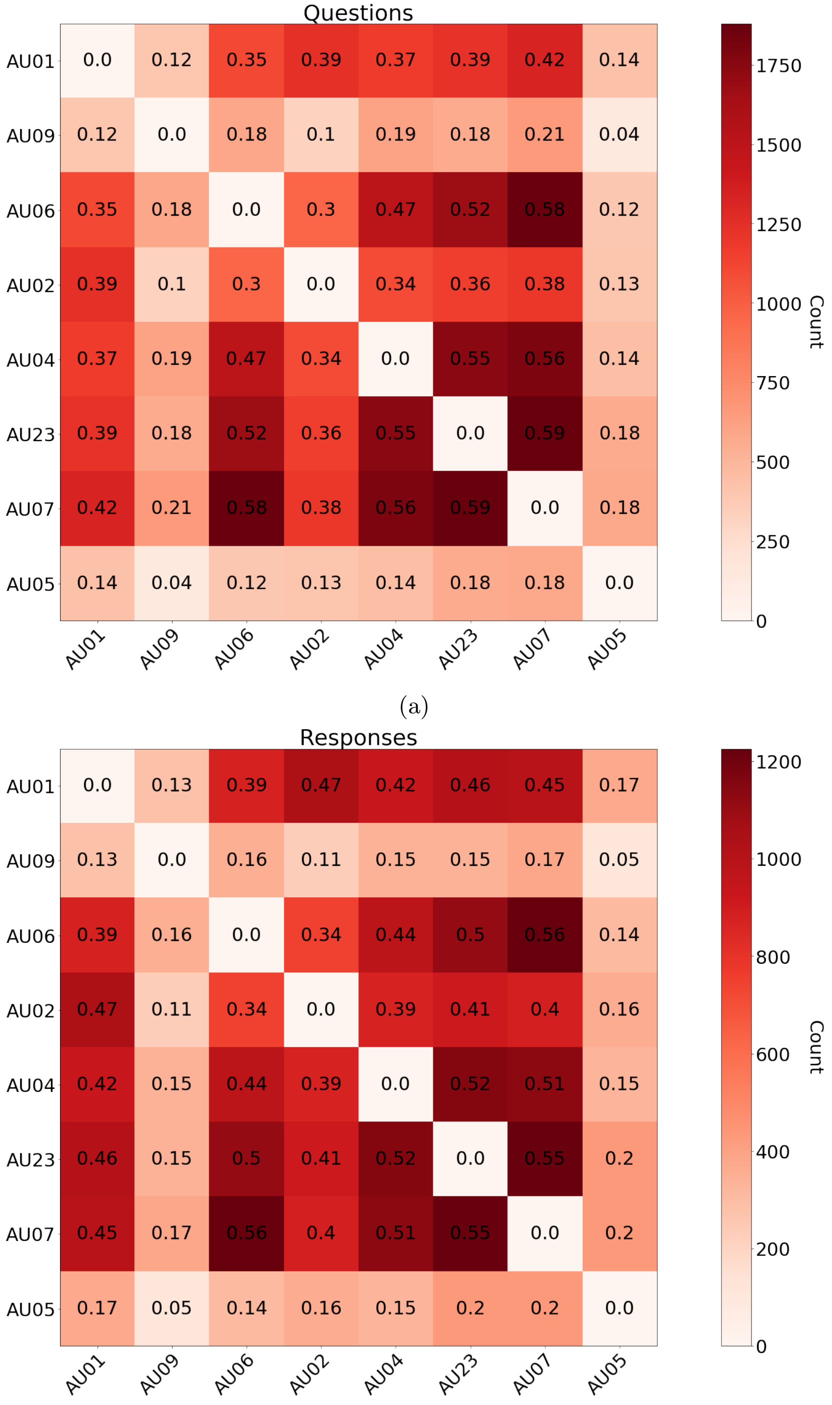
Since in large samples, p-values quickly become (close to) zero [Lin et al., 2013], a bootstrapping method is performed provide more insight about the robustness of the found effects. For every tenfold (10%, 20% etc.) a random subsample is drawn from the complete sample and the p-value is computed for that subsample. This is repeated 1000 times for every subsample and the average p-value is used. This is done so that more insight into whether the found effects are true effects or due to the large sample size. The results are shown in Figure 2.2.

### 2.2.2 Co-occurrence

Figure 2.3 presents the co-occurrence of different action units found within the questions and responses. Two facial signals co-occur when they occur within the same QRI, not necessarily at the same time or in the same frame. For each combination the co-occurrence ratio is computed, where a ratio of 1 would mean that the combination of facial signals occurs in every question or response.

### 2.2.3 N-grams

Table 2.3 presents the most occurring n-grams that are found in questions and responses. The units in the n-grams consist of the facial signals occurring in the corresponding frame. For each n-gram the size of the n-gram, the frequency of the n-gram in either the questions or responses, are presented along with the n-gram itself. Only n-grams with n = 2 and n = 3 are presented since larger n-grams would occur too little.



(a)

(b)

Figure 2.3: Co-occurrence counts of the actions units per questions/response instance (QRI). Two facial signals co-occur when they occur in the same QRI. Figure 2.3a represents the co-occurrence counts of the facial signals during questions, and Figure 2.3b the co-occurrence counts during responses. The individual values represent the co-occurrence rates.

| **Questions** | | | **Responses** | | |
|---|---|---|---|---|---|
| **N** | **Sequence** | **Freq** | **N** | **Sequence** | **Freq** |
| 2 | ('AU04', 'AU04AU07') | 380 | 2 | ('AU04AU06AU07', 'AU06AU07') | 251 |
| 2 | ('AU04AU07', 'AU04AU06AU07') | 372 | 2 | ('AU06AU07', 'AU04AU06AU07') | 246 |
| 2 | ('AU04AU06AU07', 'AU06AU07') | 361 | 2 | ('AU07', 'AU06AU07') | 198 |
| 2 | ('AU06AU07', 'AU04AU06AU07') | 353 | 2 | ('AU04', 'AU04AU07') | 186 |
| 2 | ('AU04AU06AU07', 'AU04AU07') | 349 | 2 | ('AU04AU07', 'AU04AU06AU07') | 177 |
| 3 | ('AU04AU07', 'AU04', 'AU04AU07') | 141 | 3 | ('AU04', 'Z', 'AU04') | 98 |
| 3 | ('AU04AU06AU07', 'AU06AU07', 'AU04AU06AU07') | 127 | 3 | ('AU06AU07', 'AU04AU06AU07', 'AU06AU07') | 91 |
| 3 | ('AU04AU06AU07', 'AU04AU07', 'AU04AU06AU07') | 123 | 3 | ('AU02AU01', 'AU02', 'Z') | 91 |
| 3 | ('AU04', 'AU04AU07', 'AU04') | 122 | 3 | ('AU04AU06AU07', 'AU06AU07', 'AU04AU06AU07') | 74 |
| 3 | ('AU04', 'Z', 'AU04') | 106 | 3 | ('AU07', 'Z', 'AU07') | 74 |
| 3 | ('AU07', 'Z', 'AU07') | 98 | 3 | ('AU04AU07', 'AU04', 'AU04AU07') | 60 |
| 3 | ('AU04AU06AU07', 'AU04AU07', 'AU04') | 73 | 3 | ('AU02', 'Z', 'AU02') | 60 |

Table 2.3: The most frequent n-grams found in questions and responses.

