Methodology for Studying Emotional, Conversational, and Linguistic Facial Expressions

Carla Luisa de Oliveira Viegas

CMU-LTI-24-021

December 2024

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

Jeffrey Bigham, Chair
Bhiksha Raj
Fernando de La Torre
João Magalhães (Universidade NOVA de Lisboa)
Nuno Correia (Universidade NOVA de Lisboa)
Teresa Romão (Universidade NOVA de Lisboa)
Luisa Coheur (Universidade de Lisboa)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Language and Information Technology.

Keywords: emotion recognition, facial expressions, sign language, conversational expressions, stress, enthusiasm, adjectives

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTMENT OF
COMPUTER SCIENCE

METHODOLOGY FOR STUDYING EMOTIONAL, CONVERSATIONAL, AND LINGUISTIC FACIAL EXPRESSIONS

CARLA LUISA DE OLIVEIRA VIEGAS

M.Sc. in Medical Engineering

DOCTORATE IN COMPUTER SCIENCE

NOVA University Lisbon
December, 2024

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTMENT OF
COMPUTER SCIENCE

METHODOLOGY FOR STUDYING EMOTIONAL, CONVERSATIONAL, AND LINGUISTIC FACIAL EXPRESSIONS

CARLA LUISA DE OLIVEIRA VIEGAS

M.Sc. in Medical Engineering

Adviser: João Miguel da Costa Magalhães
Full Professor, NOVA University Lisbon

Co-adviser: Jeffrey Philip Bigham
Associate Professor, Carnegie Mellon University

DOCTORATE IN COMPUTER SCIENCE

NOVA University Lisbon
December, 2024

Methodology for Studying Emotional, Conversational, and Linguistic Facial Expressions

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.1.18) [158].

To my Lord and Savior who gave me strength when I had none and to my family for their support, patience, and endurance.

Acknowledgements

I would like to thank the following people, without whom I would not have been able to complete this research, and without whom I would not have made it through my doctoral degree!

I am thankful to my supervisors João Magalhães, Alexander Hauptmann, and Jeffrey Bigham, and all the committee members for their guidance and support. Thank you to Robert Frederking, Jamie Callan, Stacey Young, Nuno Correia, and Carla Ferreira for their help to overcome the challenges in coordinating requirements in dual degree programs. I thank my colleagues and friends at Carnegie Mellon University and NOVA University Lisbon for all the conversations and for making it fun to learn and grow together. Thank you to the Foundation for Science and Technology in Portugal for their funding and the opportunity to study at Carnegie Mellon University and NOVA University Lisbon.

And my biggest thanks is to my family. I thank my husband, Christopher who endured with me all these years of work and sacrifice, my sons Emanuel and Jacob for their patience and sacrifice, especially before conference deadlines, my parents for all the years they have taken care of my children to make it possible for me to work. I am thankful to my brothers and sisters of the body of Christ who supported me with their prayers and fasting, and of course to my Lord Jesus and his Blessed Mother for providing me strength when I had none.

Abstract

With the inception of affective computing in the late 1990s, a new research field emerged to bring human affect recognition and generation into human-machine interaction. Although various types of affect have been studied using different modalities, the majority of the work over the last 20 years has focused on recognizing the basic emotions defined by Paul Ekman. Nevertheless, to achieve the goal of creating emotionally intelligent dialogue systems that can have engaging conversations, the focus needs to move away from basic emotions and towards more fine-grained expressions. As human communication is rich in different facial expressions, obtaining large datasets for one specific expression is time-consuming and costly. In this work, I present a data bootstrapping methodology that combines automatic methods and human annotation to reduce the cost and time of creating annotated data. This methodology allows to gain insights into emotional, conversational, and linguistic facial expressions, which I show in the following domains: stress, enthusiasm, and adjectives in sign language. In all cases, facial action units from the Facial Action Coding System were automatically detected and statistically evaluated. In addition, characteristic facial movements of each expression were quantitatively evaluated through user studies.

We hope that other researchers will follow up on exploring the expressions presented in this work as well as unstudied ones using our data bootstrapping methodology, as they can have a meaningful impact on the use of pedagogical agents, sign language translating agents, and rapport-building with agents.

Keywords: data bootstrapping, facial expressions, stress, enthusiasm, sign language, emotions

Resumo

Com o surgimento da computação afetiva no final da década de 1990, um novo campo de pesquisa emergiu para trazer o reconhecimento e a geração de afeto humano para a interação homem-máquina. Embora vários tipos de afeto tenham sido estudados utilizando diferentes modalidades, a maioria dos trabalhos nos últimos 20 anos tem se concentrado no reconhecimento das emoções básicas definidas por Paul Ekman. No entanto, para alcançar o objetivo de criar sistemas de diálogo emocionalmente inteligentes que possam ter conversas envolventes, é necessário desviar o foco das emoções básicas para expressões mais refinadas. Como a comunicação humana é rica em diferentes expressões faciais, obter grandes conjuntos de dados para uma expressão específica é um processo demorado e caro. Neste trabalho, apresento uma metodologia de bootstrap de dados que combina métodos automáticos e anotação humana para reduzir o custo e o tempo de criação de dados anotados. Essa metodologia permite obter insights sobre expressões faciais emocionais, conversacionais e linguísticas, que apresento nos seguintes domínios: estresse, entusiasmo e adjetivos em linguagem de sinais. Em todos os casos, as unidades de ação facial do Sistema de Codificação de Ação Facial (Facial Action Coding System) foram detectadas automaticamente e avaliadas estatisticamente. Além disso, os movimentos faciais característicos de cada expressão foram avaliados quantitativamente por meio de estudos com usuários.

Esperamos que outros pesquisadores deem continuidade à exploração das expressões apresentadas neste trabalho, bem como de outras ainda não estudadas, utilizando nossa metodologia de bootstrap de dados, pois elas podem ter um impacto significativo no uso de agentes pedagógicos, agentes de tradução de linguagem de sinais e na construção de rapport com agentes.

Palavras-chave: data bootstrapping, expressões faciais, stress, entusiasmo, linguagem gestual, emoções

Contents

List of Figures ix

List of Tables xiv

1 Introduction 1

1.1 Background . . . 2

1.2 Existing limitations and challenges . . . 4

1.3 Research Aim, Objectives, and Question . . . 6

1.4 Significance . . . 7

1.5 Contributions . . . 8

1.6 Other Contributions . . . 11

1.7 Thesis Structure Outline . . . 11

2 Background 13

2.1 Definition of Facial Expressions . . . 15

2.1.1 Facial Muscles . . . 16

2.1.2 Face Measuring Systems . . . 17

2.2 Facial Expressions and Emotions . . . 21

2.2.1 Definition of Emotion . . . 22

2.2.2 Emotion Theory . . . 22

2.2.3 Emotion Representation Models . . . 27

2.3 Facial Expressions during Stress . . . 30

2.3.1 Definition of Stress . . . 30

2.3.2 Importance of Stress detection . . . 31

2.3.3 Methods for Stress detection . . . 31

2.4 Facial Expressions during Enthusiastic Presentations . . . 35

2.4.1 Conversational Expressions . . . 35

2.4.2 Enthusiasm . . . 36

2.4.3 Charisma . . . 37

2.5 Facial Expressions in Sign Languages . . . 39

2.5.1 Overview 40
2.5.2 Roles of Facial Expressions in Sign Languages 42
2.5.3 Sign Language Notation Systems 43
2.5.4 Sign Language Datasets with Facial Expressions 43
2.5.5 Automatic Sign Language Translation with Facial Expressions . . 44

3 Data Bootstrapping for Facial Expression Analysis 45
3.1 Ideal Method 45
3.2 Generation of simulated Data 46
3.3 Actor-Posed Datasets 46
3.4 In-The-Wild Datasets 47
3.5 My Method on Data Bootstrapping for Facial Expression Dataset Creation 47

4 Facial Expressions during Cognitive Stress 49
4.1 Introduction 49
4.2 Data 50
4.2.1 Recording setup 50
4.2.2 Experimental Protocol 51
4.3 Facial Action Units for Stress Detection 51
4.3.1 Methodology 52
4.3.2 Results 56
4.3.3 Conclusion 57
4.4 Facial Action Units during Cognitive Stress 58
4.4.1 Methodology 59
4.4.2 Results 63
4.4.3 Discussion 65
4.4.4 Conclusion and Future Work 66

5 Facial Expressions during Enthusiastic Presentations 67
5.1 Introduction 67
5.2 Entheos Dataset 68
5.2.1 Data Acquisition 68
5.2.2 Label Selection and Temporal Granularity 68
5.2.3 Data Annotation Protocol 70
5.2.4 Final Data Selection 71
5.2.5 Data Statistics 72
5.3 Computational Experiments 74
5.3.1 Features 74
5.3.2 Model Architecture 75
5.4 Results 75
5.4.1 Predicting Enthusiasm level 75
5.4.2 Evaluating the Effect of Discourse Features 76

5.4.3 Investigating Visual Features . . . 78
5.4.4 Investigating Acoustic Features . . . 79
5.4.5 AU Statistical Tests . . . 79
5.4.6 Prosody Statistical Tests . . . 81
5.5 Discussion and Conclusion . . . 81

6 Facial Expressions in Sign Language 83
6.1 Introduction . . . 84
6.2 Facial Action Units for Sign Language Generation . . . 84
6.2.1 RWTH PHOENIX14T Sign Language Dataset . . . 85
6.2.2 Dual Encoder Transformer for Sign Language Generation . . . 85
6.2.3 Computational Experiments . . . 87
6.2.4 Results . . . 88
6.2.5 Discussion and Conclusion . . . 89
6.3 Facial Action Units during Sign Language Modifiers . . . 91
6.3.1 Sign Language Facial Expression Datasets . . . 92
6.3.2 Facial Action Unit Analysis . . . 97
6.3.3 Results . . . 97
6.3.4 Discussion . . . 98
6.3.5 Limitations . . . 98
6.3.6 Future Work . . . 100
6.3.7 Conclusion . . . 100

7 Conclusion 101

Bibliography 103

Appendices

A Dataset Contributions 126
A.1 Stress Dataset . . . 126
A.2 Entheos Dataset . . . 126
A.3 Sign Language Dataset . . . 127

List of Figures

1.1 From left to right: an example from the stress, enthusiasm, and German Sign Language datasets on facial expressions. 2

1.2 Data-bootstrapping method for facial expression analysis, composed of 5 steps: take an existing dataset, extract relevant excerpts through human annotation or algorithms, extract FACS features, train a classifier to verify if patterns are found, and perform statistical analysis and/or clustering combined with human annotation to identify facial patterns. 6

2.1 Charles le Brun, The Expressions, from the Trait é des Passions. Public Domain. 15

2.2 Facial expressions triggered by electric stimulation. Fig. 4, p. 277 from Mécanisme de la Physionomie Humaine by Guillaume Duchenne, 1862 [63]. Public Domain. 16

2.3 Overview of facial muscles. Although more than 40 facial muscles exist, it is important to study facial behavior based on functional and not structural anatomy. E.g. the corrugator muscle group is comprised of three muscles that usually act together when innervated, bringing the brows down and together. The frontalis muscle, on the other hand, is a single muscle spanning the forehead, but the inner and outer parts of this muscle can be moved independently of each other, allowing for just the inner or outer corners of the eyebrows to rise. Graphics from [198] 17

2.4 Examples from different Facial Action Units (AUs) [87] from the lower face relevant to the generation of mouthings in sign languages. AUs can occur with different intensity values between 0 and 5. AUs have been used in psychology and in affective computing to understand emotions expressed through facial expressions. Image from [275]. 18

2.5 Basic universal emotions defined by Paul Ekman [103] 28

2.6 Overview of compound facial expressions of emotion [61]. (A) neutral, (B) happy, (C) sad, (D) fearful, (E) angry, (F) surprised, (G) disgusted, (H) happily surprised, (I) happily disgusted, (J) sadly fearful, (K) sadly angry, (L) sadly surprised, (M) sadly disgusted, (N) fearfully angry, (O) fearfully surprised, (P) fearfully disgusted, (Q) angrily surprised, (R) angrily disgusted, (S) disgustedly surprised, (T) appalled, (U) hatred, and (V) awed. 29

2.7 Overview of four main categories of stress detection methods defined in [104]. 32

2.8 An enthusiastic sample from the Entheos dataset, showing aligned video frames, audio, and text. 35

2.9 Samples from the MPI Facial Expression Database [131]. From left to right, the pictures show an actor expressing conversational expressions on imagining something negative and positive, being impressed, and being insecure. . . . 37

2.11 Grammatical Facial Expressions of Homonomy (GEH) are essential to distinguish the meaning of signs with the same manual gesture. In a) the signs LAWYER and CRAZY are performed with the same manual gesture in Brazilian Sign Language (LIBRAS) but in the eyes, eyebrows, and the open mouth [252]. Similarly, in b) CANNOT (ability) and CANNOT (permission) in Hong Kong Sign Language (HKSL) are executed with the same manual sign but differ in FEs [270]. 40

2.12 Grammatical Facial Expressions of Sentence (GES) define the sentence type, such as WH-questions (WHQ), Yes/No questions (YNQ), doubt questions, rhetorical questions (RHQ), topics, negatives (NEG), affirmatives, conditional clauses (COND), focus and relative clauses. The figure shows how GES are used in a) Brazilian Sign Language (LIBRAS) [253], b) American Sign Language (ASL) [281], c) Irish Sign Language (ISL) [230], and d) Hong Kong Sign Language (HKSL) [270]. 41

2.13 Grammatical Facial Expressions of Intensity (GEI) take the role of quantifiers. Above in LIBRAS the same sign EXPENSIVE is accompanied by different FEs which allow us to distinguish between LITTLE EXPENSIVE, EXPENSIVE, and VERY EXPENSIVE [251]. 42

2.14 Different notations for sign language. HamNoSys [106] represents the execution of the signs in terms of the manual phonological components but lacks facial movement description. SignWriting [267] includes information about the execution of facial and manual movements. Frequently used gloss annotation lacks information about the execution of the signs. 44

3.1 Example of how simulated facial expression can be created by controlling which AUs should be activated and with which intensity over time. The annotator is asked to select an emotion that is recognized from the simulation [44]. . . . 46

3.2 Dataset examples depicting actors who posed given emotions (a) and conversational expressions (b). Annotators are asked to select the emotion or conversational expression they recognize out of a list or through an open answer. 47
3.3 Examples of the AFEW in-the-wild dataset on basic emotions. Samples are extracted from TV shows and movies and annotated with the basic emotions to obtain ground truth labels. 47
3.4 Methodology to study emotional, conversational, and linguistic facial expressions through data bootstrapping. It comprises five steps: taking an existing dataset, dataset adaption, feature extraction, classification, and statistical analysis or clustering. 48

4.2 Extracted frame from one video during typing phase. 50
4.3 Phases of the experiment recorded on video. 51
4.4 Label distribution for different classification problems. 53
4.5 Plot showing AU2, AU4, AU20 of one subject during the entire video. The start and ending of each phase are marked by vertical lines. In between the main phases, the subjects were given a short break. However, the behavior of the shown AUs was different for different subjects. Nevertheless, correlations of different AUs with the different phases were visible. 54
4.6 Visualization of the data of one subject using tSNE. 55
4.7 Confusion matrices for subject-wise classification using Random Forest algorithm on a) 6 classes, b) 4 classes, and c) 2 classes. 58
4.8 Overview of our proposed analysis methodology. Our pipeline starts with five-minute videos of 115 subjects. A random selection of 1000 frames per subject follows together with K-means clustering for each individual's data. The centroids (video frames) of each individual's clusters are selected for our user study. Only the centroids with more than 60% "Yes" annotations were selected for manual clustering of the faces. At the end of our pipeline, we obtained seven facial activity patterns of stress. 59
4.9 Examples of clusters obtained by performing clustering on the data of each individual separately. The figure shows the impact of the silhouette score on the differences shown in the clusters. Clusters with low silhouette score (0 to 0.2) show no visible differences between the AUs. Clusters with medium silhouette score (0.2 to 0.4) show two visually distinct clusters and clusters with high silhouette score (> 0.4) show visual differences among all clusters. . . . 62
4.10 Results of manual clustering of samples identified as showing a stressed person. On the right of each sample, the mean of the z-normalized AUs intensities is shown with boxplots. The AUs with the highest means are indicated on the samples of the clusters, together with the basic emotions that show the activity of the same AU. 64

5.2 Layout of the annotation interface. On the top left is the sample to be annotated and below are the different labels: perceived gender, enthusiasm, and emphasis. On the top center is the option to mark the sample as noisy if laughter or clapping is present. On the right side are reference samples for the three different levels of enthusiasm. 72

5.3 From top to bottom: Label distribution in our train-test split, among perceived gender, and ratings given by TED viewers. Top: The training set and testing set reflect the same imbalance of class labels. Center: Female speakers have proportionally fewer monotonous samples and more normal samples than males, but the same proportion of enthusiastic samples. Bottom: Samples labeled as enthusiastic have been mainly rated as fascinating, persuasive, and inspiring. They have rarely been rated negatively. 73

5.4 An overview of our proposed multimodal dataset and model for predicting levels of enthusiasm using different features extracted from video, audio, and text. 75

5.5 Label distribution of different enthusiasm levels in relation to discourse relations. In (a), most samples had no PDTB relation, however, there is a visible difference between monotonous and enthusiastic samples in the occurrence of temporal and contingency relations. In (b), RST relations show that enthusiastic samples compared to monotonous samples use more elaboration, attribution, and joint relations. 77

5.6 Label distribution of different enthusiasm levels in relation to facial action units. In (a), we can see that monotonous samples have more often low intensities for AU26 (jaw drop) than enthusiastic samples. (b) shows that monotonous samples have mostly very low standard deviation for AU12 (lip corner puller), but enthusiastic samples have a higher standard deviation. 78

5.7 Label distribution of different enthusiasm levels in relation to acoustic features. In (a), we can see that enthusiastic samples have a higher mean F0 (pitch) compared to monotonous samples. (b) shows that monotonous speech tends to have lower mean loudness compared to enthusiastic speech. 79

5.8 Label distribution of enthusiastic and non-enthusiastic samples in relation to the (a) mean AU 17 (p-value = 0.0028), mean AU 17 (p-value = 0.9715), (c) mean F0 (p-value = 0.0), (d) std F0 (p-value = 0.0), (e) mean loudness (p-value = 0.0034), (f) std loudness (p-value = 0.00). 82

6.2 Comparison of the ground truth and the generated poses with our proposed dual encoder model for the gloss annotations CLOUD and RAIN. The upper example shows that the predictions captured the correct hand shape, orientation, and movement of the sign CLOUD. In the lower example it is visible that the predictions captured the repeating hand movement meaning RAIN. Although at first glance the hand orientation seems not correct, it is a slight variation which still is correct. 90

6.3 Examples in which our model failed to generate the correct phonology of signs. Example 1 depicts inaccuracies in hand shape, orientation, and movement. Example 2 shows the difficulty of the model to capture pointing hand shapes. 91

6.4 Different samples from the DGS Modifier Facial Expression Dataset for different weather nouns and modifiers. Samples for the category "light"show frequently raised eyebrows compared to samples without modifiers (center). In the category "strong"similar eyebrow and mouth movements can be seen to describe strong rain. However, in the other weather noun groups a variety of FEs occurred. 93

6.5 Annotated sample of simultaneous co-articulation of two grammatical classes (NUM: determiner, ADV: adverb) in the sentence. The hand articulator produces the sign VARIOS (SEVERAL) while the FEs MUITO (A LOT). The figure shows the different tiers for signed annotation: translation into Portuguese (LP_P1 transcricao), literal translation into Portuguese respecting the LGP syntactic order (LGP_P1 Tran_Lite), the written corresponded meaning of the sign (gloss) (GLOSA_P1), gloss produced by one hand articulator (GLOSA_P1-M1), gloss produced by FEs (GLOSA_P1_EXPR), grammatical class of the sign produced by one hand articulator (M1_ClassGram), and the grammatical class of the sign produced by FEs (Exp_ClassGram). 95

6.6 Samples from the LGP data showing facial expressions producing the sign MUITO (A LOT). 96

6.7 Boxplot analysis of all samples for a) "strong"in DGS and b) "a lot"in LGP. The AU with the highest median values in both sign languages is AU 4 (brow lowered). Both sign languages have active AU 25 (lips apart) and AU 26 (jaw drop) which indicate an open mouth. In both sign languages AU 7 (lid tightened), AU 10 (upper lip raiser), AU 14 (dimpler), AU 15 (lip corner depressor), and AU 17 (chin raiser) are also active. However, in LGP AU 7, AU 14, and AU 17 show higher intensities than in DGS. On the right of each boxplot, a visualization of the intensity of the activated AUs is shown for the different quantiles. 99

List of Tables

1.1 Possible variations of dataset characteristics. 5

2.1 Single Action Units (AU) in the Facial Action Coding System [224]. 19

2.2 More grossly defined AUs in the Facial Action Coding System [224]. 19

2.3 Sample of different emotion theories to show the variety of proposed theories. 23

2.4 Emotion Models used in Emotion Recognition. 29

4.1 Action Units (AUs) used as features. They provide intensity information (I) which varies between 0-5. or presence/absence information (P) with a binary value. 52

4.2 Labels used for different classification problems. 56

4.3 Average accuracy results for person independent and dependent classification. 56

4.4 Frequency occurrence of facial action units (AUs) during stress, typing 1, and 2 phases. AU 9, AU 10, AU 12, and AU 28 occur during less than 5% of the frames, which is why they are not used for further clustering analysis. . . . 60

5.1 Description of the Public Speaking Competence Rubric (PSCR) [241] evaluated as a potential label to describe the use of vocal expressions and paralanguage during a talk. 69

5.2 Fine-grained description of vocal attributes derived from PSCR, evaluated as potential label categories on sentence-level. 69

5.3 Intuitive labels used to evaluate as potential categories to annotate sentence-level samples. 70

5.4 Inter-rater agreement using different labels computed with Fleiss' kappa with interpretations based on [144]. Enthusiasm, emphasis, and vocal intensity achieved almost perfect agreement. 71

5.5 Weighted average precision, recall, and F1-score for binary (**B**) and multiclass (**M**) classification. The same model architecture was used to train different feature combinations. BERT embeddings performed best in the unimodal setting. Combining acoustic with text features performed best in the multimodal setting. 76
5.6 ANOVA significance test for three levels of enthusiasm and AU mean values on the left and standard deviation of AU on the right. AUs with the lowest p-value are highlighted. 80
5.7 T-test for two levels of enthusiasm and AU mean values on the left and standard deviation of AU on the right. AUs with lowest p-value are highlighted. . . . 80
5.8 Significance test for mean and standard deviation of F0 and loudness to evaluate the dependence with the different enthusiasm levels. Left: ANOVA significance test results three enthusiasm levels shows that all p-value< 0.05, which means that all variables influence the enthusiasm level. Right: T-test significance test for two levels of enthusiasm also shows that all variables influence the enthusiasm level. 81

6.1 Translation results of the SLT model [37] used for backtranslation when trained and evaluated with ground truth hand and body skeleton joints (manual) and facial landmarks and AUs (facial). 88
6.2 Back translation results obtained from the generative models when using only manual features. Our proposed model has the highest scores in almost all metrics compared to the models using only gloss or text. 88
6.3 Back translation results obtained from the generative models when using manual features and facial landmarks and AUs. Our proposed model has the highest scores in all metrics compared to the models using only gloss or text. 88
6.4 Occurrence of different Part-of-Speech (POS) in the sign gloss annotation and the German transcripts computed with Spacy [116]. Although gloss annotations show fewer samples for all POS, the difference in the occurrence of adjectives and adverbs is statistically significant with $p < 0.05$. 92
6.5 DGS Modifier Facial Expression Dataset Statistics. The majority of the samples were found for facial expressions indicating "strong"weather. Fewer samples were found for facial expressions indicating "light"weather as signers preferred to use the manual sign LIGHT or LITTLE instead. For that purpose samples without intensity adjectives were also annotated, being shown in the column -". 93

1

Introduction

The dream of having an intelligent virtual agent able to have a dialogue with us humans has never been as close as today with the advent of large language models such as OpenAI's ChatGPT. Nevertheless, if the goal is to imitate human-to-human communication using artificial intelligence, it is essential to introduce human affect recognition and generation into the process [9, 25, 94]. Although Rosalind Picard's work on Affective Computing [208] has initiated the study of affect using computer science methods in the 1990s, much of the focus in the last decades remained in recognizing the basic emotions defined by Paul Ekman [68]: anger, sadness, disgust, surprise, happiness, contempt, and fear.

Human communication, however, can express and process a huge variety of facial expressions carrying different information even when the differences are very subtle. Kaulard et al. [131] for instance, identified 55 different conversational expressions that are not emotional but used in daily communication (e.g. bored, remembering). Therefore, a truly emotionally intelligent algorithm able to recognize, let alone imitate the enormous variety of facial expressions in the proper context, still lies in the far future mainly due to the lack of annotated data and deeper understanding of non-verbal behavior.