### 2.2.4 Questions-Response gap

Figure 2.4 shows the gap duration versus the total number of facial signals in the corresponding question. The relation between the gap duration and the total number of facial signals was evaluated by calculating the Spearman's correlation coefficient. A very weak negative correlation of $r_s = -.1171$ with $p = .0003$ was found.

A more specific analysis followed by investigating the correlations between specific facial signals and the gap duration. From Table 2.2 it followed that, although very small effect sizes, AU04, AU06, AU07, and AU09 occurred more frequently during questions than during responses. Thus for those facial signals as well, the Spearman's correlation coefficient was calculated between the occurrence count of the corresponding facial signal and the duration of the gap between the questions and responses. Since these facial signals occurred more during questions, they could help the addressee to predict information about the information of the turn content, or when the turn is going to end. Hence it could affect the duration of the gap. For none of the facial signals a correlation was found between the occurrence count and the duration of the gap. For AU04: $r_s = -.0081$, $p = .8055$; for AU06: $r_s = .0031$, $p = .9246$; for AU07: $r_s = -.0627$, $p = .0.0551$; for AU09: $r_s = .0225$, $p = .4914$.

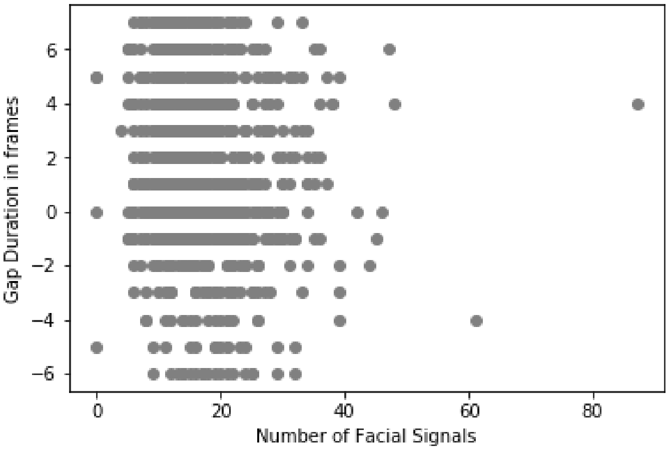


Figure 2.4: Scatter plot of the gap duration versus the total number of facial signals. The gap duration is represented in frames. The frames range from -6 till 6 which represents the interval from 250 before onset of the question until 250 after onset of the questions since the videos were 25 frames per second.

## 2.3 Discussion

This section servers as an interim discussion, for a more in depth discussion see the general discussion in chapter 4.

The goal of this research was to develop an analysis toolkit to provide more insight in facial signals during human conversation and to contribute towards unified model of communication in speech. The difference in behaviour as regards facial signals was investigated when comparing questions and responses that were collected from natural conversation in a face-to-face context. General occurrence counts and co-occurrences of various facial signals were presented and furthermore it was investigated whether certain patterns of facial signals occurred systematically. However, it should be noted that the results are not completely reliable as a result of the OpenFace output not yet being cleaned.

The general occurrence counts of facial signals were investigated when comparing questions and responses (Table 2.2). It was found that even before onset of the question, more nose wrinklers (AU09) occurred with questions than with responses. This could indicate that nose wrinklers serve as question markers. It was found that inner (AU01) and outer (AU02) parts of the eyebrows were being raised more in responses than in questions. This could reveal new insights suggesting that eyebrow movements can also be linked to responses and not only to questions. Brow lowerers (AU04) occurred more often during questions than during responses, agreeing with existing literature that eyebrow movements occur more during questions. Furthermore it was found that lid tighteners (AU07) and cheek raisers (AU06) also occurred more frequently during questions compared to responses. However, after further research (Figure 2.2), it was found that these found effects might be explained by the large sample size, rather than them being true effects.

In order to investigate whether certain sequences of facial signals occurred frequently either in questions or responses, the most frequent n-grams present in the data were extracted (Table 2.3). It was concluded that no meaningful sequences could be extracted before the data is cleaned. This consists of removing the false positives from the OpenFace output.

Finally it was investigated whether the occurrence of certain facial signals affected the gap duration between the questions and its corresponding responses. None of the tested facial signals showed a correlation with the gap duration. However, a very weak correlation was found between the total number of facial signals and the duration of the gap. Although the data contains a lot of false positives, this supports the idea that facial signals contribute to the addressee's prediction about the content and expected ending of the speaker's turn, allowing them to respond faster.

# Chapter 3

# SPUDNIG

This chapter will discuss the development of SPeeding Up the Detection of Non-iconic and Iconic Gestures (SPUDNIG): a toolkit for the automatic detection of hand movements and gestures in video data [Ripperda et al., 2019]. The prior chapters of this thesis focused on facial signals, which were recognized by OpenFace. However, since no such annotation toolkit exists for the detection of hand movements and gestures, a first attempt was made to create such toolkit. SPUDNIG was developed during this thesis project as a side project but did not contribute to answering the questions this thesis addresses. Therefore this chapter contains its own introduction, methods and results.

## 3.1 Introduction

An aspect of multimodal communication which until now has not been taken into account, in this thesis, are communicative hand gestures. Due to them being closely related to speech, manual gestures often have been the focus of multimodal research in domains of linguistics, neuroscience, anthropology and psychology [Goldin-Meadow, 2005, Kendon, 2004, McNeill, 1992]. Gestures can for example refer to objects, locations, events or ideas, but they can also be used to convey semantic information integral to the content of the speaker [Holler and Beattie, 2003, Holler and Wilkin, 2009, Hostetter, 2011, McNeill, 1992]. Previous studies have shown that this multimodal information is processed by the addressee, and that language comprehension is facilitated by it [Drijvers and Özyürek, 2017, Holler and Beattie, 2003, Kelly et al., 2004, Kelly et al., 1999, Özyürek, 2014, Kelly et al., 2010].

One of the main challenges in studies on multimodal communication is annotating the on- and offsets of those manual gestures, which is a time-consuming and labour-intensive task. Often annotation tools such as ELAN [Wittenburg et al., 2006] or ANVIL [Kipp, 2001] are used. Although such annotation tools make the annotating process considerably easier, they do not automatize or speed up the process in any way. This process consists of frame-by-frame analysis by trained researchers in order to determine the on- and offset of the gestures, and it involves multiple researchers to be able to establish inter-rater reliability. Especially for large corpora this can be extremely time-consuming. Therefore techniques that could automate this process would significantly advance research on multimodal communication.

Recently motion tracking systems such as Microsoft Kinect [Zhang, 2012] and Leap Motion (San Francisco, USA; `http://leapmotion.com`) have opened up possibilities for automatic analysis of movements. However, these techniques require you to purchase the corresponding hardware.