In this work, I aim to study unexplored facial expressions through computational experiments and statistical analysis. For that purpose, I chose facial expressions with the potential of having a high impact on social good: cognitive stress, enthusiasm, and linguistic facial expressions in sign language. According to the American Heart Association, stress has been shown to be associated with increased cardiovascular events which can lead to death [4, 228]. Monitoring and early recognition of chronic stress through computer vision instead of expensive sensors would allow individuals to take action sooner. Enthusiastic teachers have been shown to have better-performing students [30, 304]. Systems able to recognize enthusiastic speech could provide automatic feedback and allow teachers to improve their speaking skills [292] and an understanding of enthusiastic speech would allow them to develop enthusiastic virtual agents [290]. Although facial expressions have been shown to have grammatical roles in sign languages, many sign languages are understudied and lack formal dictionaries and even grammar. Formal definitions of how for instance certain adverbs are communicated solely through facial

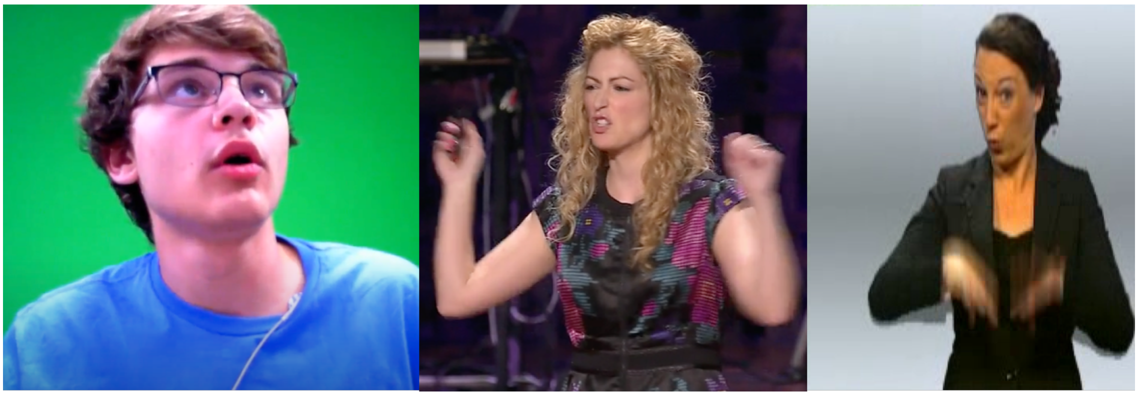


Figure 1.1: From left to right: an example from the stress, enthusiasm, and German Sign Language datasets on facial expressions.

expressions, would move sign language linguistics forward. At first glance, all three types seem to have nothing in common. At a closer look, we see that stress, enthusiasm, and some sign languages lack any formal definition of their facial expressions in literature and even datasets to study them. It is not even clear if there are typical facial expressions of individuals or even universal facial expressions for these three types. The main objective of this work is to provide new resources and statistical definitions of the facial action units active during stress, enthusiasm, and sign language.

With this work, I hope to contribute valuable knowledge and resources to the community of researchers in different fields such as perceptual and cognitive sciences, affective computing, linguistics, as well as computer vision. I believe that our research on facial expressions during stress, enthusiasm, and sign language will enable researchers in academia and industry to create more engaging human-computer interactions through extended facial expression recognizers as well as more expressive virtual agents.

The remainder of this chapter will introduce the study by first discussing the background and challenges, followed by the research aim and objectives, the significance of this research, and finally supporting publications and an outline of the thesis.

1.1 Background

For centuries artists, physiognomists, anatomists, and other scholars studied facial expressions in connection with emotions [5, 149, 27, 53] and even personality [63]. Although Darwin's work on emotions "The Expression of the Emotions in Man and Animals" was published in 1872, behaviorism emerged in the early 1900s, and studying unobservables such as emotions was largely ridiculed [160]. In addition, negative findings from flawed early research in the 1920s by Landis [142, 143] suggested that the face did not provide accurate information about emotions. It was only in the 1970s that Ekman and Izard began their research in psychology on facial expressions related to emotions, even though it was common belief in psychology that the face did not provide accurate information about internal states, especially emotions [33].

Ekman and Izard performed separate field studies, not being aware of each other's work, and were able to show that universal emotions exist which are recognizable through facial expressions independent from cultural background [73, 121]. Both very quickly recognized the need for a system able to measure the activity of facial muscles during expressions, which led to the development of the Facial Action Coding System (FACS) [75] by Ekman et al. and the Maximally Discriminative Affect Coding System (MAX) [122] by Izard. Over the years, FACS became the more established taxonomy to annotate facial expressions, and with the first annotated datasets such as the CMU-PITTSBURGH AU-Coded Face Expression Image Database from 2000 [129] it was possible to train computer vision algorithms to recognize basic emotions. Over the following years, more datasets annotated by FACS specialists were created [305, 171] and shared with the community making it today possible to use commercial applications able to detect universal emotions from facial expressions [77, 78]. Applications for emotion recognition are vast and include neuromarketing, brand perception, fragrance and aroma research, and healthcare [3].

With the increasing accuracy of detecting emotions from facial expressions, researchers in human-computer interaction started to integrate emotion recognition in dialogue planners, to create emotionally intelligent dialogue systems [161, 179]. Although rule-based systems [205, 49] were designed to detect and show basic emotions such as happiness or sadness, systems are still far from being emotionally intelligent and empathetic due to the limiting extent of understood emotions [119].

In addition, facial expressions also serve as communicative information channels, a fact established in Darwin's work on emotions in 1872 [53] (see also [85, 226]). Reference [86] also shares the view that communicative facial expressions are more often used than emotional expressions. Although [66], [24] elaborated theoretical attempts to disentangle emotions from conversational expressions, strong empirical evidence for these differences is still missing to date [131]. In 2004, [51] studied nine conversational expressions by focusing on areas of the face such as the mouth and eyebrows, and movements of the eyes and head. In 2012, [131] created the first database with 55 conversational expressions, which were recorded in a laboratory setup with subjects posing facial expressions after being given the context of the situation they were supposed to be in. Reference [42] confirmed most of the expressions in 2018 while creating another database with 62 conversational expressions. However, in the mentioned cases, objective metrics such as the analysis of active facial action units from FACS were not performed. Instead, user studies confirmed whether the facial expressions were recognizable when the situational context was given.

The importance of facial expressions in communication is most visible in sign language. Although manuals are often dominant in sign language, linguistic facial expressions also called grammatical facial expressions have been shown to define for instance wh-questions, yes/no questions, doubt questions, topics, negatives, affirmatives, conditional clauses, focus and relative clauses [252]. Although there are over 300 sign languages, they are understudied and some lack formal grammar and dictionaries (e.g. emerging sign languages) [99, 166]. Grammatical facial expressions in Brazilian Sign Language (LIBRAS) have been studied in detail in [252], revealing which facial action units are relevant to communicate for example wh-questions. Unfortunately, this analysis performed on facial expressions in sign language is an exception and not common practice. Understanding which facial action units are relevant for different grammatical functions is essential to building translation systems from sign languages to spoken language and vice versa. Current applications such as ProDeaf Translator [14] and Hand Talk Translator [271] display a signing avatar, however, lack facial expressivity.

1.2 Existing limitations and challenges

We have seen that facial expressions have three different roles in human communication: emotional (basic emotions), conversational (occur in dialogues), and linguistic (in sign languages). Given that facial expressions of emotion have been the most deeply studied and that researchers investigating conversational and linguistic facial expressions come from different research fields, two main problems exist: 1) lack of in-the-wild data especially for conversational and linguistic facial expressions, and 2) different methodologies used to study facial expressions in different fields.

Lack of in-the-wild data: The first available datasets on emotion were photographs of posed emotions in a laboratory setting with controlled environmental conditions such as lighting [129]. Questions were raised about whether these posed expressions truly represented human emotion or if they represented what individuals thought to be the proper expression of emotions given the society in which they grew up [225]. Spontaneous facial expressions of emotion were then used to create larger in-the-wild datasets [108, 172, 187]. New challenges arose, especially for computer vision recognition algorithms, as factors such as lighting, head pose, and obstructions were not controlled. Nevertheless, with larger datasets recognition algorithms were able to be trained to overcome these challenges, and currently commercially available software can reach average F1 scores of 0.76 for posed expressions and 0.51 for spontaneous expressions [64]. Although for many years, the main focus was to study the basic emotions, recent advancements focus on detecting microexpressions [151], expressions from Plutchick's wheel of emotion [214], and compound expressions [61] which are combinations of the basic emotions (e.g. happily surprised). Unfortunately, in comparison conversational and linguistic facial expressions have been studied by much fewer researchers and the number of datasets is quite low [131, 42, 253]. Given that conversational expressions are much more used in daily communication than emotional expressions [86], it is of key importance to study them in more detail and to create new datasets to improve HCI systems that can understand the subtle clues of the interlocutors (e.g. bored, interested) but also can communicate in an engaging and enthusiastic manner. Similarly, it is important to create datasets on linguistic facial

Table 1.1: Possible variations of dataset characteristics.

Emotion Elicitation	Duration	Recorded Subjects	Environment
posed	still images	professional actors	laboratory
spontaneous	videos	amateur actors	professional recording setup
artificially generated		non actors	uncontrolled condition
		avatars	

expressions for sign languages as they have grammatical and lexical roles. Although datasets exist for certain sign languages (e.g. ASL, DGS, LIBRAS) [62, 280, 250, 107] the majority remain low resource languages [99]. At the moment of writing, the only dataset focussing on facial expressions in sign language has been created for LIBRAS, however, it is not publicly available [253].

Different methodologies used to study facial expressions in different fields: Given the different roles facial expressions can have in communication, researchers from different fields such as psychology, cognitive science, computer vision, social psychophysics, and sign language linguists focused on their study. Although all investigate facial expressions, methodologies and tools vary. In the following, I will elaborate on three main components of the methodology which vary : 1) dataset creation, 2) quantitative analysis, and 3) qualitative analysis.

Table 1.1 provides an overview of how datasets can vary. As previously mentioned, datasets can show posed or spontaneously elicited emotions in laboratory or uncontrolled conditions. More recently, datasets have been created using avatar animations with different combinations of facial action units [123] to study how different expressions are perceived by study participants from different regions in the world. Although still images are more frequent in the study of facial expressions, sequences are also being explored [123].

To perform quantitative analysis of facial expressions, FACS created by Ekman [75] has a long tradition in psychology. FACS is a taxonomy that allows to measure the activity of facial muscle groups in a continuous range from zero to five. The annotation is done by specially trained annotators who underwent an examination. Researchers in the field of psychology frequently have datasets annotated by FACS specialists [98, 246, 177, 306]. Nevertheless, with the release of larger emotion datasets with FACS annotation, computational algorithms have been trained to detect FACS annotations automatically. Researchers with computer vision knowledge frequently use automatic FACS detection [253, 111, 44, 168], as it is time-effective and not costly. Other possible measures are the movement and position of facial landmarks, gaze, and head.

Qualitative analysis is mostly the only type of analysis performed in the field of cognitive science. Single frames are shown to study participants who can choose from a list of discrete emotions or conversational expressions [131, 51, 123]. The user study

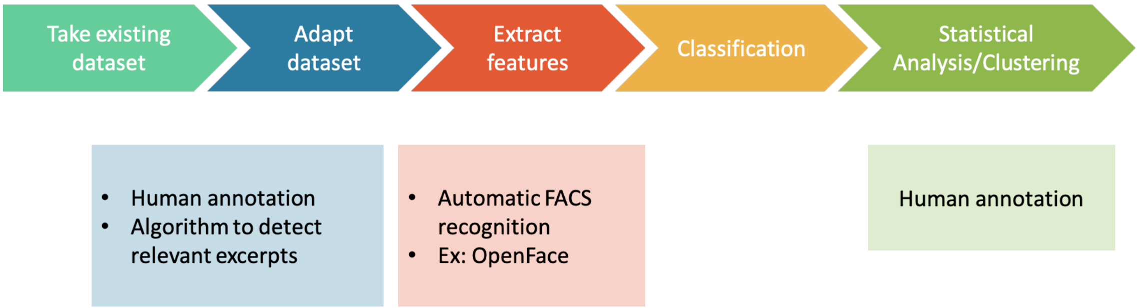


Figure 1.2: Data-bootstrapping method for facial expression analysis, composed of 5 steps: take an existing dataset, extract relevant excerpts through human annotation or algorithms, extract FACS features, train a classifier to verify if patterns are found, and perform statistical analysis and/or clustering combined with human annotation to identify facial patterns.

responses are then used to calculate for example the percentage of correctly recognized emotions or conversational expressions. Another type of qualitative analysis performed in psychology is the annotation of valence and arousal which are continuous dimensions of emotion [140, 187, 298].

1.3 Research Aim, Objectives, and Question

The research aim of this thesis is to deepen the knowledge of unexplored facial expressions.

Given the challenges mentioned before, I present a data bootstrapping method that takes available data, uses semi-automatic algorithm to find relevant expressions, and confirms expressions through annotation (see Fig. 1.2 and chapter 3).

In this work, I will apply this method to cognitive stress, enthusiasm, and sign language with the aim to answer the following research question:

How are facial expressions with different purposes such as emotional, conversational, and linguistics expressed in the examples of cognitive stress, enthusiasm, and modifiers in sign language?

Hence, the focus is on studying the facial behavior of expressions in under-explored domains, contributing to a better understanding, quantification, and qualitative analysis of stress, enthusiasm, and sign language facial expressions. Moreover, we aim at creating a novel cross-cutting dataset relevant to research communities in different fields, thus contributing to the general understanding of human facial expressions.

To tackle this thesis research question, I propose a structured approach centered on three research objectives:

Objective 1: Be the first to use different computational algorithms to evaluate whether stress, enthusiasm, and linguistic facial expressions in sign language can be distinguished by using only FACS as features.

This is our first step towards evaluating whether recognizable patterns exist in facial expressions during stress, enthusiasm, and linguistic facial expressions in sign language. While Ekman, MPI people, and Chen, put the primary focus on creating datasets and analyzing statistical characteristics, I focus on using existent datasets and creating relevant subsets to train computational models.

Objective 2: Be the first to evaluate if characteristic facial expressions exist during stress, enthusiasm, and linguistic facial expressions using FACS.

I will perform quantitative and qualitative analysis to evaluate whether stress, enthusiasm, and linguistic facial expressions in sign language have characteristic expressions.

Objective 3: Create the first curated facial expressions databases for stress, enthusiasm, and linguistic facial expressions with samples from the studied datasets, by performing user studies.

In contrast to other researchers, we want to create a library of facial expressions validated through user studies. Based on the curated frames it would be possible to generate artificial data with slight variations for further research purposes or transfer the expression to avatars using blend shapes to improve HCI.

1.4 Significance

Facial expression recognition and generation are essential components to create engaging human-computer, human-robot, and virtual interactions through avatar representation of users. The work presented in this thesis could not only provide valuable resources and insights for academic research but also for different industry sectors that aim to create emotionally intelligent technologies as well as signing avatars. The market value of such sectors has grown rapidly in the last few years. The Intelligent Virtual Assistant (IVA) market was valued at USD 2.58 billion in 2020 and is expected to reach USD 6.27 billion by 2026 [118]. The metaverse market size was valued at USD 22.79 billion in 2021 and is expected to grow at a compound annual growth rate (CAGR) of 39.8% during 2022-2030 [180]. Also, the global social robots market size is expected to expand at a CAGR of 19.93%, reaching USD 889.31 million by 2027 [257]. In the following, we will describe how a better understanding of facial expressions during cognitive stress, enthusiastic presentations, and sign language can be useful to improve technologies.

Stress is a natural process that was developed during evolution to keep us humans alive in dangerous situations. Chronic stress, however, is often unnoticed as it builds up due to daily concerns, e.g., paying bills and family obligations. Untreated chronic stress can lead to serious heart health conditions, anxiety, depression, weakened immune system, and obesity [15]. Stress detection through facial expressions is not invasive and the hardware needed is affordable.

Enthusiasm has shown to be beneficial in teaching, persuasion, and coaching [30, 229, 132, 10]; all areas of application for intelligent virtual agents. Although the benefits of enthusiastic presentations have been shown, there is a lack of datasets to study the phenomena of enthusiasm. In [289], I presented the first multimodal dataset on enthusiasm. Automatic generation of enthusiastic behavior could make virtual agents more engaging and increase engagement during teaching or coaching.

Sign languages are spoken by more than 72 million deaf. Nevertheless, there is a shortage of sign language interpreters in hospitals, governmental institutes, and universities. Navigating websites or interacting with hearing people is also quite challenging as many deaf are not as comfortable with written text as they are with their native sign language. Sign Language Dialogue Systems could help increase accessibility and inclusion. Nevertheless, sign languages bring new challenges to dialogue systems. Signs are composed of five phonological components: shape, location, movement, and orientation of hands and non-manuals (facial expressions, mouthings, body posture). Although gloss annotation (e.g. FATHER CAR EXIST.) is often used to represent the meaning of individual signs, it lacks information about the use of facial expressions. In contrast to spoken languages, facial expressions have lexical and grammatical roles in sign languages [252, 19, 291] and the lack of them causes the loss of information. In this work, I propose to create the first dataset on facial expressions with lexical meaning in German and Portuguese Sign Language. Understanding which facial expressions are used to represent adjectives of quantity will allow dialogue systems to include them, f. ex. through predefined rules. In addition, we hope to make the computer science research community aware of the importance of facial expressions in sign language to increase the research efforts.

1.5 Contributions

In the following, I provide an overview of the contributions and publications that resulted from this dissertation, organized by chapter. The first author made the major contributions of all listed articles.

Chapter 3: Data Bootstrapping for Facial Expression Analysis

- Method to create datasets for understudied emotions, conversational expressions, and linguistic facial expressions.

Chapter 4: Facial Expressions during Cognitive Stress

- Proof-of-concept on detecting stress solely from facial action units.
- First machine learning model trained solely with facial action units able to detect stress from frontal face video recordings.
- Statistical facial action unit analysis of 115 subjects during cognitive stress.
- Semi-automatic method to identify relevant frames showing stress in facial expressions.
- User study to identify facial patterns that are recognized as stressed.
- Proposal of seven facial patterns identifiable as stress.

  Publications:

– Carla Viegas, Shing-Hon Lau, Roy Maxion, and Alexander Hauptmann. "Towards independent stress detection: A dependent model using facial action units." In 2018 International Conference on content-based multimedia indexing (CBMI), pp. 1-6. IEEE, 2018.

– Carla Viegas, Shing-Hon Lau, Roy Maxion, and Alexander Hauptmann. "Distinction of stress and non-stress tasks using facial action units." In Proceedings of the 20th International Conference on multimodal interaction: Adjunct, pp. 1-6. 2018.

– Carla Viegas, Roy Maxion, Alexander Hauptmann, and João Magalhães. "The Seven Faces of Stress: Understanding Facial Activity Patterns during Cognitive Stress." In 2024 18th International Conference on Automatic Face and Gesture Recognition (FG).

Chapter 5: Facial Expressions during Enthusiastic Presentations

- First multimodal enthusiasm dataset, Entheos, built of TED talk speeches with annotated enthusiasm level. It contains sentence segments, labeled as monotonous, normal, or enthusiastic.
- Data analysis to identify attributes present in enthusiastic speech in different modalities.
- Several baseline models using different features extracted from text, speech, and video.
- Importance of identifying discourse relations in predicting enthusiasm.

  Publications:

– Carla Viegas, and Malihe Alikhani. "Entheos: A multimodal dataset for studying enthusiasm." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 2047-2060. 2021.

- Carla Viegas, and Malihe Alikhani. "Towards Designing Enthusiastic AI Agents." In Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, pp. 203-205. 2021.

- Carla Viegas, Albert Lu, Annabel Su, Carter Strear, Yi Xu, Albert Topdjian, Daniel Limon, and J. J. Xu. "Spark creativity by speaking enthusiastically: Communication training using an e-coach." In Proceedings of the 2020 International Conference on Multimodal Interaction, pp. 764-765. 2020.

Chapter 6: Facial Expressions in Sign Language

- First to use facial action units in automatic Sign Language translation.

- Novel framework that captures information from text and gloss annotation, as well as their relationship to generate continuous 3D sign pose sequences, facial landmarks, and facial action units.

- Improvement of Sign Language Generation when using facial information.

- Novel semi-automatic method to annotate facial expressions with linguistic roles.

- First dataset with facial expressions serving as adjectives in German Sign Language (DGS).

- Statistical analysis of facial expressions serving as adjectives in German Sign Language (DGS) and Portuguese Sign Language (LGP).

  Publications:

- Carla Viegas, Mert Inan, Lorna Quandt, and Malihe Alikhani. "Including facial expressions in contextual embeddings for sign language generation." In 2023 Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023).

- Carla Viegas, Lorna Quandt, and Malihe Alikhani. "Look For Adjectives In the Face: How Facial Expressions Contribute To Meaning In Signed Languages." In Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 44, no. 44. 2022.

- Carla Viegas, Mara Moita, Sofia Cavaco, Alexander Hauptmann, Jeffrey Bigham, and João Magalhães. "Unlocking Silent Conversations: Linguistic Facial Expressions in Sign Language." Submitted to FG 2025.

1.6 Other Contributions

In addition to the work presented in this thesis, the following contributions have been made to move the field of emotion recognition and gesture generation forward.

- Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. "Evaluating gesture-generation in a large-scale open challenge: The GENEA Challenge 2022". ACM Transactions on Graphics 43.3 (2024): 1-28.
- Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. "The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation." In Proceedings of the 2022 International Conference on Multimodal Interaction, pp. 736-747. 2022.
- Carla Viegas. "Two-stage emotion recognition using frame-level and video-level features." In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 912-915. IEEE, 2020.
- Sai Krishna Rallabandi, Bhavya Karki, Carla Viegas, Eric Nyberg, and Alan W. Black. "Investigating Utterance Level Representations for Detecting Intent from Acoustics." In INTERSPEECH, pp. 516-520. 2018.

1.7 Thesis Structure Outline

In this chapter, the context of this thesis has been introduced. The research objectives and questions have been identified, and the value of such research argued. The remainder of this thesis is organized as follows:

- Chapter 2 presents the background of facial expressions, how they relate to emotions, and how they are used during communication.
- Chapter 3 provides an overview of existing methods to create datasets to study facial expressions. In addition, it presents my method based on data bootstrapping which is used in the following chapters to study stress, enthusiasm, and linguistic facial expressions in sign language.
- Chapter 4 presents my work on facial expressions during cognitive stress. In chapter 4.3 we describe our experiments to recognize cognitive stress only using facial expressions. In chapter 4.4 we present a semi-automatic method to define how cognitive stress is shown through facial expressions.
- Chapter 5 presents our work on facial expressions during enthusiastic presentations. In chapter 5.3 we describe our experiments to recognize enthusiasm from facial

expressions, speech, and text. We also perform statistical analysis to understand the importance of the individual facial action units during enthusiasm.

- Chapter 6 presents our work on facial expressions during sign language. In chapter 6.2 we describe our experiments to generate facial expressions together with manual signs and in chapter 6.3 we propose a semi-automatic method to annotate facial expressions with semantic meaning in sign language. We also show how modifiers are expressed through facial expressions in German and Portuguese Sign Language.

2
Background

"Every student who examines expression itself, not its recognition, must be impressed with individual differences in the speed, magnitude, and duration of expression as well as variations in which facial expression of emotion occurs in response to a particular event. It is not known whether such differences are consistent across emotions or situations, or over time. We also do not know whether facial activity is a necessary part of any emotional experience."

Paul Ekman
FACIAL EXPRESSIONS OF EMOTION: New Findings, New Questions

When we think about facial expressions, the first association we make is with emotions. We recognize negative emotions such as anger, sadness, confusion, pain, disapproval, or fear when someone furrows their brows for example. Nevertheless, facial expressions also have other roles [170]. They can serve as *speech illustrations* – people often raise their brows when being inquisitive, and lower their brows when they lower their voices. People can cue others that they are either done talking and it's their turn, or not, through their faces (and voice), which is called *conversation regulation*. Facial expressions can serve as *emblematic gestures*. These are movements that symbolically give verbal meaning that can be conveyed by words, such as the doubtful look produced by raising the upper lip and pushing the lower lip up. In addition, facial expressions indicate *cognitive processes*, for instance when people furrow their brows when concentrating or being perplexed.

In this thesis, we want to study facial expressions outside the well studied basic emotions defined by Ekman [72]: happiness, sadness, anger, disgust, surprise, and fear. We evaluate if it is possible to recognize stress, enthusiasm, and sign language adjectives, solely from facial expressions and employ the Facial Action Coding System to objectively measure the activity of different facial muscles during the chosen settings.

In the remainder of this chapter, we will provide background to important concepts related to our work. I begin by laying the foundations of what facial expressions are (Chapter 2.1) and how emotions are defined (Chapter 2.2). In Chapter 2.3 we provide background on what stress is and different methods to detect it. In Chapter 2.4, we elaborate on what conversational expressions and the existing definitions of enthusiasm and charisma. Finally, in Chapter 2.5 we explain the linguistics of sign languages and the role of facial expressions.

2.1 Definition of Facial Expressions

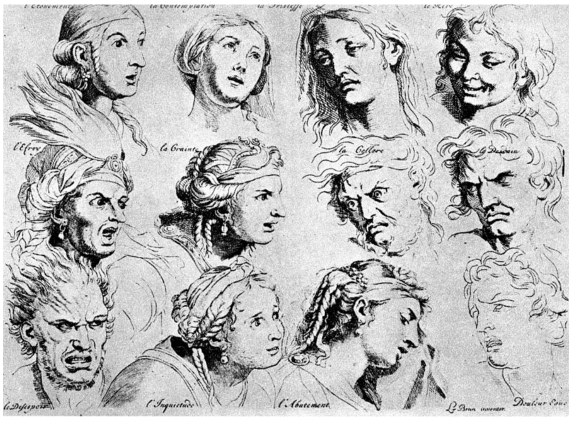


Figure 2.1: Charles le Brun, The Expressions, from the Trait é des Passions. Public Domain.

The face is one of the most complex signal systems in the human body. Each of the muscular units in the face can be activated with different timing, intensity, and laterality characteristics which allow humans to produce thousands of different expressions [170]. Since antiquity, western artists have been intrigued by facial expressions and studied them to capture moments of passion [5].

Charles Le Brun (1619-1690), artist and physiognomist, studied different expressions in detail (see Fig. 2.1) and his work *Méthode pour apprendre à dessiner les passions* [149] influenced art theory for the following two centuries. Inspired by his work, the surgeon and anatomist, Sir Charles Bell, published in 1806 *The Anatomy and Philosophy of Expression* [27] where he explains the relation of expressions and respiration.

Another important milestone in the study of facial expression was the photographic work of Guillaume-Benjamin-Amand Duchenne de Boulogne (1806-1875) called *The Workings of Human Physiognomy* [63]. Duchenne believed that facial expressions are directly linked to the soul of man. To study the variety of facial expressions, Duchenne triggered muscular contractions with electrical probes and used the recently invented camera to caption sometimes distorted and twisted faces (see Fig. 2.2).

Although the mentioned works studied facial expressions and their muscular activities, they did not try to understand what caused the expressions. Nevertheless, Bell's and Duchenne's work were essential for Charles Darwin's (1809-1882) own study on the reasons why humans have emotions which resulted in *The Expression of Emotions in Man and Animals* [53] (see more in Chapter 2.2).

Although different artists and researchers had studied facial expressions for several centuries, no objective measuring system for facial expressions existed until the late 70's. Ekman and Izard were the first to research the universality of facial expressions related to emotions. They created two different systems of facial measurements: the Facial Action Coding System (FACS) [74] and the Maximally Discriminative Facial Movement Coding System (MAX) [122]. The development of these systems were fundamental to advance the

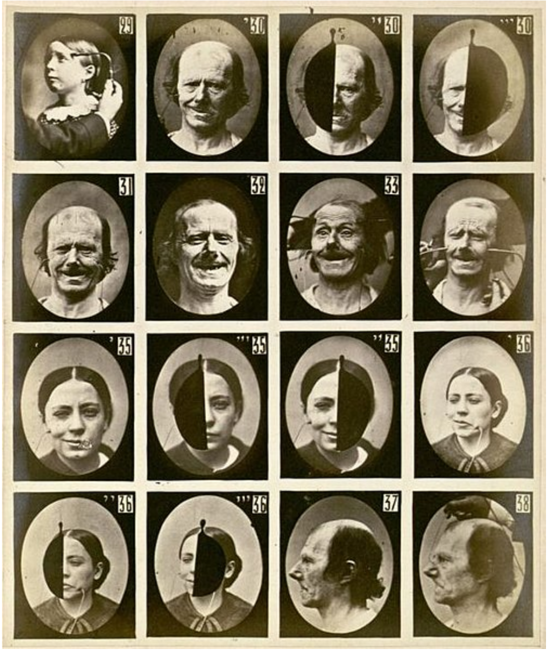


Figure 2.2: Facial expressions triggered by electric stimulation. Fig. 4, p. 277 from Mécanisme de la Physionomie Humaine by Guillaume Duchenne, 1862 [63]. Public Domain.

study of facial expression and emotion and FACS in particular has become an important tool to measure facial behavior.

To fully understand the requirements of the facial measurement system, I will provide a short overview of the anatomical characteristics of facial muscles (Chapter 2.1.1) and explain the difference of voluntary and involuntary facial movements. Then I will present existing face-measuring systems (Chapter 2.1.2) going into more detail on FACS.

2.1.1 Facial Muscles

The face includes over 40 structurally and functionally anatomically independent muscles, which can innervate independently of each other [170]. Facial muscles are the only somatic muscles which are attached on one side to bone and on the other to skin, making them extremely specialized for facial expressions. In addition, the face has some of the few muscles which are not attached to any bone at all, such as orbicularis oculi and obicularis oris (see Fig. 2.3). Another peculiarity of facial muscles is that the structure and function in some facial muscles are not linked. The corrugator muscle group is comprised of three muscles which usually act together when innervated, bringing the brows down and together. The frontalis muscle, on the other hand, is a single muscle spanning the forehead, but the inner and outer parts of this muscle can be moved independently of each other. The described characteristic shows the importance of decomposing facial behavior from the perspective of functional and not structural anatomy.

When studying facial behavior, it is also relevant to understand that voluntary and

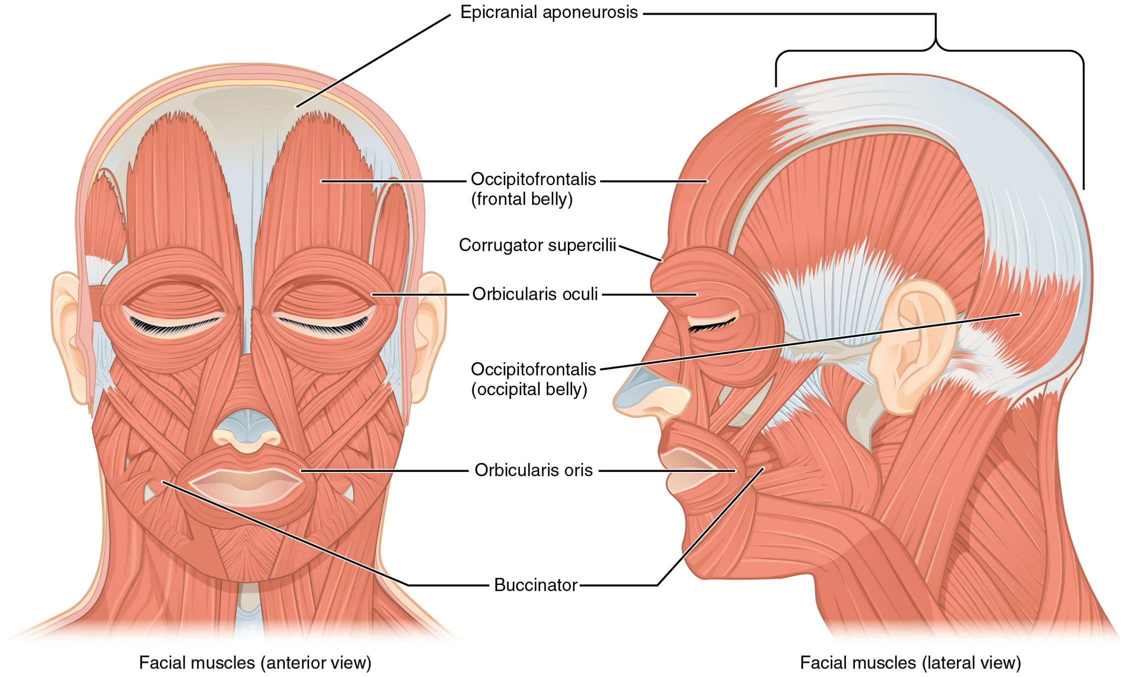

Epicranial aponeurosis
Occipitofrontalis (frontal belly)
Corrugator supercilii
Orbicularis oculi
Occipitofrontalis (occipital belly)
Orbicularis oris
Buccinator
Facial muscles (anterior view)
Facial muscles (lateral view)

Figure 2.3: Overview of facial muscles. Although more than 40 facial muscles exist, it is important to study facial behavior based on functional and not structural anatomy. E.g. the corrugator muscle group is comprised of three muscles that usually act together when innervated, bringing the brows down and together. The frontalis muscle, on the other hand, is a single muscle spanning the forehead, but the inner and outer parts of this muscle can be moved independently of each other, allowing for just the inner or outer corners of the eyebrows to rise. Graphics from [198]

involuntary expressions are controlled by different parts of the brain [219]. Voluntary expressions are controlled by impulses from the motor strip through the pyramidal tract. The lower face in particular is more fully represented in the motor cortex, which allows for more voluntary and learned control necessary for instance for speaking. The upper eyelid is innervated by the oculomotor nerve and is innervated during expressions of surprise, fear, and anger.

2.1.2 Face Measuring Systems

Facial Action Coding System FACS [74, 65] is a "comprehensive, anatomically based system for measuring all visually discernible facial movements" [224]. FACS can describe all visually distinguishable facial movements using 44 unique Action Units (AUs) (see Fig. 2.4). In addition, it comprises several categories for movements and positions of the head and eyes. In Tables 2.1 and 2.2 the different AUs with their numeric code are listed. Although the table also lists the muscle group of the AU, as seen in 2.1.1, a one-to-one correspondence between muscle groups and AUs is not possible as a given muscle can contract in different ways. FACS also allows coding of the intensity of each facial action on a five-point intensity scale, the onset, apex, and end of a facial action, and the coding of facial expressions as "events". An event is the description of a facial expression in terms of a single AU or the combination of several AUs. Other variations of FACS have also been developed:

- EMFACS: In 1982, Ekman and Friesen developed Emotion FACS [88] which identifies subgroups of AUs related to emotions.
- BabyFACS: Specialized version of FACS to classify facial movement in babies, given differences in proportions and dimensions of the bony structures of the face and specific behavior such as lip sucking [200].
- ChimpFACS: In 2007, Vick and colleagues developed a modified version of FACS for chimpanzees [288]. Although differences in muscle anatomy exist between humans and chimps, many muscles related to human facial expressions of emotion have the same location and functional effect in chimpanzees.

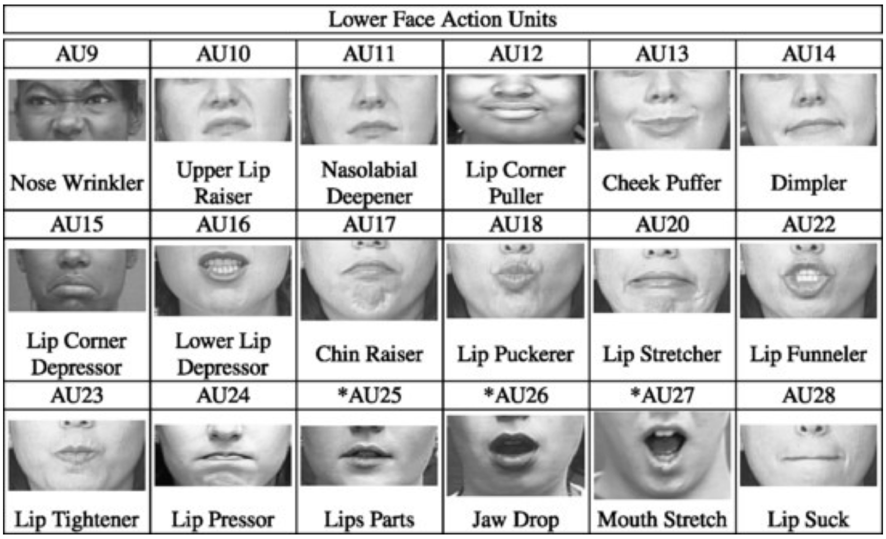


Figure 2.4: Examples from different Facial Action Units (AUs) [87] from the lower face relevant to the generation of mouthings in sign languages. AUs can occur with different intensity values between 0 and 5. AUs have been used in psychology and in affective computing to understand emotions expressed through facial expressions. Image from [275].

Maximally Discriminative Affect Coding System (MAX) Similar to FACS, MAX is an observational coding scheme that describes expression in terms of components [224]. While FACS was derived anatomically, MAX was derived theoretically which means that it codes only the facial configurations which Izard theorized to correspond to universal emotions. The disadvantage of such a system is that new facial behaviors cannot be discovered.

Electromyography (EMG) Facial EMG allows the measurement of electrical potentials elicited through muscular contractions. Although EMG is able to measure even the

Table 2.1: Single Action Units (AU) in the Facial Action Coding System [224].

AU	Descriptor	Muscular Basis
1	Inner Brow Raiser	Frontalis, Pars Medialis
2	Outer Brow Raiser	Frontalis, Pars Lateralis
4	Brow Lowerer	Depressor Glabellae, Depressor Supercilli; Corrugator
5	Upper Lid Raiser	Levator Palpebrae Superioris
6	Cheek Raiser	Orbicularis Oculi, Pars Orbitalis
7	Lid Tightener	Orbicularis Oculi, Pars Palebralis
9	Nose Wrinkler	Levator Labii Superioris, Alaeque Nasi
10	Upper Lip Raiser	Levator Labii Superioris, Caput Infraorbitalis
11	Nasolabial Fold Deepener	Zygomatic Minor
12	Lip Corner Puller	Zygomatic Major
13	Cheek Puffer	Caninus
14	Dimpler	Buccinator
15	Lip Corner Depressor	Triangularis
16	Lower Lip Depressor	Depressor Labii
17	Chin Raiser	Mentalis
18	Lip Puckerer	Incisivii Labii Superioris; Incisivii Labii Inferioris
20	Lip Stretcher	Risorius
22	Lip Funneler	Orbicularis Oris
23	Lip Tightener	Orbicularis Oris
24	Lip Pressor	Orbicularis Oris
25	Lips Part	Depressor Labii, or Relaxation of Mentalis or Orbicularis Oris
26	Jaw Drop	Masetter; Temporal and Internal Pterygoid Relaxed
27	Mouth Stretch	Pterygoids; Digastric
28	Lip Suck	Orbicularis Oris

Table 2.2: More grossly defined AUs in the Facial Action Coding System [224].

AU	FACS name
8	Lips Toward Each Other
19	Tongue Out
21	Neck Tightener
29	Jaw Thrust
30	Jaw Sideways
31	Jaw Clencher
32	Lip Bite
33	Blow
34	Puff
35	Cheek Suck
36	Tongue Bulge
37	Lip Wipe
38	Nostril Dilator
39	Nostril Compressor
43	Eyes Closure
45	Blink
46	Wink

slightest facial movements that are not observable through FACS or MAX, its obtrusive character can influence the self-conscious behavior of the subject as it knows he or she is being observed [136].

2.2 Facial Expressions and Emotions

Emotions are deeply connected with facial expressions. In the previous chapter we have seen how artists and anatomists studied facial expressions, their corresponding muscle movements, and the "passions" (emotions) they showed, but none of them studied what caused the emotions. Herbert Spencer was one of the first to study the origin of facial expressions linked to emotions and evolution in his work *Principles of Psychology* (1855) [260]: "The destructive passions are shown in a general tension of the muscular system, in gnashing of the teeth and protrusion of the claws, in dilated eyes and nostrils in growls; and these are weaker forms of the actions that accompany the killing of prey". While Sir Charles Bell believed that facial expressions had the sole task to express emotions, Charles Darwin argues that current facial expressions evolved over time with initially other purposes [53]. Darwin believed that facial expressions and gestures have been necessary for survival in "lower" species. As Darwin links emotions to their origins in animal behaviour, he believed in universal emotions (happiness, sadness, fear, anger, surprise and disgust) from a single origin for all humanity with little influence of cultural factors in the shaping of expression.

Unfortunately, after Darwin's work on emotions, behaviorism emerged in the early 1900s and studying unobservables such as emotions was largely ridiculed. In addition, negative findings from flawed early research in the 1920s by Landis [142, 143] suggested that the face did not provide accurate information about emotions. Although several researchers described the methodological flaws in Landis's work [47, 54, 89] and well-conducted studies showed that people could consistently recognize emotion in the face [100], the prevailing view in psychology was that the face did not provide accurate information about internal states, especially emotions [33].

The interest in facial expression reemerged in psychology with Silvan Tomkins theory of affect [273], which gave the face a central role in communicating emotions. A systematic study of facial expressions followed and Tomkins and McCarter were able to show that observers consistently identified facial poses to be connected to emotions [274]. Influenced by Tomkins, Ekman [71] and Izard [120] separately studied the recognition of facial expressions of emotion in literate and preliterate cultures, unaware of each others work, and showed that the universal emotions exist.

Although the study of emotions has been neglected by psychologists, as well as historians and philosophers for several decades, it has now grown into a flourishing area of research [213]. Nevertheless, the variety of phenomena related to emotions gave way to several theories on how emotions are elicited (emotion theories) and what emotions are (emotion models). In the following, I will provide an overview of different emotion theories and emotion models. This is especially important for the final subchapter which discusses whether emotions are universal or not. This question is essential to be asked if we want to develop computational systems that can recognize emotions and elicit the appropriate emotion recognition in human users when displaying specific facial expressions in a virtual or embodied agent.

2.2.1 Definition of Emotion

The following definition of emotion is based on the works from [237] and [127].

Emotion is one type of affect, with other types including mood, temperament, and sensation (such as pain). Emotions can be viewed either as states or as processes. As a state, like being angry or afraid, an emotion is a mental condition that interacts with other mental states and influences behaviors.

When considered as a process, emotion can be divided into two parts. The early part occurs between the perception of a stimulus and the initiation of a bodily response. The later part involves the bodily response itself, such as changes in heart rate, skin conductance, and facial expression. This framework is sufficient to begin analyzing emotions, though it omits aspects like the subjective awareness of emotion and associated behaviors (e.g., fighting, fleeing, hugging).

The early part of the process typically includes an evaluation of the stimulus, meaning the occurrence of an emotion depends on the individual's perception or interpretation of the stimulus. For example, one person might respond to job loss with anger, while another might feel joy, depending on their evaluation of the event. This evaluative component means that emotions are not straightforward, direct responses to stimuli, unlike reflexes such as the startle or eye-blink responses.

Several features distinguish emotions from moods. Emotions are responses to specific stimuli, which can be internal (like beliefs or memories). Emotions generally have intentional content, meaning they are about something, often the stimulus itself. In contrast, moods are typically not about anything specific and often do not appear to be caused by a particular stimulus. Emotions are also short-lived, lasting seconds or minutes, whereas moods can persist much longer. These characteristics are widely accepted in theories of emotion.

2.2.2 Emotion Theory

Different theories have been proposed by researchers, philosophers, and psychologists to explain what causes emotions. The main questions have been: 1) are emotions triggered by physiological responses within the body, 2) is brain activity responsible for emotions, or 3) do thoughts or other mental processes form emotions? Or is it a combination of all three? In the following I will present four main streams of emotion theory: evolutionary theories, social and cultural theories, theories of the emotion process, and the theory of the constructed emotion. All theories have strengths and weaknesses as they observe the phenomena of emotion from different lenses such as psychology, philosophy or cognitive sciences. It is out of the scope of this thesis to proof which theory is correct. Nevertheless, a theoretical background is necessary to refer to as I want to develop algorithms that can distinguish expressions that show stress and enthusiasm. So the question is are stress and

Table 2.3: Sample of different emotion theories to show the variety of proposed theories.

Theory	Explanation of Emotions	Example
Darwin - Evolutionary Theory	Emotions exist to serve an adaptive role. Emotions help people to respond quickly to environmental stimuli, improving the chances of success and survival.	When we encounter a hissing animal, we quickly realize that the animal is defensive and leave it alone.
James-Lange (1920-1930) [126]	Emotions arise from our awareness of our specific bodily responses to emotion-arousing stimuli.	We observe our heart racing after a threat and then feel afraid.
Cannon-Bard [38]	Emotion-arousing stimuli trigger our bodily responses and simultaneous subjective experience.	Our heart races the same time that we feel afraid.
Schachter-Singer [238]	Our experience of emotion depends on two factors: general arousal and a conscious cognitive label.	We may interpret our arousal as fear or excitement, depending on the context.
Zajonc; LeDoux	Some emotional responses happen instantly, without conscious appraisal.	We automatically feel startled by a sound in the forest before labeling it as a threat.
Lazarus - Appraisal Theory [148]	Cognitive appraisal ("Is it dangerous or not?") - sometimes without our awareness - defines emotion.	We feel fear after seeing a bear in the forest and thinking that we are in danger.
Facial-Feedback Theory [167]	Emotions are directly tied to changes in facial muscles.	Forced smiles trigger positive feelings.?

enthusiasm distinguishable from other emotions? Are facial expressions for stress and enthusiasm universal? The choice of the right theory, will help us interpret computational results and design proper experiments.

2.2.2.1 Evolutionary Theories

Following "The Origin of Species", Darwin published "The Expression of Emotions in Man and Animals" (1872) where he compares facial expressions and body posture of different animals such as dogs, cats, primates, and humans. Darwin explains the presence of emotions in humans today through natural selection. His argument is that all humans have emotions and most animals display emotion-like responses, which leads to assume that it is likely that emotions or emotion-like behavior was present in a common ancestor. Based on these claims, three different evolutionary positions have been developed: a) emotions are the result of natural selection in early hominids, b) emotions are adaptations, c) emotions are historical.

Emotions as the result of natural selection Theorists suggest that the selection occurs in response to problems that arose due to the social environment in which the organisms lived [190]. For example, being alone was problematic and the emotion of fear resulted to avoid those situations [48]. Similarly, social ridicule ilicits social anxiety [190].

Emotions as adaptations Theorists claim that the selection occured much earlier and that the adaptations are shared by a wider collection of species today. Robert Plutchik claims that there are eight basic emotions which resulted from individual adaptations, and that all eight are found in all organisms [211]. For each adaptive behavior, Plutchik assigns emotions. Example: protection is a behavior that occurs in response to pain or threat and is ilicited by the emotion of fear and terror. Destruction as a behavior designed to destroy a barrier that prevents the satisfaction of an important need is linked to anger and rage.

Emotions are historical Paul Griffith defines different psychological categories for emotions. The affect program emotions are surprise, anger, fear, sadness, joy, and disgust and all other emotions belong either to the higher-cognitive emotions or the socially constructed emotions [102, 101]. Anger for instance can have instances belonging to different categories.

2.2.2.2 Social and Cultural Theories

In this approach, theorists claim that emotions are the products of societies and culture which are acquired or learned by individuals through experience. Although defenders of this theory acknowledge that emotions are to some degree natural phenomena, the central claim is that the social influence is so significant that emotions are best understood from this perspective. Some examples used to support this theory are that people in different cultures have and experience different emotions. For example, there are cultures where there is no difference being made between anger and sadness [55, 221, 222]. Another argument, is that emotions typically occur in social settings and during interpersonal transactions rather than individual's responses to a particular stimulus [202]. In addition, emotions and their expression are regulated by social norms, values, and expectations. These norms and values influence what the appropriate objects of emotions are and how the emotions should be expressed. As an example James Averill identified the "emotion rules" that Americans follow when angry [17], e.g. "Anger should not be displaced on an innocent third party, nor should it be directed at the target for reasons other than the instigation". Claire Armon-Jones even says that the purpose of emotions is to reinforce society's norms and values [12]. James Averill describes emotions as transitory social roles and syndromes which are generated by social norms and expectations, meaning that social norms and expectations govern an individual's emotions [16].

2.2.2.3 Theories of the Emotion Process

This third category describes emotions as a process which begins with the perception of a stimulus, although the stimulus may be internal in some cases, such as a thought or a memory. It describes the activity between the perception and the triggering of the bodily response and the bodily response itself which occur as changes in heart rate, blood pressure, facial expressions skin conductivity, and so on. There are disagreements on how simple or complex the early part of the emotion process might be, which led to the development of a) cognitive, b) non-cognitive, and c) somatic theories.

Cognitive theories include in the emotion process the manipulation of information and view it as cognitive process. This is in contrast to non-cognitive theories which state that the generation of the emotion response is a direct and automatic result of perceiving the stimulus. This theory is motivated by the observation that different individuals will respond differently to the same stimulus. Being laid-off from a job may be a relief for someone and a disaster for someone else. The same person can even feel relieved first and several years later feel scared when being laid-off [223]. Secondly, the same emotion can be caused by unrelated events. Events without common physical features or properties such as the death of a parent, divorce, or not being accepted to medical school, can cause sadness. Roseman and Smith use this example to show that "theories claiming that emotions are unconditioned responses to evolutionary specified stimulus events or are learned via generalization or association" pose problems [223]. Cognitive theories propose that every individual has beliefs, goals, personal tendencies, and desires before the emotion causing event is encountered. Based on the cognitive position, philosophers developed the judgement theory, which claims that an emotion is “a basic judgment about our Selves and our place in our world, the projection of the values and ideals, structures and mythologies, according to which we live and through which we experience our lives” [259]. In Solomon’s theory “What constitutes the anger is my judging that I have been insulted and offended" [258]. Psychologists, on the other hand, developed cognitive appraisal theories, which similar to the judgment theories emphasize the idea that emotion is determined by how an individual evaluates or appraises a stimulus. The difference, however, is that they do not rely on beliefs, judgments, and so forth [223, 146, 239]. There are different variations and complexities of appraisal theories[239]. Ortony, Clore, and Collins (OCC) model says, e.g. that the emergence of emotions originates from the cognitive evaluation or appraisal of stimuli in terms of events, agents, and objects. How individuals perceive and interpret the stimuli determines how emotions might emerge [199].

Non-cognitive theories defend that judgements or appraisals are not part of the emotion process. In these theories, emotions are seen as separate from the rational or cognitive operations of the mind and the emotion process is thought to be reflex-like. Advocates of non-cognitive theories emphasize that infants and animals do not have the cognitive capabilities necessary in the judgement theories or cognitive appraisal theories to feel emotions. In 1977, Ekman described that some emotions are non-cognitive [69] and Paul Griffiths incorporated these findings into his own theory of the emotions [102]. Ekman's models proposes two mechanisms that directly interface each other: a) an automatic appraisal mechanism and an affect program. The appraisal mechanism attends to stimuli which Ekman calls elicitors, and acts very quickly and supposedly automatic in determining which emotion and activating the proper part of the affect program as it happens sometimes unaware. The affect program controls the various elements of the emotion response, such as the skeletal muscle response, facial and vocal response, and the autonomic nervous system response. In addition to the automatic appraisal mechanism, Ekman beliefs that cognitive appraisals are sometimes used and that some emotions are cognitively mediated and some are socially constructed. Other theorists such as Jenefer Robinson belief that emotions are non-cognitive. Robinson claims that cognitive processes which occur in emotion-causing situations are in addition to the core process, which is non-cognitive. Although she acknowledges that emotions can be caused by cognitive activity, she claims that the cognitive activity precedes the non-cognitive emotion process. She claims that the automatic affective appraisal mechanism can take as stimuli complex judgments or thoughts.

Somatic feedback theories differ from cognitive and non-cognitive positions as they claim that each emotion has unique bodily responses. Although there is evidence that there are specific facial expressions for sadness, anger and other emotions, there is little evidence that other bodily responses are also unique per emotion. In this theory it, it is the feedback that the mind or brain receive from the body that makes the event an emotion. William James proposed this theory and Antonio Damasio and Jesse Prinz revived and expanded it.

2.2.2.4 Theory of constructed emotion

We have seen so far theories that put major importance to the bodily responses to distinguish emotions, while cognitive appraisal theorists acknowledge bodily responses but internal cognitive activity chooses the emotion similar to the automatic appraisal mechanism of Ekman's non-cognitive model. Nevertheless, with the advance of neuroimaging, several findings of the last two decades have put into question different aspects of the classical view of emotion. Through neuroimaging, it has been shown that different emotion categories cannot be specifically and consistently localized to distinct populations of neurons within a single region of the human brain [297, 156] nor to intrinsic networks in the human brain [23, 276] (for more examples see [22]). In addition, evidence has shown that neurons do not lie dormant until stimulated by the outside world, instead ongoing brain activity influences how the brain processes incoming sensory information [236].