Examples of alternative video-based tracking systems that are able to automatically track movements or body keypoints, without being required to possess corresponding hardware, are OpenPose [Cao et al., 2018, Cao et al., 2017, Simon et al., 2017, Wei et al., 2016] and AlphaPose [Fang et al., 2017, Xiu et al., 2018]. Both systems solve the problem of pose estimation, which refers to a computer vision technique that detects human figures from video data, so that one could determine where certain body keypoints (e.g. wrist) are located. Although AlphaPose has demonstrated in their paper to outperform OpenPose at the task of pose estimation, SPUDNIG will use OpenPose for this task because OpenPose contains additional models to detect keypoints from the hands specifically, whereas AlphaPose only contains models for the body. It is required to track such movements from keypoints within the hand to be able to recognize more fine-grained finger movements, whereas when only body keypoints would be used such small movements for example could be missed because the body keypoint closest to the hand from AlphaPose (i.e. the wrist) would not move during such small movements. Such tools however only offer coordinates for certain keypoints from the body, rather than recognizing on- and offsets of gestures or movements.

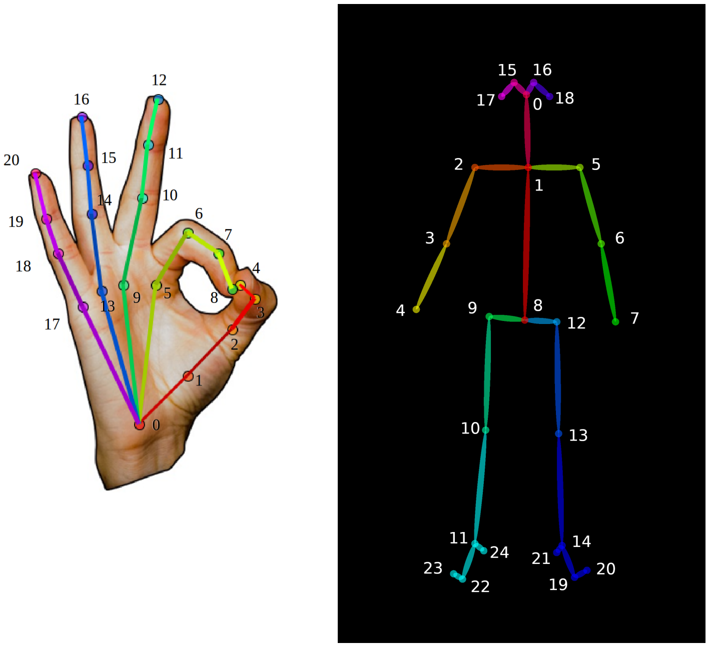
To overcome these limitations, SPUDNIG (SPeeding Up the Detection of Non-iconic and Iconic Gestures) was developed: a new open-source toolkit provided with an easy-to-use graphical user interface (GUI) for automatic detection of hand gestures and movements. Note 'detection' is used rather than 'recognition' because recognition would involve distinguishing which movements are gestures and which movements do not contribute to the message of the speaker. Similar to [De Beugher et al., 2018], detection is defined as distinguishing movement sequences from non-movement sequences. SPUDNIG uses OpenPose as input for continuous, video-based motion tracking of movements and gestures, and subsequently observes x/y coordinate changes of keypoints in the body and hands to automatically detect movement on- and offsets. Therefore SPUDNIG does not require motion capture hardware. The remainder of this chapter will discuss OpenPose, the development of SPUDNIG (i.e. how it detects movement vs. no movement), and a proof-of-principle and validation of SPUDNIG. Note that SPUDNIG is not created to eliminate the need of a human annotator entirely. The aim is to evaluate its overlap with annotations made by a human annotator, and to reduce the work of a human coder, limiting it to the removal of false positives (i.e. non-gestural movements).

## 3.2 Methods

SPUDNIG is completely developed in Python 3.7.0. It supports Windows and both GPU and CPU.

### 3.2.1 OpenPose

SPUDNIG uses the output created by OpenPose [Cao et al., 2018, Cao et al., 2017, Simon et al., 2017, Wei et al., 2016] to detect hand gestures and movements. OpenPose is capable of real-time multi-person 2D pose estimation from video data. Specifically, it uses deep learning convolutional neural networks (CNN) to recognize the location of specific keypoints (i.e. body parts) in video data. Figure 3.1 presents the keypoints that are recognized for both the points in the hand (Figure 3.1a) and in the body (Figure 3.1b). Great advantages of OpenPose are that it runs on different platforms (Windows, Mac OSX), supports both GPUs and CPUs, and that it offers a pretrained model, which means that no data needs to be annotated.



(a) 21 keypoints within the hand are recognized

(b) 25 keypoints within the body are recognized according to the Body_25 model

Figure 3.1: Keypoints OpenPose is able to recognize.

### 3.2.2 Movement/no-movement detection

Per frame a JSON file is generated by OpenPose containing the x- and y-coordinates for the 21 keypoints in the hand and 25 keypoints in the body. Note that OpenPose performs best when .avi files are used. Those JSON files are converted to three .csv files for the keypoints of the body, left-, and right-hand respectively. From theses files, a default set of 8 keypoints are selected to estimate the on- and off-sets of hand movements. The keypoints that are taken into account from the body (Figure 3.1b) are both wrists (i.e. 4 and 7) and the elbows (i.e. 3 and 6). Coordinate changes in those points mainly reflect large hand movements. From the hand, the keypoints that are taken into account are the tip of the thumb and the tip of the index finger (i.e. 4 and 8). These keypoints should be able to capture more fine-grained movements. Both the selection and the number of keypoints were the result of a careful piloting phase in which the trade-off between false positives and false negatives was optimized. It was found that adding more keypoints resulted in more false positives, whereas removing keypoints resulted in more false negatives.

SPUDNIG calculates whether movement is happening in the current frame for each keypoint separately. First SPUDNIG checks whether the reliability of OpenPose is above a certain threshold (default = 0.3, can be altered by the user). If the reliability is below the threshold, SPUDNIG assumes no movement is occurring in the respective frame and continues to the next frame. If the reliability threshold is met, SPUDNIG continues to determine whether the respective keypoint is part of a movement or part of a rest position.

Rest positions are established by checking the x- and y-coordinates of the corresponding keypoint over a span of 15 frames with the current frame being the midpoint (i.e. frame $i$-7 until frame $i$+7 are checked). If the x- and y-coordinates of these frames all differ less than 10 pixels with an overall certainty threshold of 0.7 (i.e. if 70% of the frames differ less than 10 pixels), SPUDNIG assumes that the keypoint in the current frame is part of a rest position, updates the current/last known rest position, and continues to the next frame. If the certainty threshold is not met, SPUDNIG continues to check if whether the current frame is part of a movement. It does so by checking whether the current frame differs more than 5 pixels from the last known rest position. If this is the case, SPUDNIG evaluates the next 5 pixels and checks if these also differ more than 5 pixels from the last known rest position, with a certainty threshold of 60% (i.e. 3 out of 5 pixel should differ 5 pixels). This extra check is performed to establish that the 5 pixel difference represents actual movement instead of falsely recognized keypoints or just slight shifting of such keypoints.

If the threshold is met (i.e. a movement has been initialized), SPUDNIG continues with searching for a rest position in the upcoming 300 frames ( 12 seconds with 25 fps). At this point SPUDNIG determines a movement has occurs from frame $i$ (i.e. the current frame) until the frame where the keypoint has returned to a rest position, and continues to search for new movements from this point forward.

During testing different ranges and numbers of parameters were tested (pixel difference, certainty thresholds). It was found that incrementing or decrementing those parameters would either cause more false positives or false negatives.

The above described process is repeated for each keypoint and results in a list indicating for each frame whether movement was detected or not. The resulting lists are then merged, which has the advantage that if a movement was not detected by one keypoint (e.g. because the reliability was too low), it could still be detected through another keypoint, minimizing false negatives.

### 3.2.3 Post-processing

During the post-processing phase movements smaller than 4 frames are removed in order to clean the data and minimize the number of false positives. Additionally, close consecutive movements (i.e. with 4 or less frames between them) are merged to account for small hold or pauses in movements. Based on the fps, the timing of the movements is calculated by converting the frame number to hh:mm:ss.ms format. This information is then used to generate a .csv file containing the start and end times of each movement, which is compatible with annotation tools such as ELAN [Wittenburg et al., 2006] and ANVIL [Kipp, 2001].

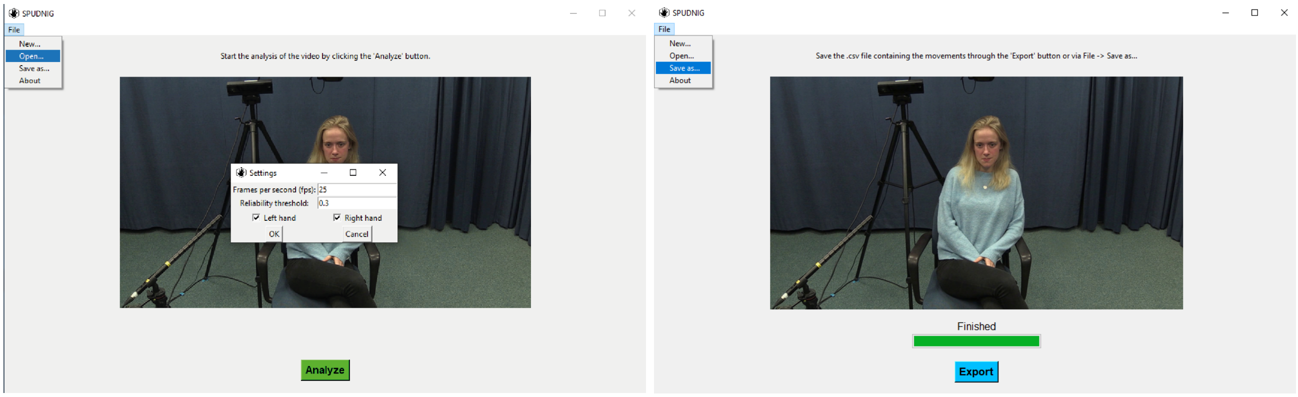
### 3.2.4 Graphical user interface

SPUDNIG comes with an easy-to-use GUI, which makes it accessible for people without technological knowledge or programming experience. The GUI is developed with Python's GUI toolkit Tkinter[1]. Figure 3.2 presents two screenshots of SPUDNIG in action.

As a first step the user should load a video through 'File → Open...' after which the analyze button will change colour from red to green, indicating the application is ready for analysis. Note that only OpenPose only works optimal with .avi files. Other, more complex video formats (e.g. MPEG) cause OpenPose to skip a non-fixed number of frames especially at the start of the video. Therefore, to prevent this disturbing the on- and off-sets of the hand movements, SPUDNIG will also only accept .avi files. Once the 'Analyze' button is clicked (Figure 3.2a), a settings screen will appear in which the user can alter the fps and the reliability threshold, and whether the left, right or both hands will be analyzed. If 'OK' is clicked in the settings screen the analysis will start and its progress will be communicated to the user through a progress bar. After the analysis is finished, the resulting .csv file can be saved by clicking the blue 'Export' button (Figure 3.2b) or through 'File → Save as...', which will prompt a window to select a desired location for saving the file.

The GUI is made fail-safe with the help of multi-threading. Multiple threads are used that monitor the user's activity as well as the state of the GUI. This way the GUI will not freeze or crash no matter which buttons are clicked.

[1] https://docs.python.org/2/library/tkinter.html



Start the analysis of the video by clicking the 'Analyze' button.

Settings

Frames per second (fps): 25

Reliability threshold: 0.3

☑ Left hand ☑ Right hand

OK Cancel

Save the .csv file containing the movements through the 'Export' button or via File -> Save as...

(a)

(b)

Figure 3.2: Screenshots of SPUDNIG GUI

### 3.2.5 Validation analyses

A validation analysis was performed by comparing the annotations of hand movements made by SPUDNIG to manual annotations by a trained researcher. Furthermore it was investigated how accurately SPUDNIG could detect iconic and non-iconic gestures. During the analyses data from the same corpus described in 1 and Figure 1.1. For the validation analyses, 20 samples of 2 minutes were used each from a different speaker.

A trained human researcher, blind to the purpose of the coding exercise, was asked to manually annotate the occurrence of all gestural movements that carried some form of meaning. This included iconic and metaphoric gestures (depicting aspects of abstract concepts, people, or actions), pragmatic gestures (including beats), deictic gestures, and interactive gestures (refer to the interlocutor rather than to the topic of conversation) [Bavelas et al., 1995, Bavelas et al., 1992, McNeill, 1992]. The annotations made by the human coder did not distinguish between those types of gestures. The annotator was asked to define the starting point of a gesture as the first frame where the hand had left its rest position and the end point as the final frame before the hand had returned to its rest position, or the last frame of the gesture stroke in case of successive gestures [Kita et al., 1997].

Gestures were annotated both in form-based coding and in meaning-based coding. In the form-based coding every stroke of a gesture was annotated as a separate movement, regardless of them depicting the same semantic meaning (e.g. a hammer gesture with successive strokes are annotated separately). In the meaning-based coding, individual (successive) strokes are annotated as one movement if they depict the same semantic concept. Researcher seem to use both approaches, hence the different coding schemes.

Furthermore, in the meaning-based coding, for all gestures a distinction was made between iconic and non-iconic gestures. This way it was tested whether iconic gestures might be more easily detected by SPUDNIG. Iconic gestures might for example be larger in size, or might contain less holds, which could result in easier detection, since SPUDNIG is not able to distinguish holds within a gesture from the hands being in an actual rest position.

To establish how the annotations made by SPUDNIG compare to human coding, the overlap between the them was calculated, for both form-based and meaning-based coding. A modified Cohen's kappa was calculated using EasyDIAg [Holle and Rein, 2015], a measure commonly used to establish agreement between two raters. EasyDIAg takes into account the categorization of the values, the temporal overlap of the annotations, and the segmentation of behaviour (e.g. on- and off-sets of annotations). An overlap criterion of 60% was used, meaning there should be a 60% temporal overlap between the annotations made by SPUDNIG and the human annotator.