Implications of these findings are profound as they indicate that it is very unlikely that perception, cognition, and emotion are localized in dedicated brain systems, making theories relying on the stimulus response mechanism highly doubtful. Lisa Feldman Barrett has analyzed these and other results of different neuroscientists and has proposed the theory of constructed emotion [22]. Barrett proposes that emotions are constructed predictively by the brain in the moment as needed, performing multimodal summaries of the state of the body (interoception), culturally emobodied knowledge (concepts), and social reality which makes the perception of emotion possible among people of the same culture. An analogy is the experience of color. Although we perceive colors due to different wavelengths which have a continuous character, we categorize them into culturally learned discrete colors such as blue or red. Similarly, emotions are thought in categories such as anger and happiness even though affect produced by interoception is continuous. The theory of constructed emotion proposes that the brain uses past experience organized as concepts and predicts and categorizes the current affect through interoceptive predictions, to construct an instance of emotion. In addition, this theory proposes that "emotions should be modeled holistically, as whole brain-body phenomena in context" and Barrett claims that "it will never be possible to measure an emotion by merely measuring facial muscle movements, changes in autonomic nervous system signals, or neural firing within the periaqueductal gray or the amygdala. To understand the nature of emotion, we must also model the brain systems that are necessary for making meaning of physical changes in the body and in the world" [22].

2.2.2.5 Conclusion

Most of the theories here presented proposed their theorems based on observations of human emotional behavior, measurements of physiological signals or facial expressions, and studies where humans recognized emotions in other humans from single images of video recordings. What is common in all theories is that the process of emotion is very fast, but very different views exist to explain the fast process. For the work presented in this thesis, I will refer to the theory of constructed emotions to explain my computational results. The evidence of not having dedicated neural networks for individual emotions is quite strong. This has several implications on my work of recognizing stress or enthusiasm. Based on this theory, there is not one single behavior that characterizes an instance of emotion. There are several possible behaviors which makes the classification problem into a one-to-many problem, meaning, one emotion can have several behaviors. Although this might not be so evident for the seven basic emotions (sadness, disgust, anger, fear, etc.) it might be the case for other instances of emotion such as stress or enthusiasm.

2.2.3 Emotion Representation Models

Emotion models are generally categorized into discrete and dimensional representations (see Table 2.4). The discrete emotion model, defined by Ekman in 1971, describes emotions

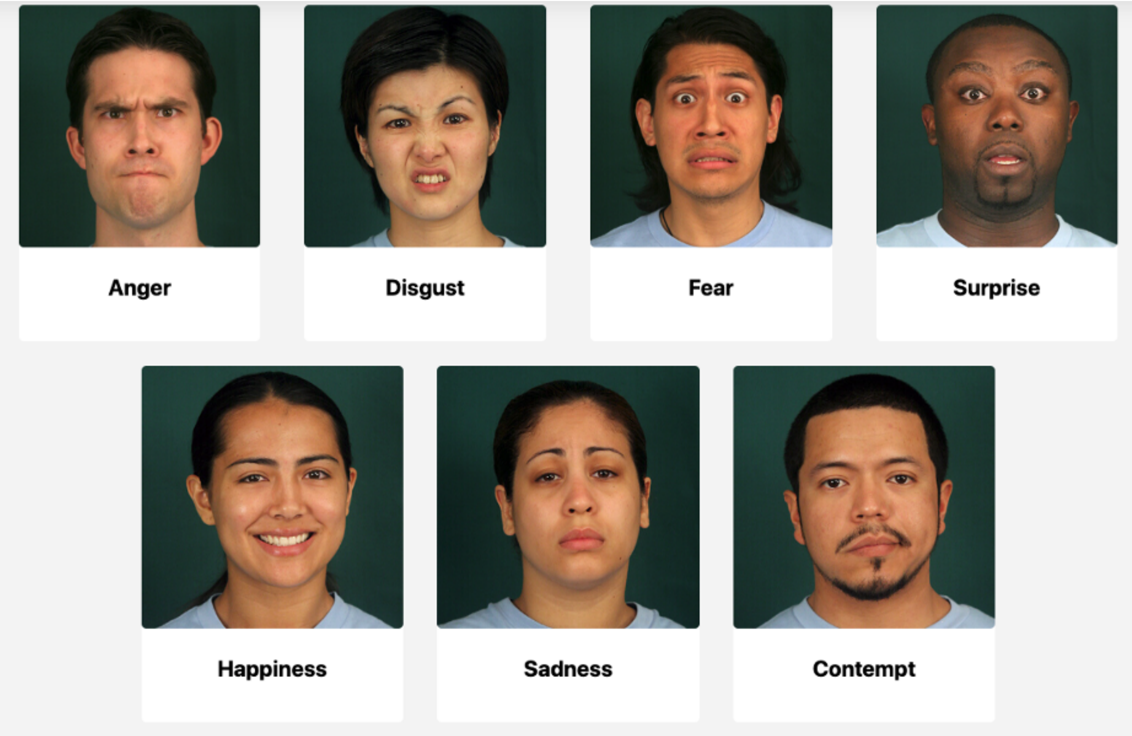


Figure 2.5: Basic universal emotions defined by Paul Ekman [103]

as six basic categories: anger, disgust, fear, happiness, sadness, and surprise [67](see Fig. 2.5).

These six basic emotions have been said to be universal across human ethnicities and cultures and can combine to form other emotions, namely compound emotions (see Fig. 2.6). Primary emotions have the following characteristics: (i) they stem from instinct; (ii) different people experience the same emotions under similar circumstances; and (iii) different people express basic emotions similarly [201]. A significant advantage of the discrete emotion model is that it can effectively describe people's emotional experiences in daily life and is intuitive to use based on the six emotion labels. Consequently, much research has focused on discrete emotion recognition. In Table 2.4 other categorical emotion models are listed, such as Mikel's wheel [182], Plutchik's wheel [210], and Parrot's emotion framework [203] that organizes emotions into primary, secondary, and tertiary emotion categories.

An alternative approach is the dimensional emotion model. Some psychologists and artificial intelligence experts believe emotions can be represented through continuous dimensions. Unlike discrete emotions, the dimensional emotion theory defines emotions as points in a dimensional space. Two typical and widely accepted dimensional emotion models are the Valence-Arousal (VA) [240] and Pleasure-Arousal-Dominance (PAD) [178] models. In the VA model, the valence dimension measures positive and negative emotional states, while the arousal dimension indicates the intensity of emotions. The PAD model builds on the VA model by adding the dominance dimension, which reflects a

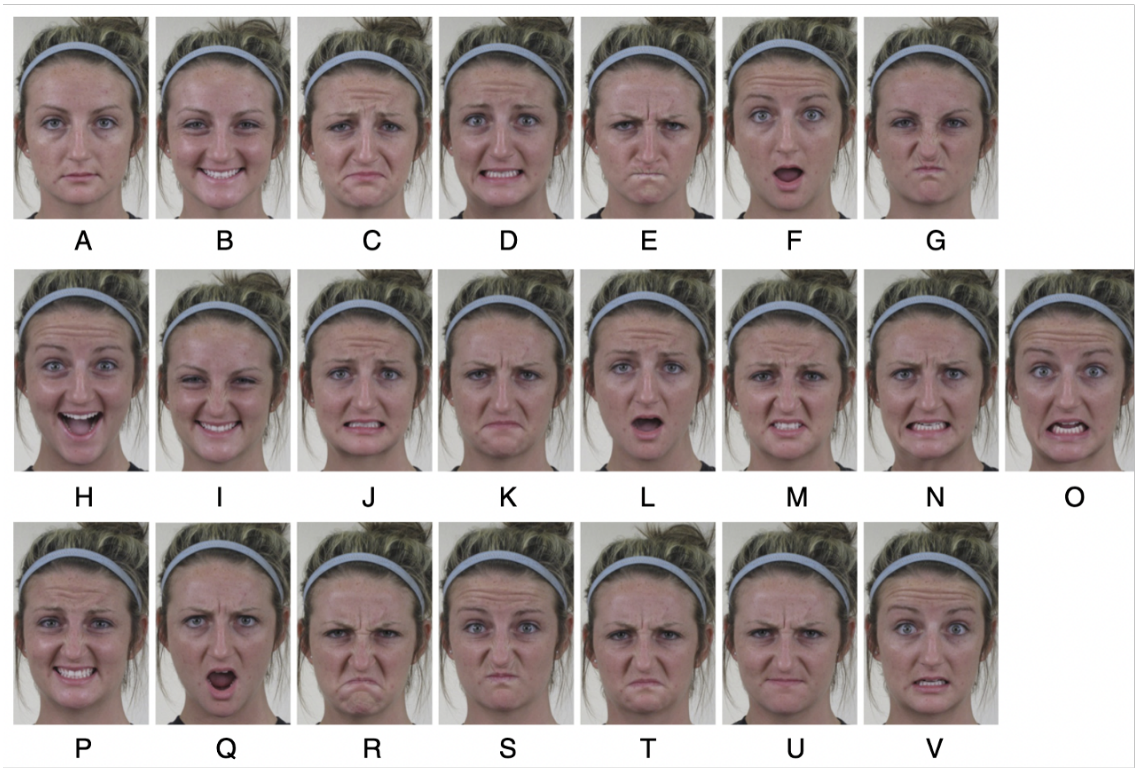


Figure 2.6: Overview of compound facial expressions of emotion [61]. (A) neutral, (B) happy, (C) sad, (D) fearful, (E) angry, (F) surprised, (G) disgusted, (H) happily surprised, (I) happily disgusted, (J) sadly fearful, (K) sadly angry, (L) sadly surprised, (M) sadly disgusted, (N) fearfully angry, (O) fearfully surprised, (P) fearfully disgusted, (Q) angrily surprised, (R) angrily disgusted, (S) disgustedly surprised, (T) appalled, (U) hatred, and (V) awed.

Table 2.4: Emotion Models used in Emotion Recognition.

Model	Type	Emotion states/dimensions
Ekman [67]	categorical	anger, contempt, disgust, fear, enjoyment, surprise, sadness
Mikels [182]	categorical	amusement, anger, awe, contentment, disgust, excitement, fear, sadness
Plutchik [210]	categorical	3 intensities of anger, anticipation, disgust, joy, sadness, surprise, fear, trust
Parrot [203]	categorical	primary, secondary, and tertiary emotion categories
VA(D) [240]	dimensional	valence, arousal, (dominance)
PAD [178]	dimensional	pleasure, arousal, dominance

feeling of control and influence over surroundings and others. Different multimodal emotion recognition datasets have been created in the last 50 years using different emotion representation models. An overview of existing multimodal datasets and algorithms developed for emotion recognition can be found in the recent review in [201].

2.3 Facial Expressions during Stress

The study of stress began during a period in which experimental psychology was still in the shadows of behaviorism (1900-1980), which assumed that the mind was just a black box between stimulus and response, not worthy to be studied [148]. Richard Lazarus played a major role in shifting the field of psychology and 1984 proposed the "Transactional Model of Stress and Coping" which emphasizes the person–environment transaction and suggests that a stress response is highly influenced by individual appraisal processes [148, 147].

In parallel Ekman et al. [72] had already defined the Facial Action Coding System (FACS) which is a taxonomy for facial muscle movement and studied facial action units (AUs) during stress-induced situations. Based on their theory of basic emotions, they found the emotion of disgust to co-occur during stress. Since then the relationship between facial expressions and stress has been studied from different angles. In [173], neurobiologists studied how gender influences the behavioral stress response, finding that men show greater stress-induced corrugator reactivity (frowning in the absence of any stimulus) than women. In [301] evolutionary psychologists found that displacement behaviors which are known to be associated with stress influence the likability of a person, indicating a benefit and potential adaptive function of displaying stress through facial expressions.

In the following, we will provide a definition of stress, describe why its detection is important, and provide an overview of existing methods to detect stress.

2.3.1 Definition of Stress

Stress is a companion of every living being. Hans Selye, the "father of stress" defined stress as the "nonspecific response of the body to any demand" [245]. Selye also said that "complete freedom from stress can only be achieved after death". Although stress can be caused by a variety of different problems (e.g. surgical trauma, fatigue, pain, emotional arousal, need for concentration, unexpected success), the biochemical, functional, and structural changes as a response are the same for positive and negative situations [83]. Any demand that causes change and requires adaptation causes stress. Although everyone experiences stress, the level at which they can cope varies from individual to individual.

There are two main forms of stress: acute stress and chronic stress. The former has a short duration, the body reacts to a new challenge or demand such as meeting new people, giving a talk, or having an accident. The body can respond through feelings (e.g. irritability, fatigue), behavior (e.g. aggressive, unmotivated), thinking (e.g. difficulties of concentration), or physical symptoms (e.g., nausea, palpitations). When acute stress is not resolved, it can accumulate over time and become chronic stress. Chronic stress is dangerous as it causes changes in neuroendocrine, cardiovascular, autonomic, and immunological functioning and can lead to mental and physical illnesses such as anxiety, depression, heart disease, and more [181].

2.3.2 Importance of Stress detection

Chronic stress is problematic not only for the individual but also for their workplace and family. Besides being linked to the six leading causes of death [228], stress has a negative impact on family life [185, 124].

Additionally, work-related stress is estimated to cause costs of $187 Billion Dollars only in the USA [109]. In order to help employers improve their working environment, companies such as Linkura [157] offer automatic stress detection of employees and strategy plans to improve the situation.

Given the risks that chronic stress brings, it is important to detect stress, observe its occurrence, and mitigate it using for instance mindfulness techniques before it develops into chronic stress causing severe problems. At the time of writing, accurate methods that individuals can access to measure stress are not available. In a clinical environment, stress is detected using costly hardware to measure physiological signals or using psychological evaluations based on subjective questionnaires. To mitigate these limitations, efforts have been made in research to evaluate alternative methods for stress detection which will be discussed in the following section.

2.3.3 Methods for Stress detection

Stress causes several biochemical, functional, and structural changes in the body [83]. Reference [104] defined four different categories of stress detection methods which are based on psychological evaluation, physiological signals, behavioral responses, and social media interactions (Fig. 2.7 shows methods for each category).

2.3.3.1 Clinical methods for stress detection

Different techniques exist to detect stress by measuring the physical reactions of the body. In clinical settings, costly hardware can be used to detect stress through anomalies of physiological signals. Through saliva or blood tests, it is possible to measure stress hormones. Skin temperature, pupil dilation, and respiration can also reveal the presence of stress. The heart's change of behavior during stress can be detected by measuring the heart rate, heart rate variability, and blood volume pulse. As stress is caused by adrenaline, heart rate is increased, and blood vessels reduce their diameter causing blood pressure to rise. R-R peak interval also gets reduced. These can be observed in an electrocardiograph (ECG). As Galvanic skin response (GSR) also called electrodermal activity (EDA) provides direct insights into the autonomous emotion regulation, it is one of the most used physiological signals in stress detection [265, 113, 296].

Most of the methods just mentioned are used in a medical environment. What they have in common is that they are invasive. This means medical instruments need to be

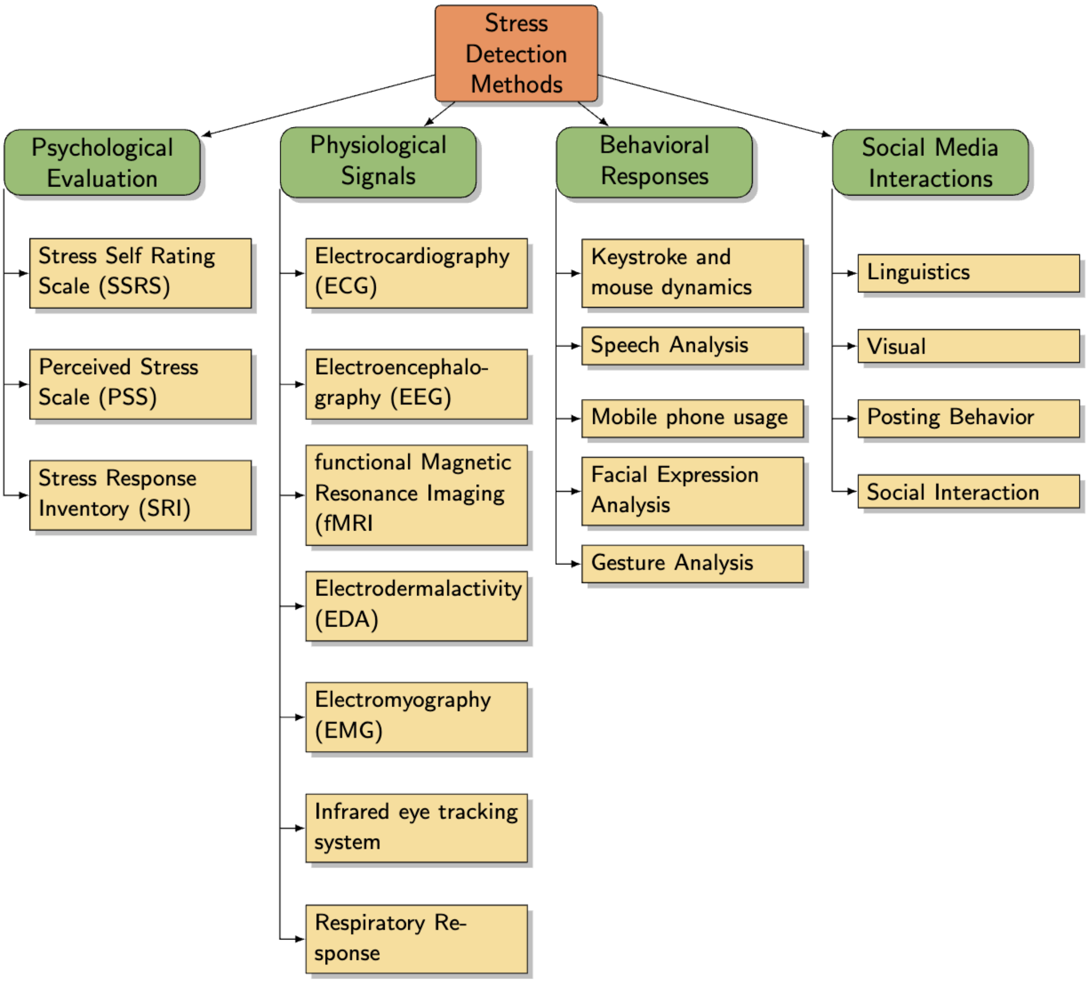

Stress Detection Methods
Psychological Evaluation
Physiological Signals
Behavioral Responses
Social Media Interactions
Stress Self Rating Scale (SSRS)
Perceived Stress Scale (PSS)
Stress Response Inventory (SRI)
Electrocardiography (ECG)
Electroencephalo-graphy (EEG)
functional Magnetic Resonance Imaging (fMRI
Electrodermalactivity (EDA)
Electromyography (EMG)
Infrared eye tracking system
Respiratory Re-sponse
Keystroke and mouse dynamics
Speech Analysis
Mobile phone usage
Facial Expression Analysis
Gesture Analysis
Linguistics
Visual
Posting Behavior
Social Interaction

Figure 2.7: Overview of four main categories of stress detection methods defined in [104].

applied to the human body, reducing freedom of movement or requiring the insertion of medical instruments. These techniques do not apply to measuring stress in our daily lives.

In addition to physiological measures, different self-report questionnaires have been designed to evaluate stress levels. As stress causes negative emotions and the feeling of losing control, different self-report questionnaires have been designed to evaluate stress levels, such as the Perceived Stress Scale (PSS), the State-Trait Anxiety Inventory (STAI), and the NASA Taskload Work Index (NASA-TLX).

2.3.3.2 Experimental Methods

With the development of smart wearables, sensors have been developed to evaluate electrodermal activity (EDA) and HRV. Several authors [115, 286, 285] studied the capabilities of using HRV measurements of the Apple Watch to distinguish between mental stress and relaxation, however, they come to different conclusions. Other mobile applications have been developed to evaluate stress all day long. In [52] physical monitoring was combined with mental stress monitoring using a wristband and sensors of the smartphone in use.

Clarke et al. [46] developed a smartphone application that besides detecting stress using a Microsoft Band 2, also suggests micro interventions. Although smartphones in combination with wearable devices are ideal to track stress continuously, average accuracies vary from 71% to 75% for binary classification of high and low stress conditions for user-specific models [269, 92, 295].

In recent years, voice features, facial expressions, eye gaze, and blink rates have gained interest as their use as indicators of stress only requires video recording. An overview of the usability of the mentioned techniques at workplaces has been evaluated in [40].

Reference [141, 254] used different features such as MFCC and pitch from speech to detect stress. Reference [114] analyzed behavioral changes using a pressure-sensitive keyboard and a capacitive mouse. His results showed that in stressful conditions, participants increase their typing pressure and the contact surface with the mouse.

2.3.3.3 Stress detection through Facial Expressions

Given stress is related to emotions, also facial cues have been used to detect stress. Lerner et al. used facial expressions that are linked to fear and anger as features [150]. She found that people react differently to the same stressor (stimulus causing stress), some showing fear and others anger. Also in [91] facial expressions were used to detect whether a person was stressed during driving or not by comparing the presence of anger and disgust, against the remaining emotions. In a similar study, cognitive load during driving was detected using facial AUs and the correlation between them [303]. In our work [294] described in Chapter 4 we were the first to use FACS and train different machine learning algorithms to detect stress solely from facial expressions. We showed that subject-dependent models were able to detect stress with an accuracy of up to 91% while subject-independent models reached an accuracy of 74% using Random Forest. Since then, different researchers created their own datasets to train machine learning models for recognizing stress from AUs [96, 266, 194]. Unfortunately, the datasets are not public.

Creating high-quality stress datasets requires a carefully designed experimental protocol. [96] collected data from 24 participants who underwent different stress-inducing activities such as social exposure, emotional recall, and mental workload. [194] and [31] collected data from 40 and 62 participants respectively during the Trier Social Stress Test also collecting neuroendocrinological stress levels, such as saliva cortisol. [266] on the other hand created a dataset of 34 subjects participating in online video meetings with self-reported stress levels. To the best of the authors knowledge, the video dataset used in this work (Chapter 4.2) is unique in the size of over 110 participants and in the careful experimental design including HRV and BP in addition to self-reported stress levels.

The features used in previous work studying AUs ([266, 194, 31, 294]) have been mainly obtained with the OpenFace toolbox [20]. [96] on the other hand, trained their own AU detector with deep learning methods. While [266] and [96] focus in their work on obtaining high classification accuracies for stress, [194] studies the differences in the performance of

ML models depending on the type of stress labels used, e.g. self-reported, live-observed, video-annotated, and neuroendocrinological. Besides obtaining an accuracy of 81.1% with an SVM, [96] performed the nonparametric Wilcoxon signed-rank test on the AUs, concluding that all AU intensities are higher during stress. [31] on the other hand, focused solely on statistical tests showing the quartiles of each AU during different experimental phases. Also, AU occurrence was analyzed, concluding that AU 5 (upper lid raiser), 7 (lid tightener), and 10 (upper lip raiser) occur more often during stress. So far, our community lacks research in which facial activity patterns are analyzed and defined, moving beyond ML classification and statistical analysis.

2.4 Facial Expressions during Enthusiastic Presentations

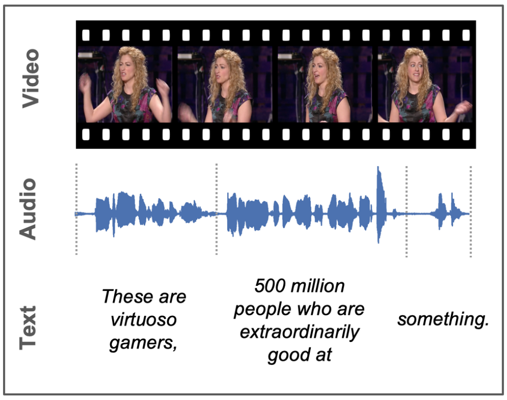

Video
Audio
Text
These are virtuoso gamers,
500 million people who are extraordinarily good at
something.

Figure 2.8: An enthusiastic sample from the Entheos dataset, showing aligned video frames, audio, and text.

Although different emotional constructs such as *anger* and *happiness* have been studied extensively in the field of natural language processing (NLP), more fine-grained emotional expressions such as enthusiasm or charisma are relatively unexplored. Such models and datasets can benefit different areas of NLP and AI. Multimodal human-machine interaction can be more effective if systems can find a deeper understanding of more complex emotional responses or generate appropriate emotionally-aware communicative presentations. Given the importance of enthusiasm in teaching [30, 304], for instance, researchers are studying the effect of virtual agents and robots that can behave in an enthusiastic manner [153, 154, 227]. The current research is far from generating natural enthusiastic behavior.

Although previous research results in psychology, education, and business have studied the importance of enthusiasm in communication [30, 229, 132, 10], it is relatively unexplored in the NLP and dialogue literature. In Chapter 5, we introduce the first multimodal dataset labeled with levels of enthusiasm following the definition that [134] provided and propose to see enthusiasm as a conversational expression. In this chapter, we provide an overview of already defined conversational expressions, the existing definitions of enthusiasm and charisma, which are related.

2.4.1 Conversational Expressions

In Chapter 2.2, we have described how emotions have been studied and how facial expressions allow us to recognize the basic emotions. In this chapter, we want to elaborate on how facial expressions also serve as communicative information channels, a fact established in Darwin's work on emotions in 1872 [53] (see also [85, 226]). This is important because communicative facial expressions are more often used than emotional expressions [86]. Although [66], and [24] elaborated theoretical attempts to disentangle emotions from conversational expressions, strong empirical evidence for these differences is still missing to date [131]. In the following, we will describe existing work on conversational expressions.

In 2004, [51] studied nine conversational expressions, performed by six actors: agreement, disagreement, disgust, thinking, pleased/happy, sadness, pleasantly surprised, clueless (as if the actor did not know the answer to a question), and confusion (as if the actor did not understand what was just said). The recorded sequences were shown to seven study participants who chose the most fitting expression out of the list of the nine expressions and a "none of the above" option. This non-forced choice methodology has been shown to be highly correlated with other identification procedures (e.g., free description of the expressions) [84]. In addition, it offers advantages over other methodologies, including the avoidance of the inflated accuracy ratings found in the absence of a "none of the above" option (i.e., in forced-choice tasks) and avoiding the subjectivity found when experimenters must categorize and analyze free description results. Overall, the participants were very good at identifying the expressions even though they did not know the actors and actresses and had no conversational context [51].

In 2012, [131] created the first database with 55 emotional and conversational expressions, which were recorded in a laboratory setup with subjects posing facial expressions after being given the context of the situation they were supposed to be in. Some of the conversational expressions are: agree (several types), bored, annoyed, confused, disbelief, don't know, compassion, etc. (a complete list can be found here[1]). A free naming method was employed to find generic terms that best summarize the obtained naming answers for facial expressions without clear labels. The created MPI Facial Expression Database[2] consists of more than 18800 samples of video sequences from 10 female and 9 male models displaying various facial expressions recorded from one frontal and two lateral views (see examples in Fig. 2.9). It contains statically and dynamically displayed facial expressions and is an important resource for studying conversational expressions. However, the database has not yet been used to analyze facial expressions using FACS or other measures.

Reference [42] confirmed most of the expressions from the MPI Facial Expression Database in 2018 while creating another database with 62 conversational expressions. However, objective metrics such as the analysis of active facial action units from FACS were not performed on this extended database. Instead, user studies confirmed whether the facial expressions were recognizable when the situational context was given.

2.4.2 Enthusiasm

Limited work exists on the automatic detection of enthusiasm and has been mainly done in the text domain. Reference [117] worked on the detection of enthusiasm in human text-based dialogues, using lexical features and word co-occurrences with conditional

[1] https://doi.org/10.1371/journal.pone.0032321.s002

[2] https://www.b-tu.de/en/graphic-systems/databases/the-large-mpi-facial-expression-database

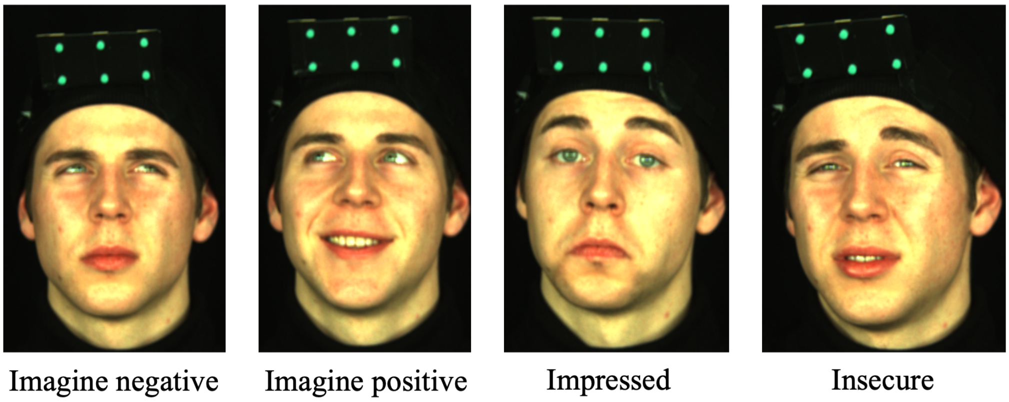


Figure 2.9: Samples from the MPI Facial Expression Database [131]. From left to right, the pictures show an actor expressing conversational expressions on imagining something negative and positive, being impressed, and being insecure.

random fields in order to distinguish enthusiastic utterances from non-enthusiastic ones. They defined enthusiasm as “the strength of each participant’s desire to continue the dialogue each time he/she makes an utterance”. In Chapter 5, we instead combine different modalities and features to detect enthusiasm and we define an enthusiastic speaker as “stimulating, energetic, and motivating” [134]. Reference [272] also worked with human-to-human conversational dialogues and annotated dialogue acts (DAs) and rhetorical relations (RRs) on a sentence-level. An enthusiasm score in the range of 10-90 was given without providing examples to the annotators. The relationship between DAs, RRs, and enthusiasm was analyzed based on the frequencies. They found that affective and cooperative utterances are significant in an enthusiastic dialogue. We detected RRs automatically and trained a feed forward network to classify enthusiasm in three levels: monotonous, normal, and enthusiastic. During data annotation, examples for each category were available as references. Twitter data have also been used to detect enthusiasm. Reference [186] created a dataset with enthusiastic and passive labels. Enthusiastic tweets had to include personal expression of emotion or call to action, whereas passive tweets lacked clear emotive content or call to action. They trained logistic regression models using salient terms. In Chapter 5 we evaluate emotional expressions in several modalities. We use acoustic features that relate to emotion such as pitch and voice quality, and also Facial Action Units extracted from videos which measure the intensity of different facial expressions.

2.4.3 Charisma

Enthusiasm is also a trait that can be displayed by charismatic speakers [261], which in addition are perceived as competent, passionate, and self-confident [192]. Charisma is a desired trait for leaders in economy and politics [10, 56] because it can influence followers to undertake personally costly yet socially beneficial actions. Reference [193] has investigated the prosodic attributes of charismatic speakers. They analyzed pitch level, pitch variation, loudness, duration of silence intervals, etc and concluded that charisma can be trained as far as melodic features are concerned. In addition to analyzing the relationship of different attributes with enthusiasm, we also trained a model that can distinguish between different levels of enthusiasm (see Chapter 5).

Although sentiment analysis and emotion detection have been studied extensively in unimodal and multimodal frameworks as shown in several surveys [165, 93, 247, 264] there is a gap in the analysis, detection and generation of enthusiastic behavior. Our dataset (Chapter 5.2) will allow to extend the work in understanding human behavior and also generate more natural virtual agents [304, 133, 154, 292].

2.5 Facial Expressions in Sign Languages

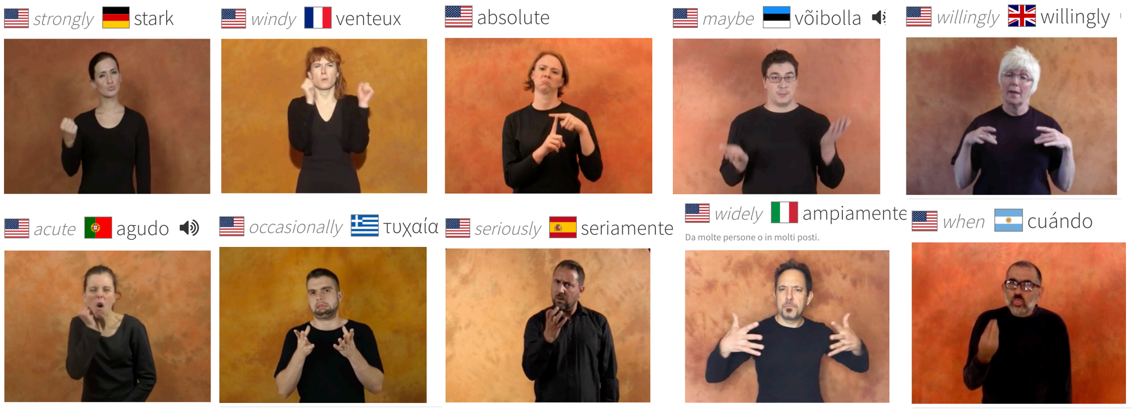


Figure 2.10: Different examples of modifiers signed in different sign languages [262]. Peculiar in sign language is that certain words would lose meaning if facial expressions were ignored.

Contrary to common belief, there is not one universal sign language, but rather over 300 different sign languages [138], such as American Sign Language (ASL) and German Sign Language (DGS) which have emerged in different Deaf communities and evolved over the years. Although linguists have shown in the last decades that sign languages are as sophisticated as spoken languages to convey information [263, 128, 232, 282], the use of sign language was suppressed for over 100 years due to the resolutions from the "Milan Conference" of 1880 [90]. At that time, oralist proponents such as Alexander Bell whose wife was deaf, advocated that learning to speak and lipread was more beneficial for the societal inclusion of deaf people [26, 191]. Recent studies, however, show that it is much more beneficial for DHH to learn sign language early in childhood to permit adequate development of communication skills [287, 283, 105].

In sign language, facial expressions are fundamental linguistic elements of communication, playing different roles. They can carry grammatical information, differentiate lexical items, participate in syntactic construction, contribute to intensification processes, and convey affective states. Although algorithms to recognize and generate sign language with avatars have evolved in recent years, the major focus has been on manual components of sign language, neglecting facial expressions. Fig. 2.10 shows how facial expressions are employed in different sign languages to communicate modifiers. Given that sign language linguistics is a relatively recent field of study, the phonological characteristics of facial expressions in sign language are still understudied.

In the following, we will provide an overview of the importance of sign language research in computer science (Chapter 2.5.1). We will also elaborate on three challenges in including grammatical facial expressions in automatic sign language translation: different roles of FEs in sign language (Chapter 2.5.2), available notation systems to describe sign language (Chapter 2.5.3), and existing sign language datasets with FE information (Chapter 2.5.4). We end this section with an overview of previous work using FEs in

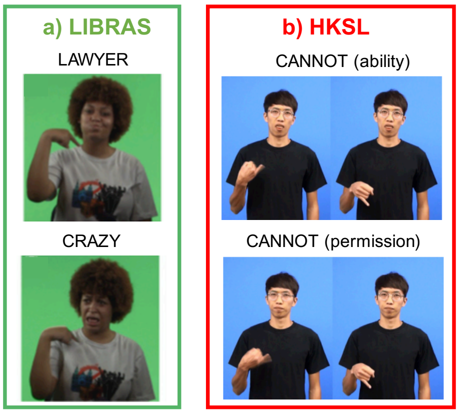


Figure 2.11: Grammatical Facial Expressions of Homonomy (GEH) are essential to distinguish the meaning of signs with the same manual gesture. In a) the signs LAWYER and CRAZY are performed with the same manual gesture in Brazilian Sign Language (LIBRAS) but in the eyes, eyebrows, and the open mouth [252]. Similarly, in b) CANNOT (ability) and CANNOT (permission) in Hong Kong Sign Language (HKSL) are executed with the same manual sign but differ in FEs [270].

automatic sign language translation (Chapter 2.5.5).

2.5.1 Overview

Approximately one in every thousand EU citizens is Deaf or Hard of Hearing (DHH) and uses one of the 31 national or regional sign languages as their first language [204]. Nevertheless, the lack of sign language users in the hearing population creates communication barriers in daily situations, such as in access to health, legal, or even school services [176, 169]. Insufficient interpreters, create the need for assistive technologies to bridge the communication gap and enhance accessibility for the DHH community. EASIER[3] and SIGNON[4] are examples of European research initiatives from Horizon 2020, intending to develop sign language translation mobile applications using signing avatars. Still, the visual-gestural multimodality of sign language and the lack of annotated data bring several challenges requiring multidisciplinary collaborations to bring forward novel and creative solutions.

In contrast to spoken languages which use sequential vowels and consonants to form words, every sign can be broken into four manual characteristics: handshape, location, movement, and orientation. Non-manual components such as facial expressions and and body movements are the fifth aspect of sign language phonology [81]. All the minimal components of signs occur sequentially and simultaneously, and slight changes in one of them can have an impact on the entire meaning of the sign (see Fig. 2.11).

In sign language, facial expressions are fundamental linguistic elements that can carry grammatical information, differentiate lexical items, participate in syntactic construction,

[3] https://www.project-easier.eu/

[4] https://signon-project.eu/

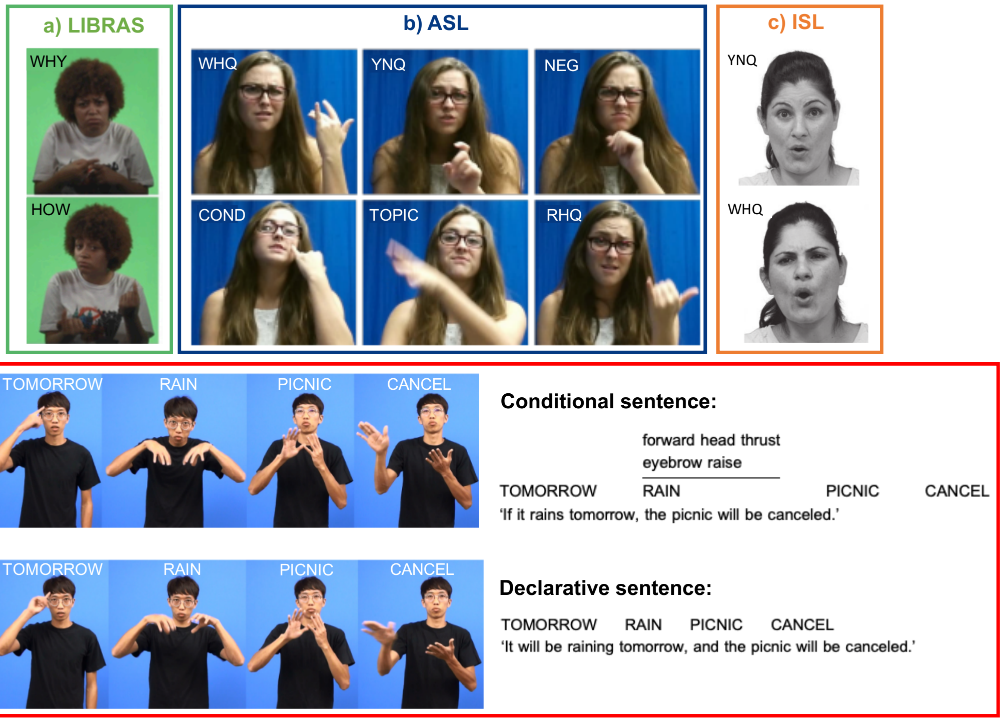


Figure 2.12: Grammatical Facial Expressions of Sentence (GES) define the sentence type, such as WH-questions (WHQ), Yes/No questions (YNQ), doubt questions, rhetorical questions (RHQ), topics, negatives (NEG), affirmatives, conditional clauses (COND), focus and relative clauses. The figure shows how GES are used in a) Brazilian Sign Language (LIBRAS) [253], b) American Sign Language (ASL) [281], c) Irish Sign Language (ISL) [230], and d) Hong Kong Sign Language (HKSL) [270].

contribute to intensification processes, and convey affective states. [207, 184]. Although facial expressions (FEs) have grammatical roles in sign languages, they are still understudied in many [197, 29]. The field of sign language linguistics is quite young compared to spoken languages. Due to resolutions of the "Milan Conference" in 1880, sign languages were degraded as inferior methods of communication, and for over 100 years deaf students were forced to learn how to speak instead of using sign language [90]. Only since 1960 with the linguistic works of William Stokoe, have sign languages started to be recognized as complete languages [263]. Nevertheless, it still required several decades for sign languages to become legally recognized as official languages – Dutch Sign Language for instance was legally recognized in 2020, and Italian Sign Language in 2021 [1].

The importance of sign languages has gained increased attention during the COVID-19 pandemic, a period during which it was essential to keep the entire population up-to-date on the public health situation [164, 215, 277]. Since then, according to the Bureau of Labor Statistics, employment of ASL interpreters and translators is projected to grow by 20 percent from 2021 to 2031 [2]. The awareness of the necessity to create automatic sign language translations has also increased in the computer science field. There has been tremendous progress from recognizing single signs [249, 6, 284, 216] to generating sentences automatically using photo-realistic avatars [233, 235, 217, 278]. However,

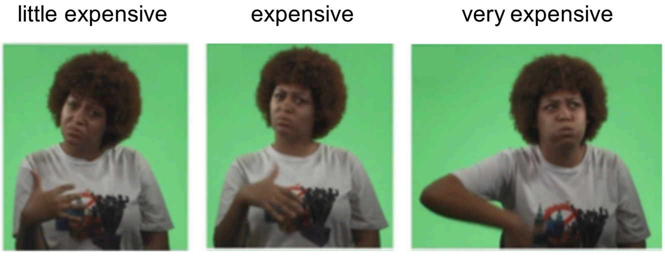


Figure 2.13: Grammatical Facial Expressions of Intensity (GEI) take the role of quantifiers. Above in LIBRAS the same sign EXPENSIVE is accompanied by different FEs which allow us to distinguish between LITTLE EXPENSIVE, EXPENSIVE, and VERY EXPENSIVE [251].

compared to spoken language resources the size and number of available annotated sign language datasets are limited. Public sign language datasets with annotated meanings of grammatical facial expressions are, at the moment of writing, not existing [218, 217, 252, 281].

2.5.2 Roles of Facial Expressions in Sign Languages

FEs are mainly associated with expressing emotions [68]. However, in sign language FEs also play a linguistic role. In addition to expressing affect, they can serve as phonetic and phonological elements [231] and have grammatical functions [76]. In this work, we follow the proposal by [209] that there are four different types of Grammatical Facial Expressions (GFEs) that occur at specific sentence points or are associated with a particular sign execution [7, 252].

Grammatical Facial Expressions of Homonymy (GEH) help to define the meaning of a sign. Without the characteristic GEH a sign is incomplete and cannot be distinguished from other signs with the same manual signal. Fig. 2.11 shows two examples in LIBRAS and HKSL in which different FEs are essential to distinguish the meaning of signs with the same manual gesture.

Grammatical Facial Expressions for Sentence (GES) define the type of sentence that is being signed, such as WH-questions, Yes/No questions, doubt questions, topics, negatives, affirmatives, conditional clauses, focus and relative clauses. Fig. 2.12 shows how facial expressions define the sentence types in different sign languages. Reference [251] has studied GES through analyzing facial action units and training a machine learning model to recognize them in LIBRAS. Reference [29], on the other hand, studied GES in ASL using manual annotation of non-manuals, such as head turns or head tilts.

Grammatical Facial Expressions of Intensity (GEI) differentiate the meaning of the sign assuming the role of a quantifier. Fig. 2.13 shows how FEs allow signers to describe different levels of EXPENSIVE. GEIs add semantic meaning to the manual sign and modify actions, having the role of modifiers. Unfortunately, these FEs are still understudied in sign languages.

Grammatical Facial Expressions of Norm (GEN) are part of the sign by norm and have the function to complete the manual sign. When a GEN sign is performed without the facial expression, the signal loses its meaning.

2.5.3 Sign Language Notation Systems

Sign languages are visual-spatial languages and are articulated by using the hands, face, and other parts of the body. Given the characteristic of transmitting information through different channels in parallel, different notations have been developed to capture in written form the execution of signs (see Fig. 2.14).

Similar to the International Phonetic Alphabet (IPA), the **Hamburg Notation System (HamNoSys)** [106] provides a direct correspondence between symbols and gesture aspects, such as hand location, shape, and movement to capture the complexity of signs in written text. Reference [255] proposes a method to annotate automatically existing SL datasets and [11] animates an avatar based on HamNoSyS, however, HamNoSys cannot represent facial expressions and mouthing (production of visual syllables with the mouth while signing).

SignWriting [267], on the other hand, uses iconic symbols for handshapes, orientation, body locations, facial expressions, contacts, and movements to represent words in sign language (SL).

Nevertheless, most SL datasets are transcribed using **gloss annotation**. Glosses are written in small capitals to be differentiated from spoken words and capture the main meaning of signs. Unfortunately, glosses do not provide any information about the execution of the sign, only about its meaning. Although, gloss annotation can include information about the FEs as in Fig. 2.12 d) above the gloss RAIN, the majority of the SL datasets lack FE information. Additionally challenging is that glosses are taken from a different language, providing only an approximation of the sign meaning and representing only one possible translation. Despite the existence of the well-studied Facial Action Coding System (FACS) created by [74], it has not yet been employed by sign language linguists to automatically and systematically create notations describing linguistic FEs.

2.5.4 Sign Language Datasets with Facial Expressions

Although SL datasets with non-manual labels, such as head turns, head tilts, raised eyebrows, etc. exist [29, 35, 125], the grammatical function of the facial expressions is not annotated. At the time of writing, the dataset created by [253] is an isolated example with manually annotated grammatical FEs of sentences. The dataset was manually annotated by FACS experts, one deaf and one hearing LIBRAS speaker. Although recently novel large-scale ASL datasets have been created such as How2Sign and Youtube-ASL, facial video data is not available [62] or the videos only contain English captions [279]. Given the rare expertise of sign language linguists, the manual annotation of the grammatical role

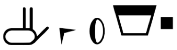
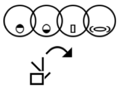
Notation System	"three" in DGS
HamNoSys	
SignWriting	
Gloss	DREI

Figure 2.14: Different notations for sign language. HamNoSys [106] represents the execution of the signs in terms of the manual phonological components but lacks facial movement description. SignWriting [267] includes information about the execution of facial and manual movements. Frequently used gloss annotation lacks information about the execution of the signs.

of FEs in sign language datasets as well as the intensity of active FAUs by FACS experts is cost-intensive and time-consuming.

2.5.5 Automatic Sign Language Translation with Facial Expressions

Although FEs have clear grammatical roles in sign languages, we are far from having sign language translation (SLT) systems able to correctly capture, interpret, and produce them. Given that SLT systems are based on machine learning, they are bottlenecked by the data used for training. Nevertheless, impressive progress has been achieved in the recent years in SLT [43, 308, 307, 302] (see this survey for more information [195]). Systems have moved from word-level rule-based animated avatars [50, 130, 174] to high-resolution photo-realistic continuous sign language videos [235, 233, 235, 217, 278]. Although human evaluation is being used to evaluate the quality of SLT systems, so far the focus mostly lies on the quality of the manual signs and not the appropriate use of grammatical facial expressions [62, 235].

3

Data Bootstrapping for Facial Expression Analysis

As mentioned in chapter 2.1.1, humans have more than 40 facial muscles that allow us to reproduce a variety of facial expressions. Compound emotions exist besides the basic universal emotions, and many more emotions have been defined in different emotion theories (chapter 2.2.3). We have also shown how facial expressions allow us to make our communication more efficient through conversational expressions (chapter 2.4.1), and that facial expressions even have linguistic functions in sign language (chapter 2.5.2). Although several datasets have been created in the last decades to study the basic emotions, there is a lack of datasets for studying other emotions, conversational expressions, and linguistic facial expressions in sign language. In this chapter, I want to present different methods that have been used to create datasets to study facial expressions. Given the variety of facial expressions that are still unstudied, conventional methods to create datasets can require considerable amounts of time and financial resources. In this thesis, I want to present my method of data bootstrapping for facial expression analysis. Compared to previously used methods, data creation is more efficient and requires fewer resources.

3.1 Ideal Method

If resources were unlimited, the ideal method to study facial expressions would be data-driven and brute-force. To achieve this goal, it is necessary to annotate very large datasets, rich in different labels with annotators from different cultural backgrounds. However, as described in the previous chapter, many emotions, conversational expressions, and linguistic facial expressions in sign language exist. An ideal dataset would have labels for a variety of emotions and conversational expressions. Nevertheless, such endeavors would be very expensive and time-consuming.

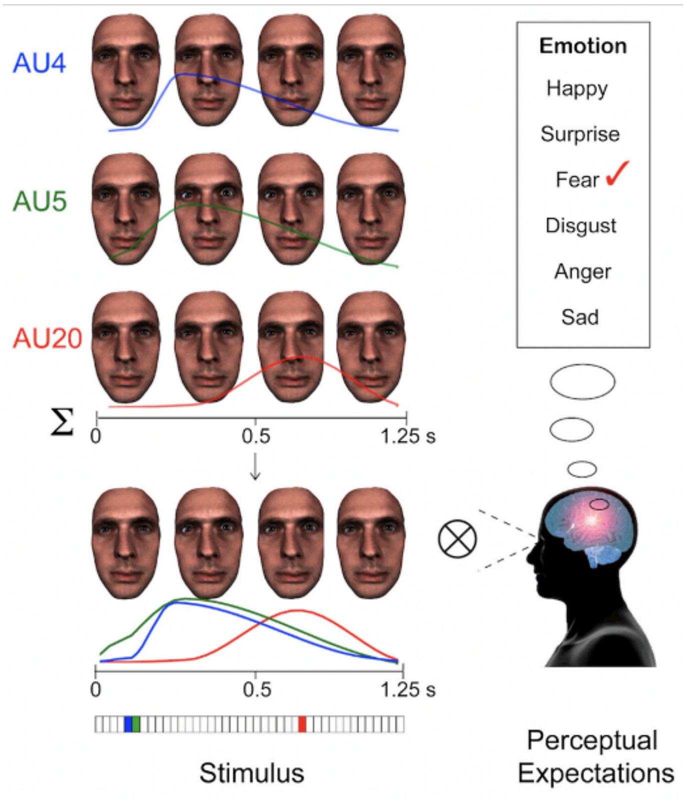


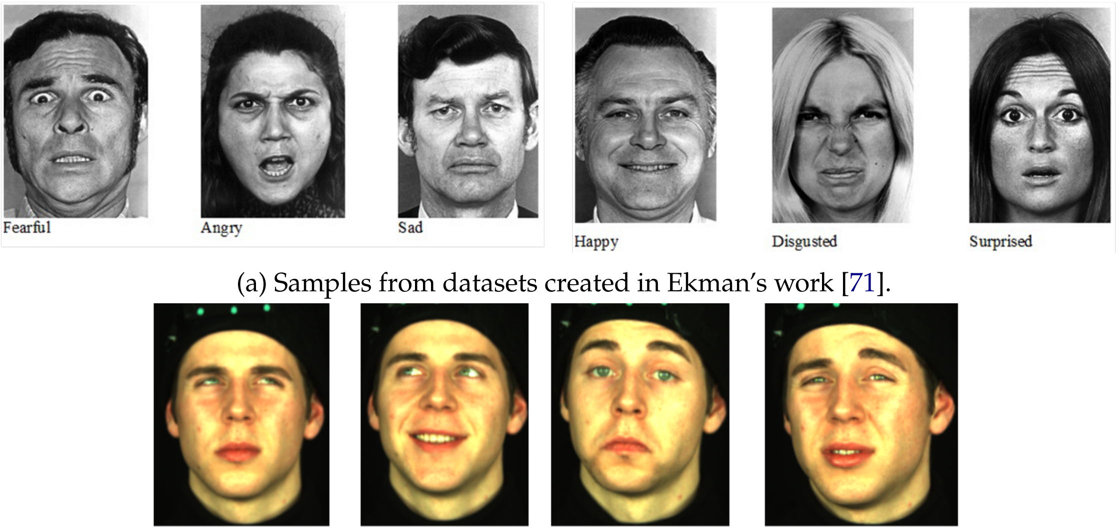
Figure 3.1: Example of how simulated facial expression can be created by controlling which AUs should be activated and with which intensity over time. The annotator is asked to select an emotion that is recognized from the simulation [44].