Furthermore, to identify how many gestures identified by the human would also be captured by SPUDNIG, it was investigated how much movement detected by SPUDNIG was not gestural. All the output of SPUDNIG was compared to the output of the human annotator to investigate how many annotations by SPUDNIG did not overlap with a gesture annotation. This should indicate how much false positives SPUDNIG produces, meaning how much of SPUDNIG's output should be filtered out, and also how many gestures SPUDNIG would miss (false negatives).

### 3.2.6 Does SPUDNIG accelerate the annotation process?

Four additional human coders were asked to compare the time it takes to manually annotate the data compared to validating SPUDNIG's output by removing non-gestural movements, and by checking the on- and offsets of the annotations. The human coders were presented 20 2-minute snippets of videos where they were instructed to annotate hand gestures manually, and 20 different 2-minute snippets of videos where they were instructed to annotate gestural movements by validating SPUDNIG's output. In the first test phase human coders were not informed about how well SPUDNIG is able to capture hand movements. This meant that the human coders would look for both false negatives and positives, and misaligned on- and offsets. In the second phase the human coders were informed that SPUDNIG manages to capture all hand movements and they did not have to check for false negatives. The conditions and order of the videos were randomized.

## 3.3 Results

### 3.3.1 Form-based annotations

First the overlap between the annotations from SPUDNIG and the human annotator's form-based coding was compared. A raw agreement of 87% and a modified Cohen's Kappa maximum value of .86 were found, indicating a very high agreement [Cohen, 1960, Landis and Koch, 1977].

The manual analysis showed that out of the 207 gestures identified by the human annotator, SPUDNIG identified 206. Note that this method ignores the amount of overlap between the annotation from SPUDNIG and the human annotator, in contrast to the modified Cohen's Kappa value.

### 3.3.2 Meaning-based annotations

Next the overlap between the annotations from SPUDNIG and the human annotator's meaning-based coding was compared. Here a raw agreement of 86% was observed, together with a modified Cohen's Kappa maximum value of .77, indicating high agreement.

The manual analysis showed that from the 185 gestures identified by the human annotator, SPUDNIG identified 184.

### 3.3.3 Iconic gestures

Next the focus was on the iconic gestures. It was investigated how many meaning-based annotations were actually iconic gestures. 45 out of the 185 in total were iconic. The overlap between SPUDNIG and the human annotator was calculated and a raw agreement of 93%, and a modified Cohen's Kappa of 1 was observed, which indicates near perfect agreement [Cohen, 1960, Landis and Koch, 1977].

### 3.3.4 Non-iconic gestures

From the 185 meaning-based annotations, 140 of them were non-iconic. Again the overlap was compared between SPUDNIG and the human annotator and a raw agreement of 84%, and a modified Cohen's Kappa maximum value of .74 was found, indicating high level of agreement [Cohen, 1960, Landis and Koch, 1977].

### 3.3.5 Movement/gesture detection

Finally, it was investigated how many of SPUDNIG's annotation were actually not part of a gesture. Although SPUDNIG seems to detect gestural movements highly accurately, it is not capable of distinguishing gestural from non-gestural movements. The result is that SPUDNIG produces considerably more annotations than a human annotator would produce: SPUDNIG produces 311 annotations, of which 217 were non-gestural movements.

### 3.3.6 Acceleration

In the first phase the human coders were not informed about how well SPUDNIG is able to capture movements. This means the human coder looked for both false positives as well as false negatives. Per coder it was calculated how many milliseconds it took to annotate a millisecond of data. Descriptives of the outcomes of the test are presented, since the low number of annotators and the low number of videos per condition would not result in reliable statistical results. Human coders were quicker on average when using SPUDNIG (M = 76.8, SD = 83.4) than when manually annotating the videos (M = 81.2, SD = 77.06).

In the second phase the human coders were informed about how well SPUDNIG is able to capture movements, and asked to analyse another set of videos. During this phase human coders were almost twice as quick when using SPUDNIG (M = 19.25, SD = 12.63) as compared to manually annotating the data (M = 35.4, SD = 25.9).

## 3.4 Discussion

This chapter presented SPUDNIG: SPeeding Up the Detection of Non-iconic and Iconic Gestures, a toolkit for automatic detection of hand movements and gestures in video data. An easy-to-use graphical user interface is presented, and a proof-of-principle is provided, where the annotations of SPUDNIG are compared to form-based and meaning-based annotations of a trained human researcher.

The results showed that SPUDNIG can very accurately, over 99%, annotate the occurrence of hand gestures, for both form- and meaning-based coding, and for both iconic and non-iconic gestures. Although SPUDNIG manages to capture almost all gestures, it is noted that SPUDNIG produces an overabundance of annotations, based on non-gestural movements. However, SPUDNIG advances the most labour-intensive part of many multimodal communication research, the annotation process of manual gestures. Removing the false positives from SPUDNIG's output is comparatively a much faster and easier process compared to manually going through the video and detecting all gestures from scratch.

### 3.4.1 Performance

SPUDNIG achieved overall high to very high Cohen's Kappa scores [Cohen, 1960, Holle and Rein, 2015, Landis and Koch, 1977]. However, it should be noted that SPUDNIG does not differentiate between gestures and non-gestural movements, comparable to other semi-automatic annotation tools [De Beugher et al., 2018]. In the 20 times 2 minutes from different videos that were used, SPUDNIG produced 311 annotations, of which 217 included non-gestural movements, such as head scratches and fidgeting for example. Although the need of a human researcher is not eliminated, since a substantial part of the annotations made by SPUDNIG are false positives, it is demonstrated that SPUDNIG achieved very high gesture annotation overlap, thus speeding up the gesture annotation process significantly.

Overall, SPUDNIG managed to detect 184 gestures out of the 185 detected by the human researcher.

### 3.4.2 Limitations

One of SPUDNIG's limitations is its dependency on OpenPose. If OpenPose for some reason fails to recognize (some of) the keypoints, SPUDNIG will be unable to detect movements. Examples of why OpenPose would fail to detect certain keypoints are occlusion of certain body parts (e.g. the speaker puts his hands between his legs), or the video quality being too poor. For example, if the speaker would move his hands too fast, the frames would become blurry which lead to reliability drops for the corresponding keypoints.

A second limitation is that SPUDNIG is unable to differentiate between different types of gestures. Future work could for example use SPUDNIG's open source code to detect certain patterns in x- and y-coordinate changes and recognize recurrent gestures [Bressem and Müller, 2014, Bressem and Müller, 2017, Ladewig, 2011].

Third, SPUDNIG uses 2D data, but does not provide information about movements in three-dimensional space. Therefore it is less suited for studying more complex movement dynamics related to space or directionality [Trujillo et al., 2019].

A final limitation is that SPUDNIG uses pixel differences for detecting movements. An alternative option to analyze OpenPose's output would be to use millimeters as threshold, which can be achieved by using multiple cameras and the camera calibration procedure that is already integrated in OpenPose. This could result in a more robust threshold, independent of camera dimensions.

# Chapter 4

# General Discussion

This chapter first will discuss the main questions of this thesis, which try to identify regularities in facial signals during questions and responses in face-to-face communication. This is followed by a brief discussion of the development and validation of SPUDNIG, which was developed as a side project during this thesis, to facilitate and accelerate the annotation process of hand movements and gestures in video data. During the discussion of the results regarding the facial signals, it should however be noted that those results are not completely reliable as a result of the OpenFace output not being cleaned yet.