3.2 Generation of simulated Data

Instead of annotating data from humans, another possibility is to simulate facial expressions with human graphical models. This allows to have controlled facial expressions which can be displayed with different appearances and genders [44]. The controlled facial expressions can be shown to annotators from different cultures. A disadvantage is the missing context and that unnatural expressions might be generated. It can also be time-consuming. Fig. 3.1 shows an example where the activation of different action units can be controlled. The animation is shown to annotators who are asked to select the emotion they recognize out of a list.

3.3 Actor-Posed Datasets

The most common method found to create emotion databases is creating actor-posed data. You tell the actor what expression to show and annotators are asked to identify the expression. Datasets from laboratory settings have been criticized because the expressions do not occur naturally. In addition, it is time-consuming and expensive. This is how Ekman defined the basic emotions and also how the MPI database of conversational expressions was created (see Fig. 3.2).



(a) Samples from datasets created in Ekman's work [71].

(b) Samples from the MPI facial expression dataset [131].

Figure 3.2: Dataset examples depicting actors who posed given emotions (a) and conversational expressions (b). Annotators are asked to select the emotion or conversational expression they recognize out of a list or through an open answer.

Figure 3.3: Examples of the AFEW in-the-wild dataset on basic emotions. Samples are extracted from TV shows and movies and annotated with the basic emotions to obtain ground truth labels.

3.4 In-The-Wild Datasets

To have more natural expressions, in-the-wild datasets have been created out of movies and TV shows. Although the interactions are acted upon, they occur within context and not in an isolated manner. Here excerpts are created and manually annotated. This allows for expressions in context with other modalities but the manual annotation of each excerpt is still time-consuming and expensive. Some examples are the AFEW and SFEW datasets for basic emotions shown in Fig. 3.3.

3.5 My Method on Data Bootstrapping for Facial Expression Dataset Creation

In this thesis, I present my data bootstrapping method for facial expression analysis. The idea is to take existing datasets, adapt them, and use semi-automatic algorithms to

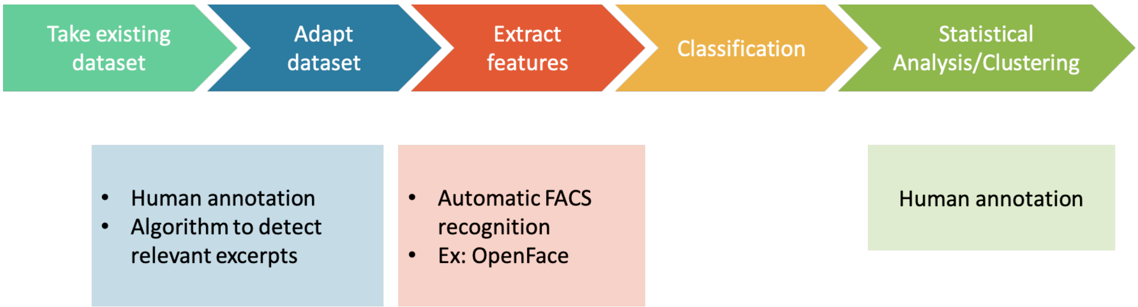


Figure 3.4: Methodology to study emotional, conversational, and linguistic facial expressions through data bootstrapping. It comprises five steps: taking an existing dataset, dataset adaption, feature extraction, classification, and statistical analysis or clustering.

find relevant expressions, which are then confirmed through annotation. This method is faster and cheaper than the previously presented ones. It allows us to start exploring emotional, conversational, and linguistic FE that have not been studied yet with few resources. However, the annotation is focused on a very reduced number of labels. Also adapting the dataset for one's needs requires some engineering.

Fig. 3.4 shows the overview of the data bootstrapping method. It is composed of five steps. 1) First, an existing dataset that suits the domain and problem is chosen. *Example: Sign language data can be obtained from news videos with sign language interpreters.* 2) Next, the dataset needs to be adapted. This might require human annotation to obtain some ground truth labels or the use of algorithms that detect relevant excerpts based on other available modalities such as text transcripts. *Example: Text transcripts can be obtained from the news speaker. Human annotation is needed to align sign language with the text transcripts. Nevertheless, through consultation with sign language experts, it might be possible to identify features that allow to write an algorithm able to align the data.* As our focus is on facial expressions, 3) the features we use are facial action units. Their high-level characteristic allows humans to interpret the results and verify if they make sense. As manual annotation is very expensive, I use an automatic FACS recognition algorithm, namely OpenFace. If the data is multimodal, it can also be beneficent to extract features from other modalities (see chapter 5).

4) Next, the extracted features are used to train a classifier or generative model, depending on the problem, to evaluate if the action units of the expressions allow to find patterns. 5) The last step is to obtain more understanding of what occurs during the expression that is being studied. I suggest using statistical analysis or clustering. Also here some human annotation might be necessary.

To show the generalization of this method, I applied it to different roles of facial expressions in the remainder of this thesis: 1) chapter 4 stress (emotion), chapter 5 enthusiasm (conversational expression), chapter 6 quantifiers in sign language (linguistic).

4

Facial Expressions during Cognitive Stress

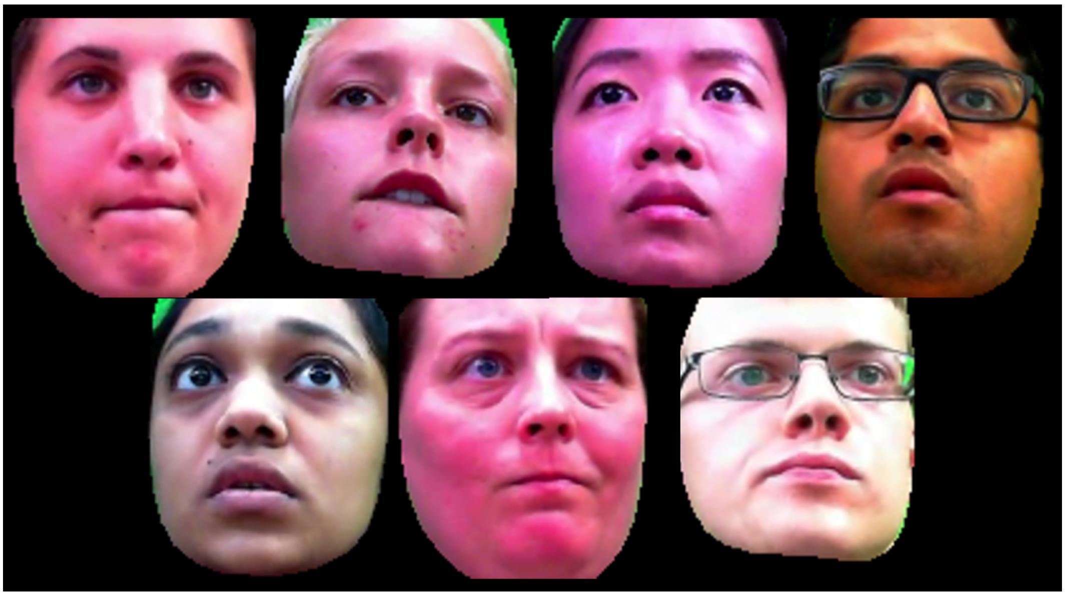


Figure 4.1: Examples from our seven different stress facial activity patterns, defined through semi-automatic methods (Chapter 4.4). We named the clusters as follows (from top left to bottom right): *"pressed lips"*, *"biting lip"*, *"stoic face"*, *"open mouth"*, *"lifted eyebrows"*, *"frowning"*, and *"dimpler"*. All samples used to define the seven facial activity patterns of stress were identified by user study participants as showing stressed individuals.

4.1 Introduction

In this chapter, we want to evaluate if a combination of different Action Units (AUs) from the Facial Action Coding System (FACS)[75] can be used to indicate the presence of cognitive stress. To the best of our knowledge, this work is the first to use a variety of facial action units to detect cognitive stress on a dataset with more than 100 subjects. Compared to previous work, physiological data such as ECG and BP together with standard stress

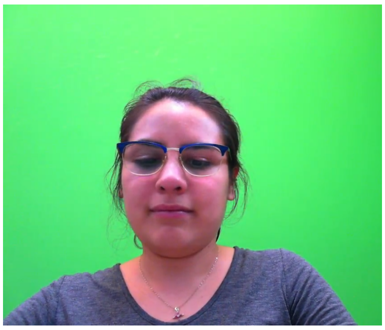


Figure 4.2: Extracted frame from one video during typing phase.

questionnaires confirm that the stressors used during the experiment, indeed caused cognitive stress in the subjects.

In Chapter 4.3, we used different classifiers to perform subject-dependent and subject-independent classifications. We show that facial action units from FACS carry relevant information, allowing the distinction from stress and non-stress states. As the results of classifiers trained subjectwise perform better than classifiers trained using the leave-one-subject-out (LOSO) method, we suspected that cognitive stress is not expressed as a universal expression.

In Chapter 4.4, we perform further analysis to understand how cognitive stress is expressed in the face. We propose a semi-automatic method to identify potential frames that express stress and perform a user study to identify facial expressions that show cognitive stress. In the end, we are able to propose seven facial activity patterns that show cognitive stress (Fig. 4.1).

4.2 Data

In this work, we utilize the video recordings of the dataset created by Lau [145] which is composed of data from 115 subjects (48 male and 67 female). The original goal of the dataset was to determine if detecting stress through keystroke dynamics is possible. Nevertheless, frontal video recordings of the entire experiment, as well as, qualitative and quantitative evidence of actual stress state during the stressor task, make the dataset a unique resource to study facial expressions during stress.

4.2.1 Recording setup

The primary experiment detailed in [145] involved capturing four distinct video streams. Three of these streams exclusively captured the subjects' keyboard typing activities from different angles: left, right, and above. The fourth stream featured a frontal view capturing the subject's face during the experiment (see Fig. 4.2), which is the specific video data analyzed in this work. The recording equipment was a Microsoft Life Studio Pro webcam with 1080p resolution, capturing at 30 frames per second. The frontal camera

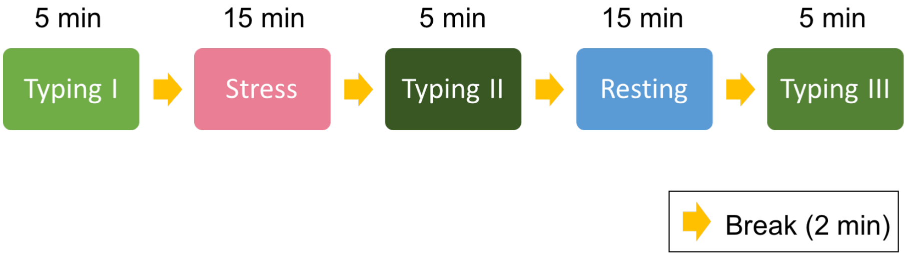


Figure 4.3: Phases of the experiment recorded on video.

was positioned beneath the monitor utilized by participants throughout the experiment. Each video has an approximate duration of one hour, with slight variations depending on the typing speed of the subject.

4.2.2 Experimental Protocol

The facial video data formed a component of a broader experiment with the primary objective of gathering typing data from subjects under both neutral and stressed conditions. The overarching aim was to determine the feasibility of discerning between subjects' typing in neutral and stressed states. Alongside typing data, additional ancillary information such as BP and ECG readings was collected to assess participants' stress levels induced by the stressor.

All participants in the experiment followed the same protocol. Initially, each subject underwent a 30-minute rest period to establish a neutral baseline. Next, the subject provided a neutral typing sample ("*Typing 1*"), capturing the initial neutral face video. Subsequently, the subject engaged in a 15-minute stressor task, involving a multitasking exercise accompanied by a negative social evaluation from the experimenter. After completing the stressor task, the subject provided a stress-induced typing sample ("*Typing 2*"), followed by a second 15-minute rest period to return them to a neutral state. A second neutral typing sample was then obtained ("*Typing 3*"). Between each of these phases, subjects had a 2-minute break to fill out the State-Trait Anxiety Inventory (STAI) and NASA Taskload Work Index (NASA-TLX) questionnaires regarding their stress state. The overview of the phases recorded on video is shown in Fig. 4.3.

4.3 Facial Action Units for Stress Detection

In this chapter, we propose the detection of cognitive stress, using only facial action units detected from video recordings. The advantage of using video for personal stress detection is the easy accessibility to webcams when working on computers. In contrast to previous work [150, 91], where stress was detected using facial expressions that are known to occur with fear, here no restrictions are used. We extracted 17 different Action Units (AUs) from upper-level to lower-level face frame-wise. We were able to distinguish stress

Table 4.1: Action Units (AUs) used as features. They provide intensity information (I) which varies between 0-5. or presence/absence information (P) with a binary value.

AU	Description	Prediction
AU1	Inner brow raiser	I
AU2	Outer brow raiser	I
AU4	Brow lowerer	I
AU5	Upper lid raiser	I
AU6	Cheek raiser	I
AU7	Lid tightener	P
AU9	Nose wrinkler	I
AU10	Upper lip raiser	I
AU12	Lip corner puller	I
AU14	Dimpler	I
AU15	Lip corner depressor	I
AU17	Chin raiser	I
AU20	Lip stretcher	I
AU23	Lip tightener	P
AU25	Lips part	I
AU26	Jaw drop	I
AU45	Blink	P

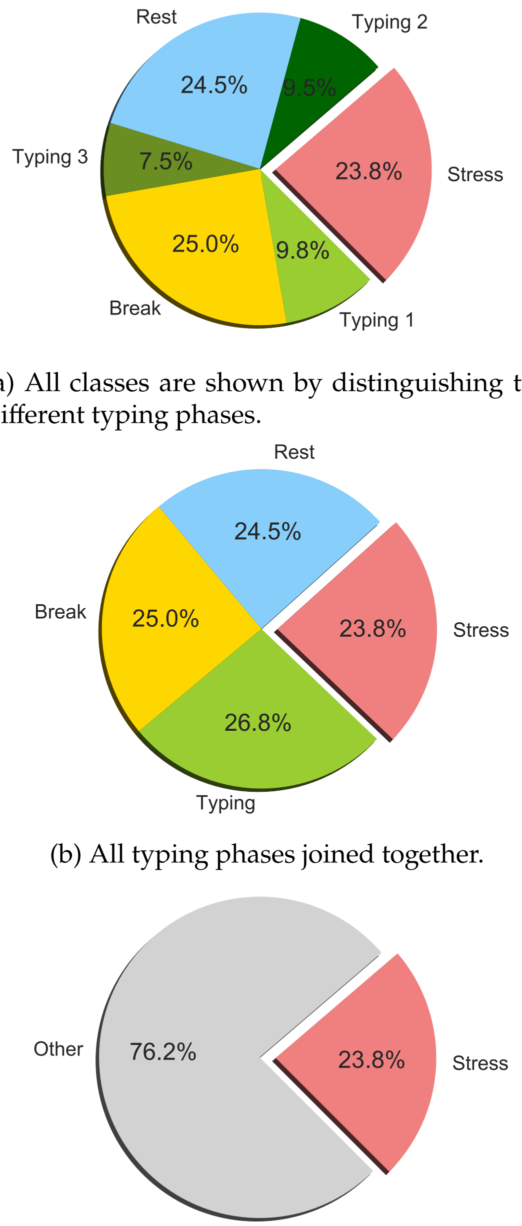
from non-stress situations with an accuracy of 74% for person-independent classification using leave-one-subject-out (LOSO). For person-dependent classification we reached an accuracy of 91% using 5-fold cross-validation. Besides detecting stress through video during work for personal tracking, this method can also be useful for security applications, such as ATM surveillance. It can also be used as a feedback tool to monitor stress during interview training as well as public speaking.

4.3.1 Methodology

4.3.1.1 Features

We focused on using a variety of facial Action Units (AUs) from the Facial Action Coding System (FACS) to distinguish stressed from non-stressed participants. Given the nature of the data, only cognitive stress can be evaluated. To extract AUs from each frame of the videos, the toolbox OpenFace [20] was used to detect 17 different AUs. The AUs used are shown in Table 4.1. Most of the AUs are intensity values between 0-5, except AU7, AU23, and AU45 which are binary (present/absent). In this work, we used for each frame only the AU values as feature (see Eq.4.1). To account for person-specific differences, the features were standardized per subject. Additionally, 2 minutes in between each phase were removed to avoid data from transitions.

$$\mathbf{f} = (f_1 = \text{AU1}, f_2 = \text{AU2}, \ldots, f_{17} = \text{AU45}) \tag{4.1}$$



(a) All classes are shown by distinguishing the different typing phases.

(b) All typing phases joined together.

(c) Label distribution for binary problem, showing stress and grouping all remaining phases together.

Figure 4.4: Label distribution for different classification problems.

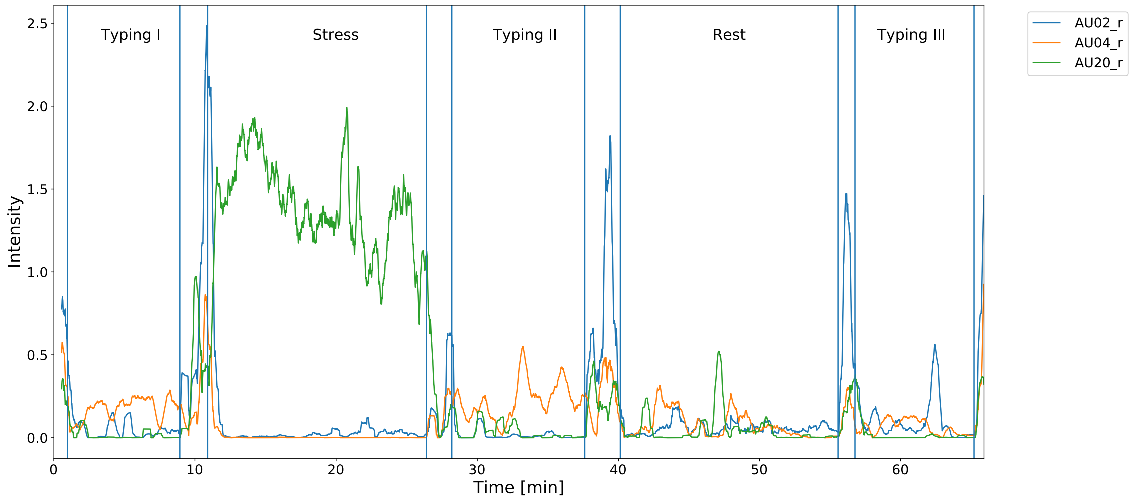


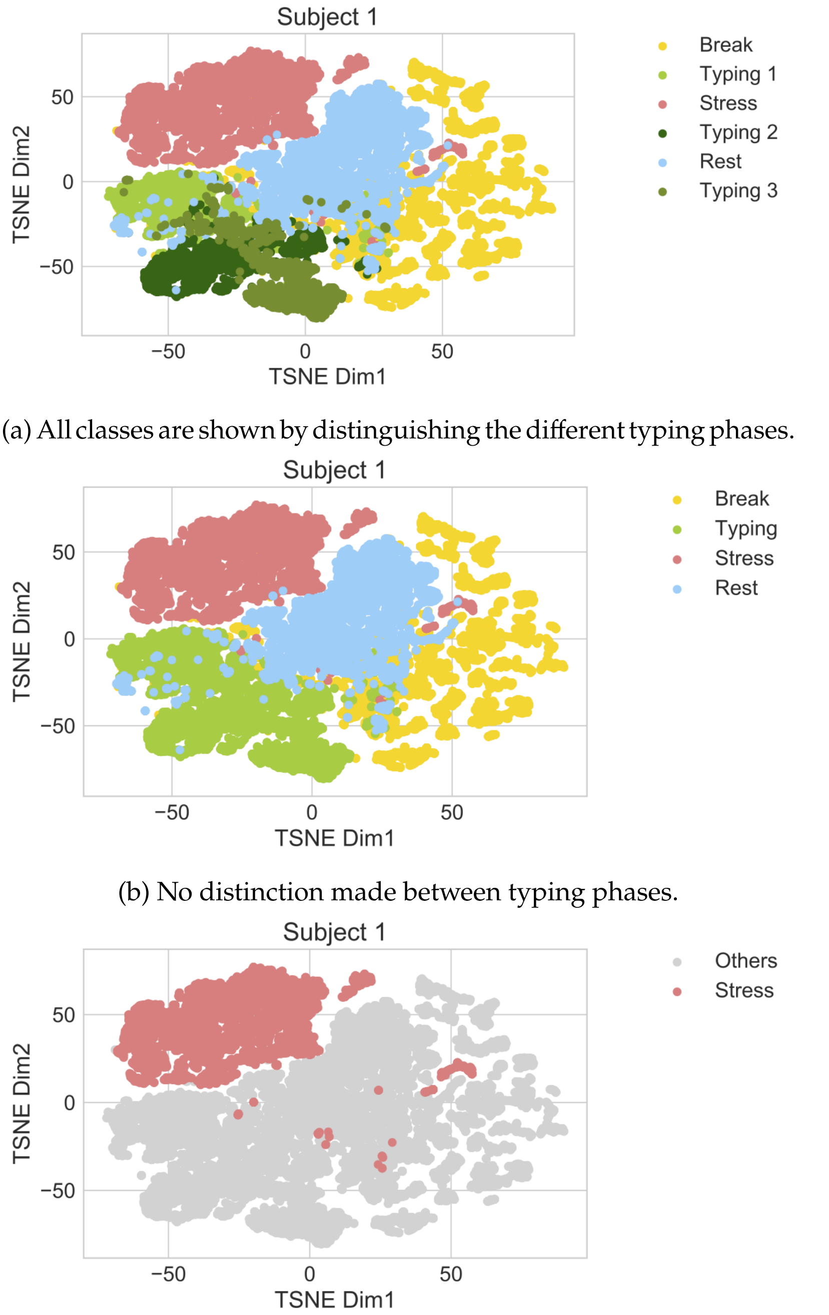
Figure 4.5: Plot showing AU2, AU4, AU20 of one subject during the entire video. The start and ending of each phase are marked by vertical lines. In between the main phases, the subjects were given a short break. However, the behavior of the shown AUs was different for different subjects. Nevertheless, correlations of different AUs with the different phases were visible.

4.3.1.2 Data Visualization

As an initial analysis, we visualized the behavior of each AU overtime per subject. Each plot showed a different behavior of the AUs. Nevertheless, for each subject different single AUs were visibly correlated with the stress phase as is shown in Fig. 4.5. For this subject, AU20 which in [159] is also described as an indicator of fear is highly correlated to the stress phase. AU2, the outer brow raiser, is activated during the breaks. Also, correlations with other phases were visible in other subjects. For that reason, we used unsupervised clustering for each subject. In Fig. 4.6 results of t-distributed Stochastic Neighbor Embedding (tSNE)[162] are shown for one subject. In Fig. 4.6a features recorded during different phases are colored differently, including the different typing phases. It is visible that the typing phase before and after the stressor are grouped separately. The final typing phase after resting, however, overlaps with the other two. In Fig. 4.6b the same data points are shown without distinguishing between the typing phases. It can be seen that differences exist between the phases. However, features during the resting phase overlap with features from the breaks in between main phases which is comprehensible. When distinguishing only between the stress and non-stress phases, we can see that some features from the stress phase overlap the other phases (Fig. 4.6c).

4.3.1.3 Classification

Given the clusters shown in Fig. 4.6, we defined three different classification problems (see Table 4.2): 1) 6-class problem, distinguishing between all phases including each typing phase, 2) 4-class problem considering all typing phases as the same, and 3) binary



(a) All classes are shown by distinguishing the different typing phases.

(b) No distinction made between typing phases.

(c) Binary problem visualization, showing stress and remaining phases grouped together.

Figure 4.6: Visualization of the data of one subject using tSNE.

Table 4.2: Labels used for different classification problems.

2 Classes	4 Classes	6 Classed
1) Stress	1) Break	1) Break
2) All others	2) Typing 1,2,3	2) Typing 1
	3) Stress	3) Stress
	4) Resting	4) Typing 2
		5) Resting
		6) Typing 3

Table 4.3: Average accuracy results for person independent and dependent classification.

Person independent			
Classifier	2 Classes	4 Classes	6 Classes
Random Forest	0.75	0.49	0.41
LDA	0.74	0.47	0.33
Gaussian Naive Bayes	0.48	0.4	0.29
Decision Tree	0.68	0.34	0.29
Person dependent			
Classifier	2 Classes	4 Classes	6 Classes
Random Forest	0.93	0.83	0.83
LDA	0.89	0.74	0.65
Gaussian Naive Bayes	0.84	0.79	0.75
Decision Tree	0.89	0.74	0.67

classification problem distinguishing between stress phases and non-stress phase. For each classification problem, the label distribution varies (see Fig. 4.4). As the label distribution is only balanced in the 4 class problem, we used weighted accuracy on all classification results. Given that the behavior of the AUs was not visibly correlated between the five subjects, we trained a person-independent model using the LOSO method. We also trained 5 different models, one per subject, to evaluate person-dependent classification using 5-fold cross-validation. For person-independent classification, the training set contained approximately 112,000 samples and testing 28,000 (total samples of one subject). During subject-dependent classification 5-fold cross-validation was used, meaning that 80% of the data of one subject was used to train and the remaining 20% of the same subject was used for testing. Different simple classifiers were used: Random Forest, LDA, Gaussian Naive Bayes, and Decision Tree.

4.3.2 Results

The results obtained with each classifier for the different classification problems are shown in Table 4.3. In the subject-independent classification, the results vary from 29% for the 6-class problem to 75% for the binary classification. The random forest (RF) algorithm performed best out of all classifiers. The distinction between stressed facial expression and not stressed achieved an average accuracy of 75%. The accuracy drops to 49% when distinguishing between stress state, break, typing, and resting. When distinguishing further the individual typing phases, the accuracy is lowest with 41%. LDA performs similarly to RF in the 2-class and 4-class problem but is 8 percent points lower in the 6-class problem.

In the person-dependent classification, the performance of all algorithms is better. The results vary from 65% for the 6-class problem to 93% for the binary problem. In this case, the best-performing algorithm is also RF with an average accuracy of 93% for the binary problem and 83% for the 4-class and 6-class problems.

As the best results were obtained using RF for person-dependent classification, we computed the confusion matrix for each of the three classification problems shown in Fig. 4.7. Fig. 4.7a shows that stress and resting phases were recognized with an average accuracy of 99% and 95% respectively. The highest misclassification occurred between similar states. 31% of the break samples were confused with resting state and 48% of typing I were misclassified as typing III. Interestingly, also 39% of typing II were confused with the resting state. In Fig. 4.7b we can see much lower misclassification between classes. The highest misclassification occurred for break samples: 17% were classified as resting state. Also, 14% of resting samples were classified as typing. Fig. 4.7c shows that 11% of stress samples were misclassified as nonstress. However, all nonstress samples were correctly classified as not showing stress.

4.3.3 Conclusion

In this work, videos collected by Lau [145] were used for the first time to detect the stress state of 115 different participants using only facial AUs. Each participant was recorded while performing different tasks. By visualizing the features over time and through a tSNE plot, it was visible that AUs contain relevant information to distinguish the tasks. We formulated different classification problems to evaluate whether it was possible to correctly distinguish between each phase. As in previous work [303], subject-independent classification achieved lower accuracy results than subject-specific classification. Our preliminary analysis suggests that facial expressions during stress are not universal.

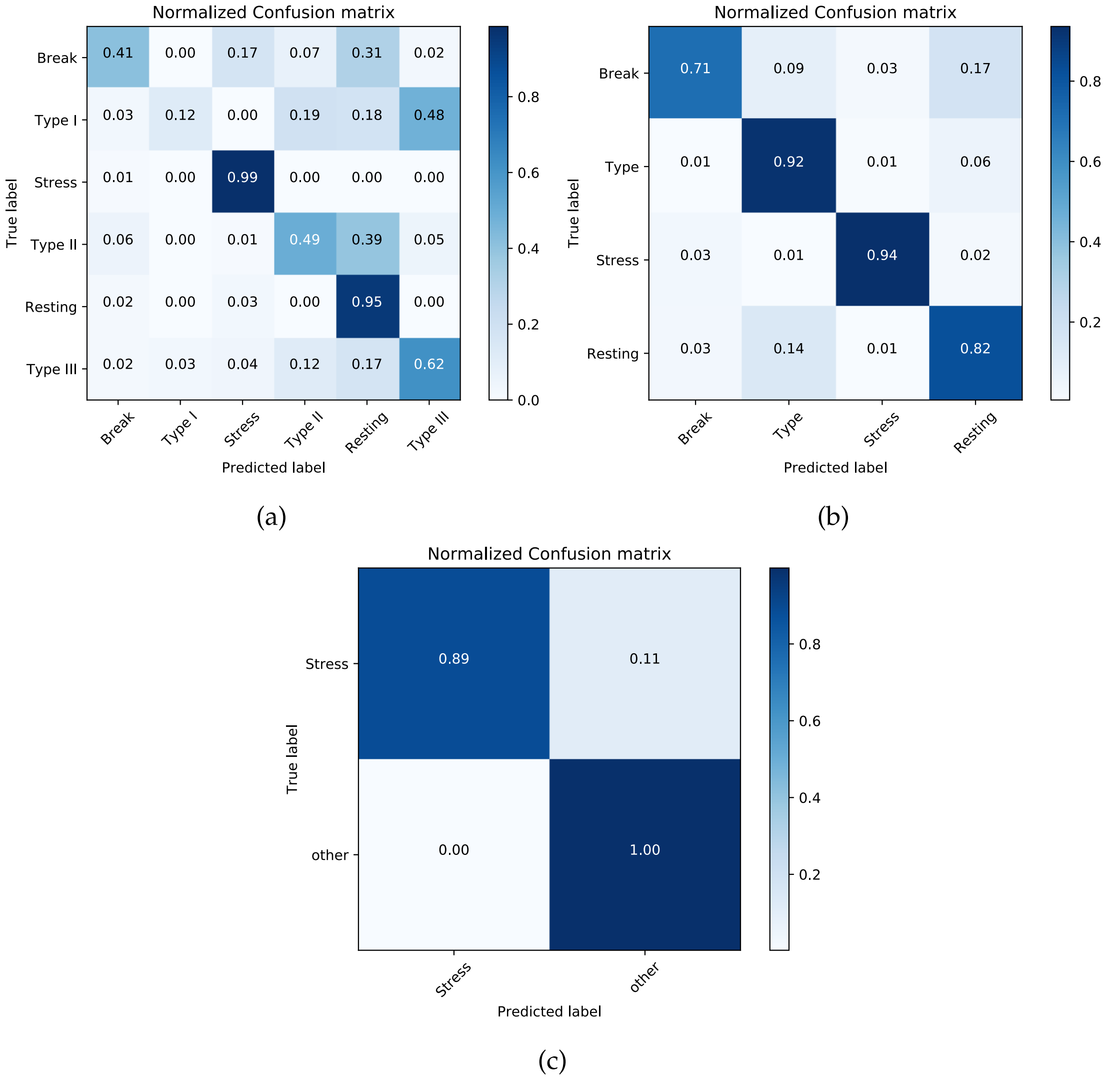


Figure 4.7: Confusion matrices for subject-wise classification using Random Forest algorithm on a) 6 classes, b) 4 classes, and c) 2 classes.

4.4 Facial Action Units during Cognitive Stress

In Chapter 4.3, we have shown that it is possible to train different classifiers able to recognize stress only through the intensities of 17 AUs. Classifiers performed with an accuracy of an average of 40% higher in the person-dependent setting compared to the person-independent setting. These results make us deduce that facial expressions during cognitive stress are not universal, but person-dependent. As discussed in Chapter 2.3, studies using AUs for stress detection, such as [294, 96, 95, 97, 31] lack a deeper analysis that includes a user study recognizing stress from the facial expressions as well as an analysis showing the relationship with emotion-related AUs for the different facial activity patterns that occur during stress. Several challenges exist in defining facial activity patterns of stress: 1) facial expressions of stress are person-specific, 2) one individual could show stress through different facial expressions, 3) the human face allows for an immense

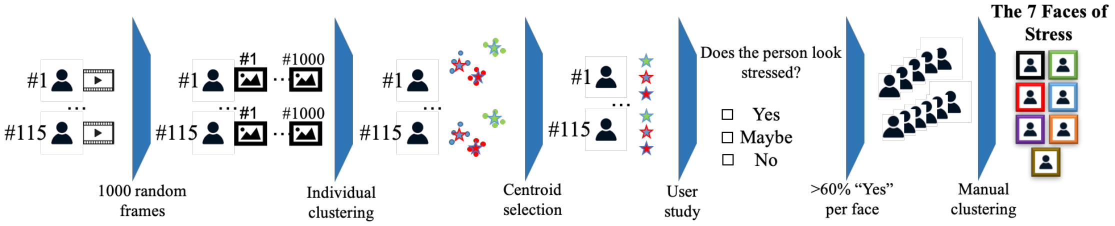


Figure 4.8: Overview of our proposed analysis methodology. Our pipeline starts with five-minute videos of 115 subjects. A random selection of 1000 frames per subject follows together with K-means clustering for each individual's data. The centroids (video frames) of each individual's clusters are selected for our user study. Only the centroids with more than 60% "Yes" annotations were selected for manual clustering of the faces. At the end of our pipeline, we obtained seven facial activity patterns of stress.

number of facial expressions, 4) we do not know which characteristics of facial expression are relevant during stress, nor 5) do we know how many facial activity patterns we are looking for.

In this chapter, we present an analysis methodology that figuratively speaking, allows us to find several needles in a haystack. We analyzed a total of 575 minutes of video recordings from 115 subjects (data described in Chapter 4.2), combining clustering with human annotation to define facial activity patterns that show stress. Our methodology does not require trained Facial Action Coding System (FACS) specialists or psychologists for data annotation. Instead, we employed automatic AU detectors, computed clusters from a large amount of data points, and used cluster centroids for human annotation. With our proposed method, we identified seven facial activity patterns during stress. As facial expressions of stress are often related to disgust and anger [91], we show the relationship between active AUs during stress and the basic emotions. With this work, we aim to provide a novel resource to recognize stress and simulate stress-related facial behavior to make human-machine interactions more natural. For this purpose, characteristic facial activity patterns during stress are made publicly available[1].

4.4.1 Methodology

The goal of this work is to define different facial expressions of stress that can be recognized as such by humans. For this purpose, we propose a novel semi-automatic methodology that allows the unveiling of relevant facial activity patterns of stress out of a large amount of data. Fig. 4.8 shows the overview of our proposed methodology. First, we randomly extracted 1000 frames per subject recording and employed K-means clustering to categorize the data for each individual. The results of the cluster analysis provided us with frequently occurring combinations of AUs during stress for each individual. Next, we used the centroids (video frames) from each individual's clusters for our user study. Only centroids that obtained more than 60% "Yes" annotations were chosen for manual clustering of the

[1] https://github.com/clviegas/SevenFacesOfStress

Table 4.4: Frequency occurrence of facial action units (AUs) during stress, typing 1, and 2 phases. AU 9, AU 10, AU 12, and AU 28 occur during less than 5% of the frames, which is why they are not used for further clustering analysis.

Action	**AU**	**Stress**			**Typing 1**			**Typing 2**		
Units	**Description**	f	m	both	f	m	both	f	m	both
AU01	Inner brow raiser	13.81	16.37	14.87	13.32	14.32	13.74	13.22	13.67	13.41
AU02	Outer brow raiser	18.1	21.15	19.37	14.28	15.45	14.77	14.93	15.82	15.3
AU04	Brow lowerer	21.19	20.87	21.06	34.45	22.75	29.57	31.07	27.75	29.68
AU05	Upper lid raiser	75.5	80.66	77.65	61.43	69.81	64.93	60.42	67.22	63.26
AU06	Cheek raiser	16.38	24.12	19.61	6.2	16.67	10.57	**4.63**	18.29	10.33
AU07	Lid tightener	41.06	41.37	41.19	35.13	35.86	35.43	32.3	31.95	32.16
AU09	Nose wrinkler	**3.27**	**3.16**	**3.22**	**2.45**	**2.66**	**2.54**	**1.9**	**3.62**	**2.62**
AU10	Upper lip raiser	**3.97**	**4.62**	**4.24**	5.14	**0.97**	**3.4**	5.29	**1.69**	**3.79**
AU12	Lip corner puller	**4.17**	**2.44**	**3.45**	**4.68**	**2.27**	**3.67**	5.41	**1.85**	**3.92**
AU14	Dimpler	6.08	5.35	5.78	8.76	13.48	10.73	9.04	12.46	10.47
AU15	Lip corner depressor	13.19	20.09	16.07	7.69	13.43	10.09	7.5	15.81	10.97
AU17	Chin raiser	27.3	27.38	27.34	17.72	16.87	17.37	16.51	18.84	17.48
AU20	Lip stretcher	12.0	19.05	14.94	7.56	8.85	8.1	7.21	11.16	8.86
AU23	Lip tightener	24.45	33.44	28.2	29.75	43.57	35.52	23.94	44.5	32.52
AU25	Lips part	16.34	14.04	15.38	9.77	9.0	9.45	10.09	9.05	9.65
AU26	Jaw drop	12.05	10.52	11.41	7.45	7.57	7.5	8.14	6.94	7.64
AU28	Lip suck	**0.87**	**1.55**	**1.15**	**0.69**	**0.63**	**0.66**	**0.6**	**0.91**	**0.73**
AU45	Blink	16.24	14.56	15.54	17.86	18.66	18.19	18.52	18.73	18.61

facial expressions. Finally, we identified relevant AUs that co-occur during the different facial stress patterns and the basic emotions.

4.4.1.1 Feature Extraction

We extracted 18 distinct AUs using the OpenFace toolbox [20] (see Table 4.4). The intensity values of AUs range from zero to five, with five being the maximum intensity.

To mitigate variations in facial expressiveness among participants, we applied z-normalization to the AUs within each individual. This step was crucial to address the potential skewing effect of highly expressive individuals, who might otherwise disproportionately influence the formation of facial expression clusters. Our objective was not to delineate clusters solely based on different levels of expressiveness (referring to the intensity and frequency of facial responses). Instead, we aimed to extract distinct facial activity patterns that remain consistent across a group of individuals, irrespective of whether the expressions were subtle or more pronounced. The z-normalization within each participant effectively eliminated individual differences in expressiveness, allowing us to focus on the extraction of consistent facial activity patterns. Kunz et al. [139] also employed z-normalization on AUs to cluster pain facial expressions.

For our frequency occurrence analysis, we used binary features of the AUs (Eq. 4.2)

$$\mathbf{f}_b = (f_{b1} = AU01, f_{b2} = AU02, \cdots, f_{b17} = AU45) \\ \text{with} \begin{cases} f_{bx} = 1, & \text{if } f_{cx} \geq 1 \\ f_{bx} = 0, & \text{otherwise} \end{cases} \tag{4.2}$$

and for the cluster analysis we used the z-normalized features of continuous AUs (Eq. 4.3)

$$\begin{aligned} \mathbf{f}_c = (f_{c1} = AU01, f_{c2} = AU02, \cdots, f_{c17} = AU45) \\ \text{with } f \in [0,5] \end{aligned} \tag{4.3}$$

4.4.1.2 Frequency Occurrence of Facial Action Units

To determine the AUs for inclusion in our cluster analysis, we assessed the frequency of each AU during stress. For comparison, we also computed the frequency occurrences (FOs) during the typing 1 and 2 phases. Only AUs that manifested in at least 5% of the recorded stress segments were chosen for additional analyses. Fourteen of the eighteen AUs occurred more often than 5% and are detailed in Table 4.4. We also computed the statistical significance of the differences between the frequency occurrences during stress and the typing phases. Similar AUs showed a p-value < 0.05 when computing the statistical significance of the AUs FOs during stress and typing 1 and stress and typing 2. In both analyses, AU 2, 4, 5, 6, 14, 15, 17, 20, 25, 26 show significantly different occurrences during stress and the typing phases. In the analysis with typing 1, AU 23 also showed a p-value < 0.05, while in the analysis with typing 2, AU 7 and AU 45 were additionally statistically significant.

4.4.1.3 Initial Cluster Analysis

For our initial cluster analysis, we applied the z-normalization on the AUs with FO above 5%. We randomly chose 100 samples per subject from the stress video segment and applied the K-means algorithm using different cluster numbers (K=1 to 21) on a total of 1500 samples. The clustering results were visually not different from each other, which is why we proceeded with performing clustering on data of each individual separately. We selected 1000 samples per subject randomly and also performed K-means clustering with K varying between 1 and 10. We applied the elbow method [268] to choose the ideal cluster number, which for most subjects was three. We also computed the silhouette score [248] for each clustering of the subjects' samples. Fig. 4.9 shows the clustering results for nine of the 115 subjects and how the silhouette score indicates the visual difference between the clusters. Clusters with low silhouette scores show minimal differences in facial expressions. Clusters with high silhouette scores, on the other hand, show well-distinguishable facial expressions. What is also noticeable, is the variety of facial expressions recognized by performing clustering for each individual. The clusters show unexpressive faces, smiles, pressed lips, open mouths, and asymmetric dimplers.

4.4.1.4 User Study

Given that our initial clusters were automatically created from randomly extracted frames of the stress videos, it is not guaranteed that the obtained clusters show facial expressions that can be recognized as stress. To further refine the data to obtain facial activity patterns

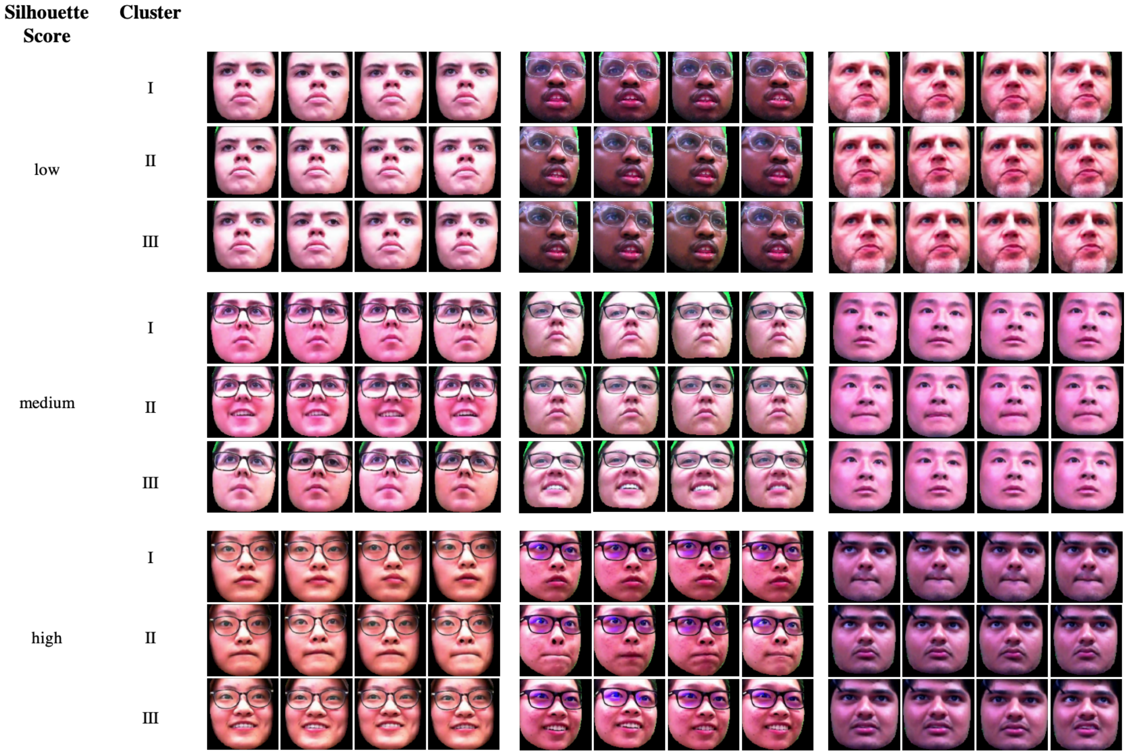


Figure 4.9: Examples of clusters obtained by performing clustering on the data of each individual separately. The figure shows the impact of the silhouette score on the differences shown in the clusters. Clusters with low silhouette score (0 to 0.2) show no visible differences between the AUs. Clusters with medium silhouette score (0.2 to 0.4) show two visually distinct clusters and clusters with high silhouette score (> 0.4) show visual differences among all clusters.

that show stress, we performed a user study. Instead of using all 1000 frames per individual, we show the cluster centroids (frames in the center of the clusters) of each subject. Given that the silhouette score of the clusters indicates the visual difference between them, we chose either one frame from one cluster, two clusters, or each cluster to avoid including frames with similar facial expressions. With this selection criteria, we obtained 208 cluster centroids to be shown in our user study.

A total of 17 people (eight female and nine male) took our study online with ages between 25 and 44. In the introduction of the study, we inform that pictures of different people working on a computer will be shown during this study and that they should indicate whether the person looks stressed or not. During the study, individual frames are shown with the question “Does the person look stressed?”. The participants can answer with “Yes”, “Maybe”, or “No”. The study was split into four batches to allow breaks in between, showing the 208 frames in a randomized order. Answering each batch took an average of 6 minutes.

In crowd-sourced annotation, the presence of noisy annotations is common, especially when non-expert annotators are involved, as spammers and malicious workers may contribute [34]. To address this, we assessed the percentage agreement of each annotator’s annotations by comparing them to a preliminary majority vote. The analysis revealed that 12 annotators had an agreement lower than 60% with the preliminary majority vote. To ensure the highest annotation quality, we selected the remaining 5 annotators. To validate the high inter-rater agreement, we computed Cohen's kappa [175] pairwise for the 5 annotators, obtaining an average agreement score of 0.42, which is considered moderate.

4.4.1.5 Final Clustering

For final clustering, we selected the data samples from the user study that had a "Yes" annotation for more than 60% of the answers. We applied again K-means clustering on the final data selection, however, only two clusters showed visually consistent similarities over the samples. For this reason, we decided to manually group the samples following common similarities. Two authors of this work performed the manual clustering separately and reached the same clusters. For each cluster, we computed the mean of the z-normalized AU intensities. Following the definitions of which AUs are present during the basic emotions [88, 60, 188] we identified the AUs that co-occur during our defined facial activity patterns of stress and the basic emotions.

4.4.2 Results

In the following, we will describe the results obtained through manual clustering of the samples that were identified to show stressed individuals during the user study. Fig. 4.10 provides an overview of samples belonging to the individual clusters, as well as boxplots of the z-normalized AUs occurring in each cluster. Each cluster shows one very characteristic expression which we used to describe the clusters for easier comprehension.

4.4.2.1 Pressed lips - Cluster 1

This cluster had the most samples in our analysis (24.2 %). The characteristic pressed lips show high intensities for AU 23 (lip tightener), followed by AU 14 (dimpler) and AU 20 (lip stretcher). AU 20 is also a characteristic AU during fear and AU 23 during anger. The focused eyes as well as the pressed lips transmit a tensioned person. This facial activity pattern also can be seen frequently in Fig. 4.9 in the row with the highest silhouette score.

4.4.2.2 Biting lip - Cluster 2

In cluster two, individuals are biting their lower lip. In FACS however, there is a dedicated AU for biting the lip (AU 32), the tool we used cannot detect that AU. The AU with the highest intensity is instead AU 23 (lip tightener) followed by AU 17 (chin raiser) and AU 15 (lip corner depressor). The mentioned AUs are also active during anger, disgust, and sadness respectively.

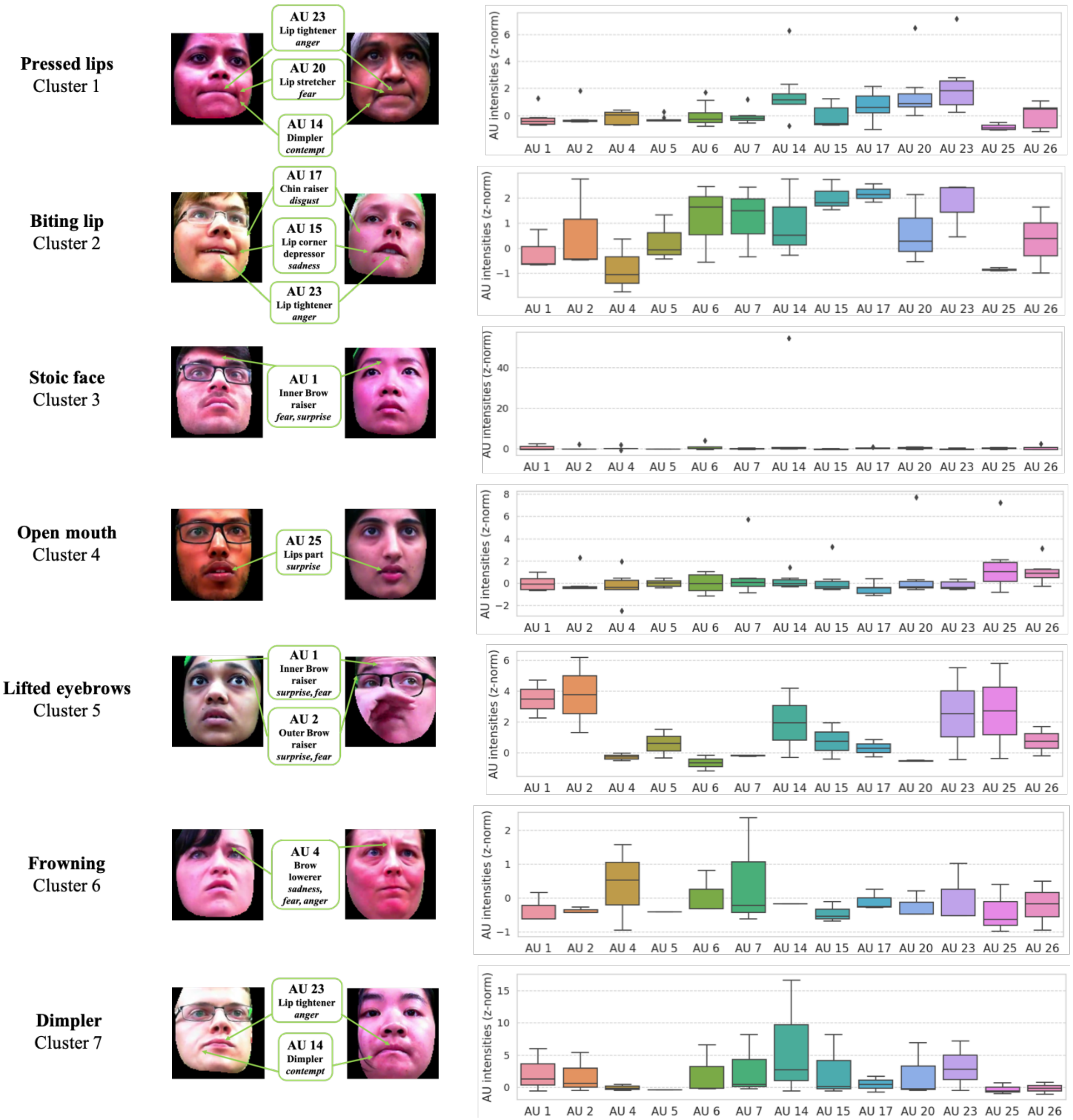


Figure 4.10: Results of manual clustering of samples identified as showing a stressed person. On the right of each sample, the mean of the z-normalized AUs intensities is shown with boxplots. The AUs with the highest means are indicated on the samples of the clusters, together with the basic emotions that show the activity of the same AU.