## 4.1 Regularities and their timing

The first two research questions were "Are there any regularities between speech acts (i.e. questions and responses) and facial signals (e.g. eyebrow raises might occur more in questions whereas lip tighteners might occur more in responses)?" and "What is the timing of the occurrences/sequences of facial signals with regards to questions and responses?". It was hypothesized that there would be regularities between questions and responses and certain facial signals. More specifically, based on [Chovil, 1991, Flecha-García, 2002], it was hypothesized that eyebrow movements occurred more during questions. Furthermore this thesis served as an exploratory study to provide more insights about more potential regularities as regards facial expressions during questions and responses.

First the general occurrence of facial signals was investigated where a distinction was made between questions and responses, and three different intervals, of which the results are shown in Table 2.2. It follows that, even before onset of the question, nose wrinklers (AU09) occur more frequently with questions than with responses. Nose wrinklers are often paired with emotions of disgust or anger [Jack et al., 2014]. This could indicate that nose wrinklers serve as an indicator for upcoming questions.

For AU01 and AU02, which represent raising the inner and outer parts of the eyebrows, are found to occur more frequently during responses, contradicting previous results from [Chovil, 1991, Flecha-García, 2002], who found that eyebrow movements occur more during questions than in responses. For outer brow raises there was only a statistical difference found towards the start of the response, whereas for inner brow raises it was more shifted towards the end of the response. Since outer brow raises are paired with emotions of surprise, this could indicate that a response to a possible unexpected question would start with an outer brow raiser. The fact that inner brow raises occurred more towards the end of responses, indicates that responses could be completed by inner brow raises, serving as an indicator that the turn is going to end for example.

These findings could indicate, although the effect sizes are really small, that brow movements could also be linked to responses, and not only to questions, hence extending existing literature.

Corresponding to existing literature that states eyebrow movements are paired with questions, it was found that eyebrow lowerers occurred more frequently in questions than in responses, independent of the timing in the question or response. Brow lowerers are found to be paired with emotions of confusion and frustration [Grafsgaard et al., 2011, Bosch et al., 2014]. These kind of emotions can also be followed by questions, in order to relieve the confusion or frustration, thus explaining why they occurred more in questions.

For AU07, which represents the eyelid tightener, it was found to occur more during questions, independent of the timing in the questions. However, when looking at Figure 2.2, it follows that this found effect might be a cause of the large sample size rather than it being a true effect. The effect only starts to reach the significance level when approximately 60% of the data is used, whereas when less is used the difference remains not significant.

A similar effect is observed for AU06, which represent the cheek raiser. Table 2.2 suggests that cheek raises occur more frequently in questions than in responses, when focusing on the end of the questions and responses. However, when looking at Figure 2.2c, it is observed that this effect is not significant until approximately 90% of the complete sample is used, indicating this effect might not be a robust effect but due too the large sample size.

Besides investigating general occurrence counts, it was also investigated which facial signals co-occurred during questions and responses. Figure 2.3 presents the results from the co-occurring facial signals. When comparing questions, Figure 2.3a, to the responses, Figure 2.3b, it is observed that both Figures have a similar distribution of intensities. It concludes that there's no difference in co-occurrence pairs of facial signals between questions and responses.

## 4.2 N-grams

The next question was "Are there any particular sequences of facial signals within questions/responses that occur systematically?". To investigate this, the most frequent n-grams were extracted from the data. Each unit of the n-grams consists of the facial signals occurring in the corresponding frame. Table 2.3 presents the most frequent n-grams with n = 2 and n = 3. Because of the messy output of OpenFace, most n-grams sometimes even contain 3 facial signals per unit, meaning that 3 facial signals would occur simultaneously. Therefore it is hard to interpret these results without cleaning the OpenFace data first. The most frequent sequence during questions is ('AU04', 'AU04AU07'), meaning that first the brows are being lowered (AU04), and somewhere during this process the eyelids are being tightened (AU07). Some more interpretable n-grams that are found are ('AU04', 'Z', 'AU04') during questions, and ('AU07', 'Z', 'AU07') during responses. Note that 'Z' indicates that no facial signals are occurring. The first n-gram represents the brows being lowered, followed by nothing, followed by the brows again being lowered. Note that due to the post-processing there is no information about how long the duration of the brow lowerer or the pause in-between. The second n-gram represents the eyelids being tightened, followed by nothing, followed by the eyelids being tightened again. These results indicate that consecutively lowering the eye brows occurs more during questions, agreeing with existing literature [Chovil, 1991], and with the results presented in Table 2.2. They also indicate that consecutively tightening the eyelids could be linked to responses.

In order to conclude more meaningful conclusions from these results, the OpenFace data will first need to be cleaned. It is believed that once the data, more clear and interesting patterns will emerge.

## 4.3 Gap duration

The final research question investigated the gap duration between corresponding questions and responses, and if the occurrence of certain facial signals during the question affected the gap duration between the question and its corresponding response. For none of the checked facial signals, a (significant) correlation was found between the occurrence count of the particular facial signal and the duration of the gap. This would indicate that the occurrence of none of the facial signals would lead to shorter gaps, hence no specific facial signals would contribute to the addressee's prediction about the content and expected ending of the incoming turn. However, a significant, very weak negative correlation was found between the total number of facial signals, indicating that more facial signals lead to a shorter gap duration between questions and responses. This yet indicates that the occurrence of facial signals in general does contribute to the aforementioned prediction. Further research could reveal which specific facial signals, if any, cause this effect, by cleaning OpenFace's data or investigating other specific facial signals.

## 4.4 SPUDNIG

SPUDNIG was developed in order to facilitate and accelerate the annotation process of hand movements and gestures, since this is such an important aspect of research on multimodal communication, and the process of manually annotating such movements and gestures is a very time-consuming task.

It was shown SPUDNIG is able to accurately annotate the occurrence of hand gestures, since SPUDNIG detected 184 hand gestures from the 185 gestures annotated by a trained human researcher. Although SPUDNIG manages to capture almost all gestures, it should be noted that it does not eliminate the use of a human researcher, since it also produces an overabundance of annotations. SPUDNIG is developed in such a manner that it focuses on capturing all movements. This way users do not have to inspect the complete video and look for gestures from scratch, but can go through the annotations created by SPUDNIG and adjust or remove them. It was shown that this is an easier and faster process.

Another great strength of SPUDNIG is that no knowledge of programming, no motion capture systems, not even GPU hardware (although this speeds up the process) is required, and it comes with an easy-to-use graphical user interface. All of this makes it accessible to a large community of users. Furthermore, since the code is open source it has potential for being developed further, with the ultimate goal of being a tool able to distinguish gestural from non-gestural movement. Although this distinction is far from SPUDNIG's current state, it does currently provide a possibility to the non-programming user to significantly facilitate the gesture annotation process.

# Chapter 5

# Conclusion

## 5.1 Facial signals

The aim of this thesis was to provide more insight in the facial signals accompanying questions and responses in face-to-face communication, and whether they contribute to the addressee's prediction on information about the content and expected ending of the speaker's turn. This eventually can be used to compose a unified model of communication in speech. Unlike most studies in this domain, this study used data of participants that engaged in dyadic conversations of free speech in an effort to take a more naturalistic approach, rather than having conversations in highly controlled environments. The questions and responses were annotated by trained human researcher of the Max Planck Institute (MPI) for Psycholinguistics in Nijmegen. The facial signals were recognized by OpenFace, an automatic facial signal detecting toolkit. Although the output of OpenFace contains quite some false positives, the goal of this thesis was to create an analysis tool for analyzing the facial signals together with questions and responses, rather than cleaning up the output of OpenFace. It is believed this caused the small effect sizes for some of the results, and also sometimes uninterpretable results. Statistical differences between the questions and responses were found, agreeing as well as contradicting existing literature. However, the results that were found could change completely once the output is cleaned, or agree and even magnify the found effects. The main conclusion of this thesis is that the data first needs to be cleaned in order to produce meaningful results, but that this thesis has succeeded in developing an analysis toolkit that is able to investigate different properties of the data regarding facial signals within question-response sequences. The cleaning of OpenFace's output is currently in progress by researchers at the MPI, and the analyses performed during this thesis can and will be re-used once the data is cleaned. Although the most important recommendation for future work has already been made, i.e., cleaning OpenFace's output, more recommendations will follow.

First, this research does not distinguish between different types of questions.

In [Couper-Kuhlen, 2012] it was shown that different intonation patterns are found when differentiating between polar, Wh-, declarative, tag and repeat questions. It is possible that different patterns or regularities of facial signals are found when distinguishing those different types of questions, or even that these different layers neutralize each other. Next, OpenFace also is able to extract information about the head pose and eye gaze direction, which in this research has been ignored. This information however does contain valuable information during communication [Hanna and Brennan, 2007, Staudte et al., 2014], so it could also be interesting to investigate how this differs in questions compared to responses. Furthermore, if cleaned data is used and more meaningful co-occurrences or even sequences of facial signals are found, these could be used in the gap duration research. Instead of testing whether the occurrence of a single facial signal affects the duration of the gap between questions and responses, one could check if the occurrence of multiple, or even sequences of facial signals affect the gap duration. Finally it is recommended that when investigating question-response sequences with large sample sizes, one should be cautious of small p-values, and always validate them by performing some kind of bootstrapping method.

This research has provided new opportunities with regards to investigating facial signals in question-response sequences. A possible implication would be to apply the results regarding which facial signals occur during questions and responses from this or future research to the field of robotics and virtual agents, in order to let them communicate in a more human-like manner.

## 5.2 SPUDNIG

The prior discussed research was possible due to OpenFace, which is able to recognize facial signals from video data, increasing the research possibilities regarding facial signals. On the other hand, this kind of toolkit is lacking for the automatic detection of hand gestures. This thesis has presented SPUDNIG: SPeeding Up the Detection of Non-iconic and Iconic Gestures, a first attempt to create such toolkit. It is demonstrated that SPUDNIG detects both iconic and non-iconic gestures highly accurately from video data. SPUDNIG comes with an easy-to-use graphical user interface and aims to speed up the process of annotating hand movements and gestures, rather than eliminating the need of a human researcher. It is presented that removing the false positives from SPUDNIG is an easier process compared to manually going through the full video and annotating all gestures from scratch. The source code is open source and it is encouraged to try and improve the toolkit.

# Bibliography

[Baltrusaitis et al., 2015] Baltrusaitis, T., Mahmoud, M., and 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)*, volume 6, pages 1–6. IEEE.

[Baltrusaitis et al., 2013] Baltrusaitis, T., Robinson, P., and Morency, L.-P. (2013). Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 354–361.

[Baltrusaitis et al., 2018] Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE.

[Bavelas et al., 1995] Bavelas, J. B., Chovil, N., Coates, L., and Roe, L. (1995). Gestures specialized for dialogue. *Personality and social psychology bulletin*, 21(4):394–405.

[Bavelas et al., 1992] Bavelas, J. B., Chovil, N., Lawrie, D. A., and Wade, A. (1992). Interactive gestures. *Discourse processes*, 15(4):469–489.

[Bellegarda, 2000] Bellegarda, J. R. (2000). Large vocabulary speech recognition with multispan statistical language models. *IEEE Transactions on Speech and Audio Processing*, 8(1):76–84.

[Benitez-Quiroz et al., 2016] Benitez-Quiroz, C. F., Wilbur, R. B., and Martinez, A. M. (2016). The not face: A grammaticalization of facial expressions of emotion. *Cognition*, 150:77–84.

[Borras-Comes et al., 2014] Borras-Comes, J., Kaland, C., Prieto, P., and Swerts, M. (2014). Audiovisual correlates of interrogativity: A comparative analysis of catalan and dutch. *Journal of Nonverbal Behavior*, 38(1):53–66.

[Bosch et al., 2014] Bosch, N., Chen, Y., and D'Mello, S. (2014). It's written on your face: detecting affective states from facial expressions while learning computer programming. In *International Conference on Intelligent Tutoring Systems*, pages 39–44. Springer.

[Bressem and Müller, 2014] Bressem, J. and Müller, C. (2014). The family of away gestures: Negation, refusal, and negative assessment. *Body–language–communication: An international handbook on multimodality in human interaction*, 2:1592–1604.

[Bressem and Müller, 2017] Bressem, J. and Müller, C. (2017). The "negative-assessment-construction"–a multimodal pattern based on a recurrent gesture? *Linguistics Vanguard*, 3(s1).

[Cao et al., 2018] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

[Cao et al., 2017] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*.

[Cassell et al., 2001] Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., and Rich, C. (2001). Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 114–123. Association for Computational Linguistics.

[Cavé et al., 2002] Cavé, C., Guaïtella, I., and Santi, S. (2002). Eyebrow movements and voice variations in dialogue situations: an experimental investigation. In *Seventh International Conference on Spoken Language Processing*.

[Chovil, 1991] Chovil, N. (1991). Discourse-oriented facial displays in conversation. *Research on Language & Social Interaction*, 25(1-4):163–194.

[Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

[Couper-Kuhlen, 2012] Couper-Kuhlen, E. (2012). Some truths and untruths about final intonation in conversational questions. In *Questions: Formal, functional and interactional perspectives*. Cambridge University Press.

[De Beugher et al., 2018] De Beugher, S., Brône, G., and Goedemé, T. (2018). A semi-automatic annotation tool for unobtrusive gesture analysis. *Language Resources and Evaluation*, 52(2):433–460.

[DeLong et al., 2005] DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, 8(8):1117.

[Drijvers and Özyürek, 2017] Drijvers, L. and Özyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research*, 60(1):212–222.

[Ekman, 2004] Ekman, P. (2004). Emotional and conversational nonverbal signals. In *Language, knowledge, and representation*, pages 39–50. Springer.

[Ekman, 2009] Ekman, P. (2009). *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition).* WW Norton & Company.

[Ekman and Friesen, 1978] Ekman, P. and Friesen, W. V. (1978). Facial action coding system consulting psychologists press. *Palo Alto, CA.*

[Fang et al., 2017] Fang, H.-S., Xie, S., Tai, Y.-W., and Lu, C. (2017). RMPE: Regional multi-person pose estimation. In *ICCV*.

[Federmeier and Kutas, 1999] Federmeier, K. D. and Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of memory and Language*, 41(4):469–495.

[Flecha-García, 2002] Flecha-García, M. L. (2002). Eyebrow raising and communication in map task dialogues. *Gesture: The Living Medium.*

[Flecha-García, 2010] Flecha-García, M. L. (2010). Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in english. *Speech communication*, 52(6):542–554.

[Goldin-Meadow, 2005] Goldin-Meadow, S. (2005). *Hearing gesture: How our hands help us think*. Harvard University Press.

[Grafsgaard et al., 2011] Grafsgaard, J. F., Boyer, K. E., and Lester, J. C. (2011). Predicting facial indicators of confusion with hidden markov models. In *International Conference on Affective computing and intelligent interaction*, pages 97–106. Springer.

[Hanna and Brennan, 2007] Hanna, J. E. and 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.

[Hesch and Roumeliotis, 2011] Hesch, J. A. and Roumeliotis, S. I. (2011). A direct least-squares (dls) method for pnp. In *2011 International Conference on Computer Vision*, pages 383–390. IEEE.

[Hirsimaki et al., 2009] Hirsimaki, T., Pylkkonen, J., and Kurimo, M. (2009). Importance of high-order n-gram models in morph-based speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):724–732.

[Holle and Rein, 2015] Holle, H. and Rein, R. (2015). Easydiag: A tool for easy determination of interrater agreement. *Behavior research methods*, 47(3):837–847.

[Holler and Beattie, 2003] Holler, J. and Beattie, G. (2003). How iconic gestures and speech interact in the representation of meaning: Are both aspects really integral to the process? *Semiotica*, 146:81–116.

[Holler et al., 2016] Holler, J., Kendrick, K. H., Casillas, M., and Levinson, S. C. (2016). *Turn-taking in human communicative interaction*. Frontiers Media.

[Holler et al., 2018] Holler, J., Kendrick, K. H., and Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic bulletin & review*, 25(5):1900–1908.

[Holler and Levinson, 2019] Holler, J. and Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*.

[Holler and Wilkin, 2009] Holler, J. and Wilkin, K. (2009). Communicating common ground: How mutually shared knowledge influences speech and gesture in a narrative task. *Language and cognitive processes*, 24(2):267–289.

[Hostetter, 2011] Hostetter, A. B. (2011). When do gestures communicate? a meta-analysis. *Psychological bulletin*, 137(2):297.

[Indefrey and Levelt, 2004] Indefrey, P. and Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92(1-2):101–144.

[Jack et al., 2014] Jack, R. E., Garrod, O. G., and Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology*, 24(2):187–192.

[Katz, 1987] Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3):400–401.

[Kelly et al., 1999] Kelly, S. D., Barr, D. J., Church, R. B., and Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of memory and Language*, 40(4):577–592.

[Kelly et al., 2004] Kelly, S. D., Kravitz, C., and Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and language*, 89(1):253–260.

[Kelly et al., 2010] Kelly, S. D., Özyürek, A., and Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, 21(2):260–267.

[Kendon, 2004] Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

[Kipp, 2001] Kipp, M. (2001). Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*.

[Kita et al., 1997] Kita, S., Van Gijn, I., and Van der Hulst, H. (1997). Movement phases in signs and co-speech gestures, and their transcription by human coders. In *International Gesture Workshop*, pages 23–35. Springer.

[Kluyver et al., 2016] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S., et al. (2016). Jupyter notebooks-a publishing format for reproducible computational workflows. In *ELPUB*, pages 87–90.

[Kouloumpis et al., 2011] Kouloumpis, E., Wilson, T., and Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! In *Fifth International AAAI conference on weblogs and social media.*

[Ladewig, 2011] Ladewig, S. H. (2011). Putting the cyclic gesture on a cognitive basis. *CogniTextes. Revue de l'Association française de linguistique cognitive*, (Volume 6).

[Landis and Koch, 1977] Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

[Levinson, 2016] Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. *Trends in cognitive sciences*, 20(1):6–14.

[Levinson and Torreira, 2015] Levinson, S. C. and Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731.

[Lin et al., 2013] Lin, M., Lucas Jr, H. C., and Shmueli, G. (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research*, 24(4):906–917.

[McKinney, 2011] McKinney, W. (2011). pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 14.

[McNeill, 1992] McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* University of Chicago press.

[Özyürek, 2014] Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651):20130296.

[Pagliardini et al., 2017] Pagliardini, M., Gupta, P., and Jaggi, M. (2017). Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507.*

[Pak and Paroubek, 2010] Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.

[Pickering and Garrod, 2013] Pickering, M. J. and Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and brain sciences*, 36(4):329–347.

[Ripperda et al., 2019] Ripperda, J., Drijvers, L., and Holler, J. (2019). Speeding up the detection of non-iconic and iconic gestures (spudnig): a toolkit for the automatic detection of hand movements and gestures in video data.

[Sacks et al., 1978] Sacks, H., Schegloff, E. A., and Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

[Santos et al., 2009] Santos, I., Penya, Y. K., Devesa, J., and Bringas, P. G. (2009). N-grams-based file signatures for malware detection. *ICEIS (2)*, 9:317–320.

[Sendra et al., 2013] Sendra, V. C., Kaland, C., Swerts, M., and Prieto, P. (2013). Perceiving incredulity: The role of intonation and facial gestures. *Journal of Pragmatics*, 47(1):1–13.

[Sidorov et al., 2014] Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., and Chanona-Hernández, L. (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3):853–860.

[Simon et al., 2017] Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*.

[Staudte et al., 2014] Staudte, M., Crocker, M. W., Heloir, A., and Kipp, M. (2014). The influence of speaker gaze on listener comprehension: Contrasting visual versus intentional accounts. *Cognition*, 133(1):317–328.

[Stivers et al., 2009] Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

[Trujillo et al., 2019] Trujillo, J. P., Vaitonyte, J., Simanova, I., and 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.

[Van Berkum et al., 2005] Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3):443.

[Wang et al., 2007] Wang, X., McCallum, A., and Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 697–702. IEEE.

[Wei et al., 2016] Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. In *CVPR*.

[Welch, 1947] Welch, B. L. (1947). The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

[Wittenburg et al., 2006] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.

[Wood et al., 2015] Wood, E., Baltrusaitis, T., Zhang, X., Sugano, Y., Robinson, P., and Bulling, A. (2015). Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3756–3764.

[Xiu et al., 2018] Xiu, Y., Li, J., Wang, H., Fang, Y., and Lu, C. (2018). Pose Flow: Efficient online pose tracking. In *BMVC*.

[Young, 1996] Young, S. (1996). Large vocabulary continuous speech recognition: A review. *IEEE signal processing magazine*, 13(5):45–57.

[Zadeh et al., 2017] Zadeh, A., Chong Lim, Y., Baltrusaitis, T., and Morency, L.-P. (2017). Convolutional experts constrained local model for 3d facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2519–2528.

[Zhang et al., 2016] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

[Zhang, 2012] Zhang, Z. (2012). Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10.