4.4.2.3 Stoic face - Cluster 3

Cluster three shows a stoic face with mostly low AU intensities. Nevertheless, a slight activity of AU 1 (inner brow raiser) is visible. This facial activity pattern was present in 15% of our final clustering samples. It is also frequent in the individual clustering results in Fig. 4.9. The inner brow raiser is also characteristic during fear and surprise.

4.4.2.4 Open mouth - Cluster 4

In cluster four individuals have all their mouths open. Although similar to cluster three all AU intensities are low, AU 25 (lips part) shows the highest mean. This cluster was the second most frequent among the samples containing 18% of the samples. Parted lips are also characteristic of the basic emotion of surprise.

4.4.2.5 Lifted Eyebrows - Cluster 5

In cluster five samples show strongly lifted eyebrows. The highest mean values are shown in AU 2 (outer brow raiser) and AU 1 (inner brow raiser). Both AUs are characteristic of the basic emotions of fear and surprise.

4.4.2.6 Frowning - Cluster 6

Cluster six contains samples with predominantly frowned eyebrows. AU 4 (brow lowered) has the highest mean compared to the remaining AUs. Eyebrow frowning is also characteristic during sadness, fear, and anger. In this cluster slight variations of the lower face are visible, showing parted or pressed lips.

4.4.2.7 Dimpler - Cluster 7

The final identified cluster shows a predominantly unilateral dimpler expression which has been associated with the basic emotion of contempt. AU 23 (lip tightener) has a similarly high mean value and is also present in anger.

4.4.3 Discussion

Through our proposed semi-automatic methods we distilled seven facial activity patterns of stress from initially almost 600 minutes of video recording from 115 subjects. Our user study ensured that our proposed facial patterns are recognized as stressed individuals. We also were able to show that the characteristic AUs for each pattern also co-occur during the basic emotions of fear, anger, surprise, sadness, and contempt which indicates that human annotators associate stress with negative emotions.

Although previous work trained successful stress classifiers using solely AUs, it remained unclear how stress is expressed through facial expressions. A major challenge is to find relevant frames that show facial expressions of stress, especially as stress-related facial patterns are not universal. Similar to previous work, we can confirm that components of negative emotions co-occur during our seven facial activity patterns of stress [91, 31]. Compared with previous work that focused on statistical analysis of the occurrence of AUs during stress, we can confirm the result from Blasberg et al. [31] that AU 5 (upper lid raiser) occurred more often during stress than the typing phases, however, it is not one of the main characteristics of any of our proposed seven facial activity patterns of stress.

Although we assumed that stress is shown through different facial activity patterns, we were surprised to clearly distinguish seven facial activity patterns during stress. Our initial clustering of facial expressions of individuals during stress had already shown that some of the subjects in the videos show three very distinct facial patterns during stress. Given that per subject, we chose a maximum of three different facial expressions to be shown to human annotators, it was unexpected to find seven instead of three - more than twice as many - clearly distinguishable facial patterns showing stress. Although our initial clustering results showed that several individuals smiled during the stress activity, human annotators did not relate the smiles with stress. We believe that the occurring smiles are a result of nervousness which has been shown in previous research papers [8, 28]. However, given that the user study was designed using single frames instead of video segments, we hypothesize that the lack of temporal context impeded annotators from recognizing nervousness in the smiles.

4.4.4 Conclusion and Future Work

In this work, we propose a novel semi-automatic method that allows to obtain relevant facial patterns of stress from a large data pool with high variability between individuals. We combined clustering, statistical analysis, and human annotation to obtain seven facial activity patterns of stress by analyzing almost 600 min of video material of 115 different individuals during a stressful task. The resulting seven facial expressions of stress show different characteristic facial expressions and emphasize the variety of facial activity patterns during stressful situations. Although we cannot state that these are the only facial activity patterns showing stress, this work provides the first proposal of stress facial patterns in literature that is not solely based on statistical analysis, but that contains human annotation confirming the recognition of stress. We believe that our work will serve researchers from computer science, human-machine interaction, as well as psychology in advancing systems that allow to improve not only stress detection but also the interaction with stressed individuals. In this work, we did not take into account the temporal sequence of AUs within each activity pattern. Addressing the temporal sequence should be considered the next step to comprehensively capture the facial language of stress.

5

Facial Expressions during Enthusiastic Presentations

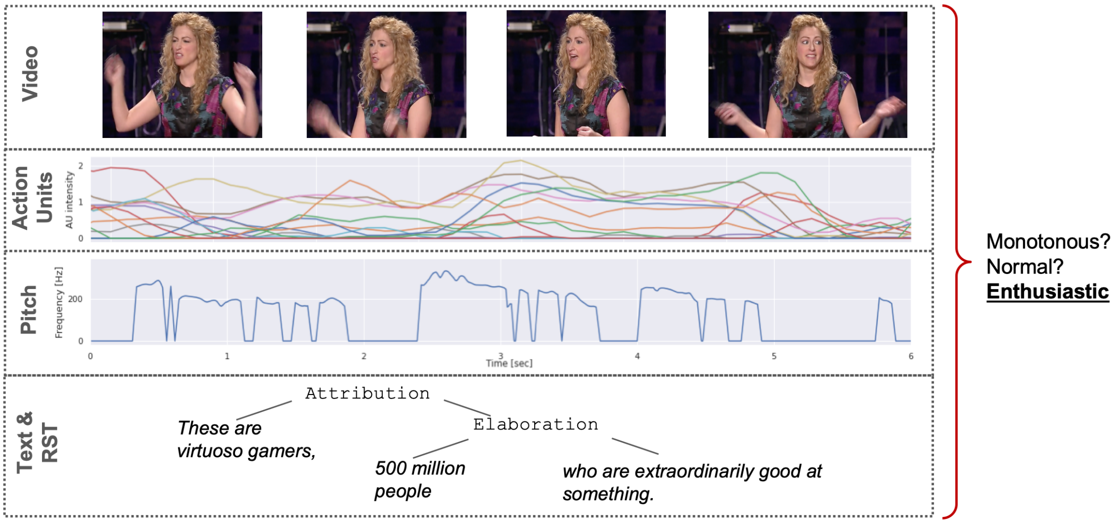


Figure 5.1: An enthusiastic sample from our Entheos dataset with the corresponding variation of facial action units and pitch, as well as the discourse relations from text. Entheos is composed of sentences from different TED talks and is the first dataset that allows studying enthusiasm in video, audio, and text.

5.1 Introduction

The computer science community has learned to detect the six basic emotions in a multimodal manner, however, there is a need to study more fine-grained conversational expressions (see Chapter 2.4). In this chapter, we focus on studying enthusiastic presentation taking into account facial expressions but also speech, text, and discourse relations.

Enthusiasm plays an important role in engaging communication. It enables speakers to be distinguished and remembered, creating an emotional bond that inspires and motivates their addressees to act, listen, and coordinate [30]. Although people can easily identify enthusiasm, this is a rather difficult task for machines due to the lack of resources and models that can help them understand or generate enthusiastic behavior.

Our contributions are as follows: First, we present Entheos (*greek*: being possessed by a god, root for enthusiasm), the first multimodal dataset of TED talk speeches with annotated enthusiasm level[1] (Chapter 5.2). It contains sentence segments, labeled as either monotonous, normal, or enthusiastic. Figure 5.1 shows an example of an enthusiastic sample. Second, in search of finding multimodal signals for understating enthusiasm, we present an analysis of our data to identify attributes present in enthusiastic speech in different modalities (Chapter 5.4). We provide several baseline models using different kinds of features extracted from text, speech, and video. We also show the importance of identifying discourse relations in predicting enthusiasm (Chapter 5.4.2).

5.2 Entheos Dataset

In this section we present our Entheos dataset. We describe our domain choice and label selection, the annotation process, extracted features, as well as statistics of the dataset.

5.2.1 Data Acquisition

Enthusiastic speakers are passionate about their message, wanting to gain their audience for their purpose and persuading them to change their perspective or take action. Given that TED is well-known for spreading powerful messages that can change attitudes and behavior, we use TED talk speeches as our domain for creating a multimodal enthusiasm dataset. We randomly selected 52 male and female speakers from the TEDLIUM corpus release 3 [112], which contains audio of 2351 talks. Transcripts were obtained through the Google cloud transcription service[2]. The talks were segmented into sentences, based on punctuation. We extend the samples from the TEDLIUM corpus with aligned video segments downloaded from the official TED website.

5.2.2 Label Selection and Temporal Granularity

In order to define the temporal granularity for annotation and what labels to use, we performed preliminary annotation experiments with three annotators.

Three audio recordings of talks were chosen from speakers with different proficiency level. One recording was a TED talk by Al Gore[3], and the remaining were recordings of participants in a pilot study with our institution in which they introduce themselves and describe their skills.

We evaluated two different temporal granularities: sentence-level and entire talk. In addition, we explored the use of three different sets of labels, which will be described in the following.

[1] https://github.com/clviegas/Entheos-Dataset

[2] https://cloud.google.com/speech-to-text

[3] https://www.ted.com/talks/al_gore_averting_the_climate_crisis

Table 5.1: Description of the Public Speaking Competence Rubric (PSCR) [241] evaluated as a potential label to describe the use of vocal expressions and paralanguage during a talk.

Rating	Description
4: Advanced	Excellent use of vocal variation, intensity and pacing; vocal expression natural and enthusiastic; avoids fillers
3: Proficient	Good vocal variation and pace; vocal expression suited to assignment; few if any fillers
2: Basic	Demonstrates some vocal variation; enunciates clearly and speaks audibly; generally avoids fillers (e.g. um, uh, like)
1: Minimal	Sometimes uses a voice too soft or articulation too indistinct for listeners to comfortably hear; often uses fillers
0: Deficient	Speaks inaudibly; enunciates poorly; speaks in monotone; poor pacing; distracts listeners with fillers

Table 5.2: Fine-grained description of vocal attributes derived from PSCR, evaluated as potential label categories on sentence-level.

Vocal Attributes	Description	Rating
Variation	Vocal variety is the spice of speech. Tone, pace, and volume should all be varied over the course of a presentation.	4: excellent, 3: good, 2: some, 1: almost no vocal variation, 0: speaks in monotone
Intensity	Speaks loudly and clearly enough for listeners to hear and understand what is being said.	4: excellent use, 3: good, 2: enunciates clearly and speaks audibly , 1: sometimes voice too soft or articulation too indistinct for listeners to comfortably hear, 0: inaudibly, enunciates poorly
Pacing	Speaks in an understandable rate and places pauses for emphasis.	4: excellent use including well-placed pauses, 3: good, 2: pace is appropriate but could have more/fewer pauses, 1: poor pacing, 0: poor pacing with no/too many pauses
Expression	Emotion delivered by the voice.	4: natural and enthusiastic, 3: suited to assignment, 2: some expressions, 1: few expressions, 0: no expressions)

PSCR (Public Speaking Competence Rubric) PSCR [241] was developed to effectively assess students' skills in public speaking. It is composed of eleven skills that are assessed during speaking with a 0-4 scale. We focused on the seventh, which evaluates the effective use of vocal expression and paralanguage to engage the audience. During annotation, annotators had Table 5.1 available for a detailed description of how the speaker articulates for the corresponding rating.

Vocal Attributes Based on the PSCR descriptions we crystallized four main components of the effective use of the voice: vocal variation, intensity, pacing, and expression. Each one was evaluated with a score of 0-4 and described as depicted in Table 5.2.

Table 5.3: Intuitive labels used to evaluate as potential categories to annotate sentence-level samples.

Category	Description	Rating
Enthusiasm	Speaker is passionate, energetic, stimulating, and motivating.	0: monotonous, 1: normal, 2: enthusiastic
Emphasis	One or more words are emphasized by speaking louder or pronouncing them slowly.	0: no emphasis, 1: emphasis existent

Enthusiasm and Emphasis As a final set of labels, we decided to use intuitive categories, namely enthusiasm and emphasis. For enthusiasm, we chose the definition provided by Keller et al. [134] as they study enthusiasm in the context of spoken monologues (similar to our data) while Inaba et al. [117] studied written dialogues. We also asked annotators to label enthusiasm in three levels: monotonous, normal, and enthusiastic. As Table 5.3 shows, annotators were asked to label emphasis as existent or not, depending on whether words were emphasized by speaking louder or pronouncing words slowly.

Experiment Description The experiment was composed of two parts. First, the entire audio recordings were played and the annotators were asked to use only the PSCR annotation scheme, rating each talk with a single score. Afterward, seven sentences of each talk were played with pauses in between to allow annotation using vocal attributes, enthusiasm, and emphasis labels. Each sentence was annotated with six scores. For both parts, the annotators had access to the description of the labels during annotation as shown in Tables 5.1,5.2,5.3. Once all annotators finished labeling a sample, the next one was played.

Results and Conclusion In Table 5.4 the inter-rater agreement for the different annotation schemes is shown in terms of Fleiss' kappa [144]. We can see that PSCR, which rated the entire talk, has the lowest agreement. Vocal variation and pacing have moderate agreement, while vocal intensity, enthusiasm, and emphasis show almost perfect agreement.

Given these results, we annotated audio recordings on a sentence level using enthusiasm and emphasis labels.

5.2.3 Data Annotation Protocol

Our study was approved by our institution's human subject board and annotators were paid $20/h. Seventeen subjects participated in data annotation and signed the consent form before the study. For data annotation, an internal tool was created that enabled annotators to listen to audio samples and annotate them through their web browser at their time of convenience. As labeling availability fluctuated, instead of randomly choosing samples from the entire dataset, we decided to release small batches of data to obtain as many annotations per sample as possible. In a bi-weekly rhythm, small batches of 200 samples were available to annotate in a randomly chosen order for each annotator. As

Table 5.4: Inter-rater agreement using different labels computed with Fleiss' kappa with interpretations based on [144]. Enthusiasm, emphasis, and vocal intensity achieved almost perfect agreement.

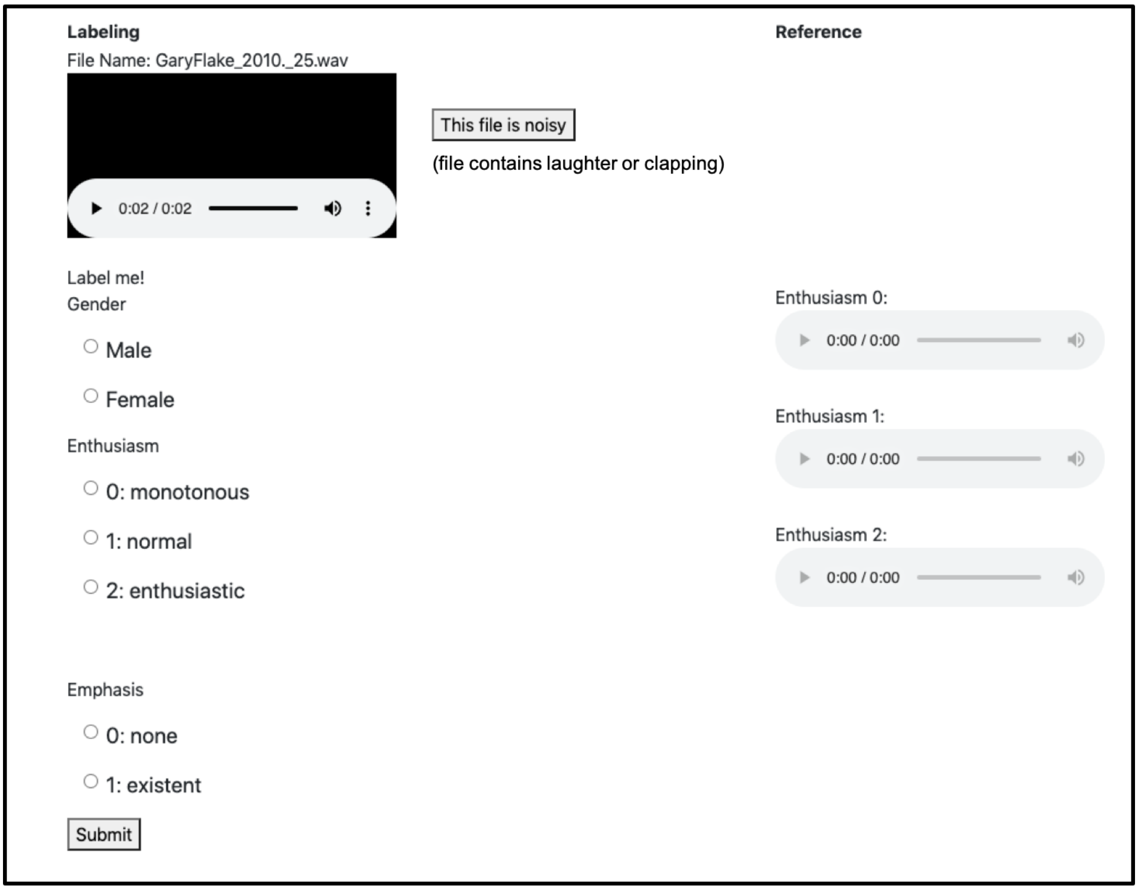
Label	**Fleiss' κ**	**Agreement**
PSCR	0.31	fair
Variation	0.56	moderate
Intensity	0.81	almost perfect
Pacing	0.55	moderate
Expression	0.63	substantial
Enthusiasm	**0.82**	**almost perfect**
Emphasis	**0.87**	**almost perfect**

our definition for enthusiasm (Table 5.3) allows subjective interpretations, we included three reference audio files for each enthusiasm level in the web interface of our annotation tool as depicted in Figure 5.2. Annotators were indicated to listen to the reference files after every 10 labeled samples and when insecure on how to label a sample. In addition, annotators were given the definition of enthusiasm and emphasis shown in Table 5.3. Besides enthusiasm and emphasis, the corresponding perceived gender was annotated. We limited the options for perceived gender to female and male, based on prior work which used these two genders to improve the performance in emotion detection [152]. Samples with laughter or clapping were asked to be labeled as noisy files.

Annotator Quality Assessment: Annotation was performed by 17 different annotators. As noisy annotations are common when crowdsourcing and not using expert annotators due to spammers and malicious workers [34], we compared the percentage agreement of each individual's annotations with a preliminary majority vote. The analysis showed that 12 annotators had lower agreement than 30%. The same annotators had also labeled less than 17% of the data. To ensure high quality of annotation we used the remaining five annotators who labeled more than 50% of the data. The remaining annotators identify themselves as Latino, Asian, and white. We removed all samples that had only one or two different annotations and computed the final majority vote for the remaining 1,126 samples. To confirm the high inter-rater agreement, we computed Cohen's kappa [175] in a pairwise manner for the five annotators and obtained an average agreement of 0.66.

5.2.4 Final Data Selection

Out of 1,819 labeled samples, we kept 1,126 which had more than one annotation. The selected samples are from 113 different TED talk speeches, 60 from male and 53 from female speakers. We created a test split with 108 samples from five speakers of each perceived gender. The training set, composed of 55 male and 48 female speakers, has a total of 1,018 samples. There is no overlap of speakers between the training and test set. In Figure 5.3 (top) we can see the label distribution in our train-test split.



Labeling

File Name: GaryFlake_2010._25.wav

(file contains laughter or clapping)

Reference

Label me!

Gender

○ Male

○ Female

Enthusiasm

○ 0: monotonous

○ 1: normal

○ 2: enthusiastic

Emphasis

○ 0: none

○ 1: existent

Figure 5.2: Layout of the annotation interface. On the top left is the sample to be annotated and below are the different labels: perceived gender, enthusiasm, and emphasis. On the top center is the option to mark the sample as noisy if laughter or clapping is present. On the right side are reference samples for the three different levels of enthusiasm.

5.2.5 Data Statistics

In the following, we will describe the relationship between the different enthusiasm levels and other attributes of the talks such as viewer ratings, number of views and comments, and perceived gender of the speakers. This metadata was obtained from a Kaggle competition[4] that collected data about TED talks until September 21st, 2017.

In Figure 5.3 (center), we can see that the enthusiasm levels are similarly distributed for both gender labels. We computed the Pearson's chi-squared test for independence to evaluate if there is a significant difference in enthusiasm level between genders. With a significance level of 5%, we obtained $p = 0.04$, meaning that the gender of the speaker and enthusiasm level are dependent on each other. In Figure 5.3 (bottom), the label distribution among the different ratings that were given by viewers is shown. There are nine positive ratings (funny, beautiful, ingenious, courageous, informative, fascinating, inspiring, persuasive, jaw-dropping) and five negative ratings (longwinded, confusing, unconvincing, ok, obnoxious) that viewers could select. The ratings have been sorted by

[4] https://www.kaggle.com/rounakbanik/ted-talks

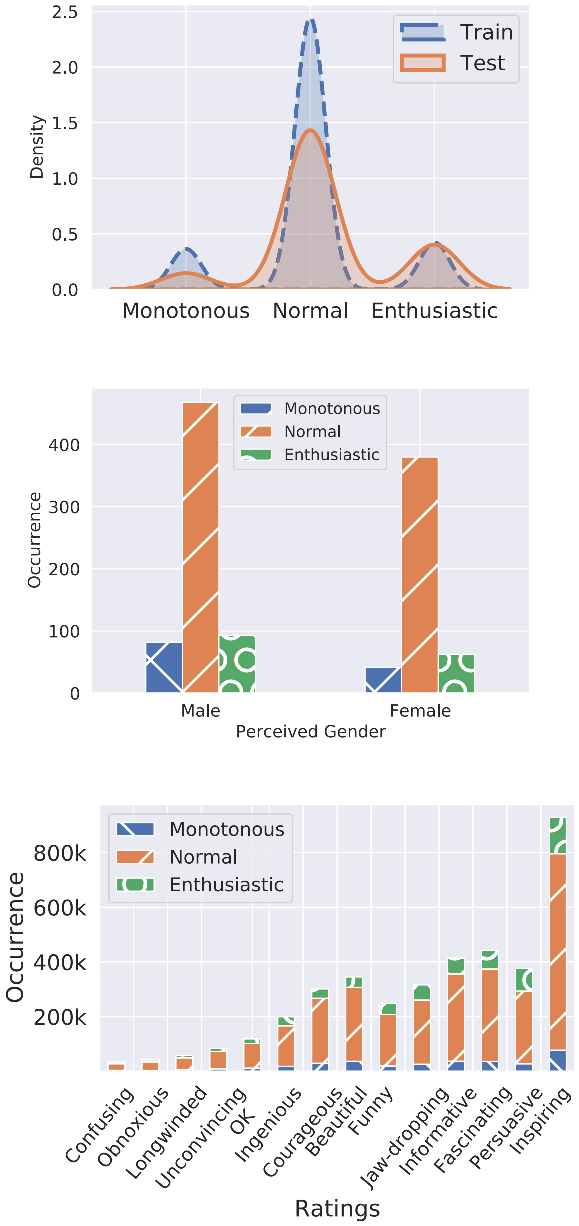


Figure 5.3: From top to bottom: Label distribution in our train-test split, among perceived gender, and ratings given by TED viewers. Top: The training set and testing set reflect the same imbalance of class labels. Center: Female speakers have proportionally fewer monotonous samples and more normal samples than males, but the same proportion of enthusiastic samples. Bottom: Samples labeled as enthusiastic have been mainly rated as fascinating, persuasive, and inspiring. They have rarely been rated negatively.

increasing the number of enthusiastic samples. We can see that the negative ratings have the least number of enthusiastic samples. The ratings with the three highest numbers of enthusiastic samples are fascinating, persuasive, and inspiring. We also performed two one-way ANOVAs to evaluate if the number of views and comments depends on the enthusiasm level. The resulting p-values were correspondingly $p = 0.3844$ and $p = 0.6892$

which means that views and comments are not influenced by the enthusiasm level of the speaker.

5.3 Computational Experiments

In the experiments of this paper, we aim to establish a performance baseline for the Entheos dataset using only the enthusiasm annotations. We train our model with different feature combinations to understand the role of different modalities in enthusiasm detection (see Figure 5.4). In the following, we describe different features that were extracted and the model architecture that we used.

5.3.1 Features

Given the small number of labeled samples, instead of training an end-to-end model, we extract different features that will serve as input for our model. In the following, we will describe the features used per modality.

Video: As enthusiasm is related to emotions, we extracted Facial Action Units (FAUs) which describe the intensity of muscular movements in the face based on the Facial Action Coding System (FACS) [87]. We used OpenFace [21] to obtain the intensity of 18 FAUs on a scale of 0-5. As FAUs vary over time, we computed the average and standard deviation for each AU and concatenated them in a feature of 36 dimensions per sample.

Acoustic: We extracted different audio features using OpenSMILE [79], a toolbox that can extract over 27k features. We extracted four different feature combinations, which have been thoroughly studied in the speech community in affective computing tasks: a) eGEMAPS (88 attributes) [80], b) Interspeech 2009 Emotion Challenge (384 attributes) [242], c) Interspeech 2010 Paralinguistic Challenge (1582 attributes) [243], and d) Interspeech 2013 Compare (6373 attributes) [244]. Each feature collection differs in the selection of features, functionals, and statistical measures. Examples of features covered are voice quality (jitter and shimmer), pitch (F0), energy, spectral, cepstral (MFCC), and voicing-related low-level features (LLDs) as well as a few LLDs including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness.

Text: As a low-level feature, we used the bert-large-uncased model[5] to obtain word embeddings on a sentence level. For each sample, we obtained a feature of 768 dimensions. As high-level features, we extracted two types of discourse relations: Rhetorical Structure Theory (RST) [163] and Penn Discourse Treebank (PDTB) [183, 212] relations. We used the RST parser from Wang et al. [299] and the PDTB parser from Lin et al .[155] for automated discourse relation annotation. Elementary discourse units (EDU) were obtained by using

[5] https://huggingface.co/bert-large-uncased

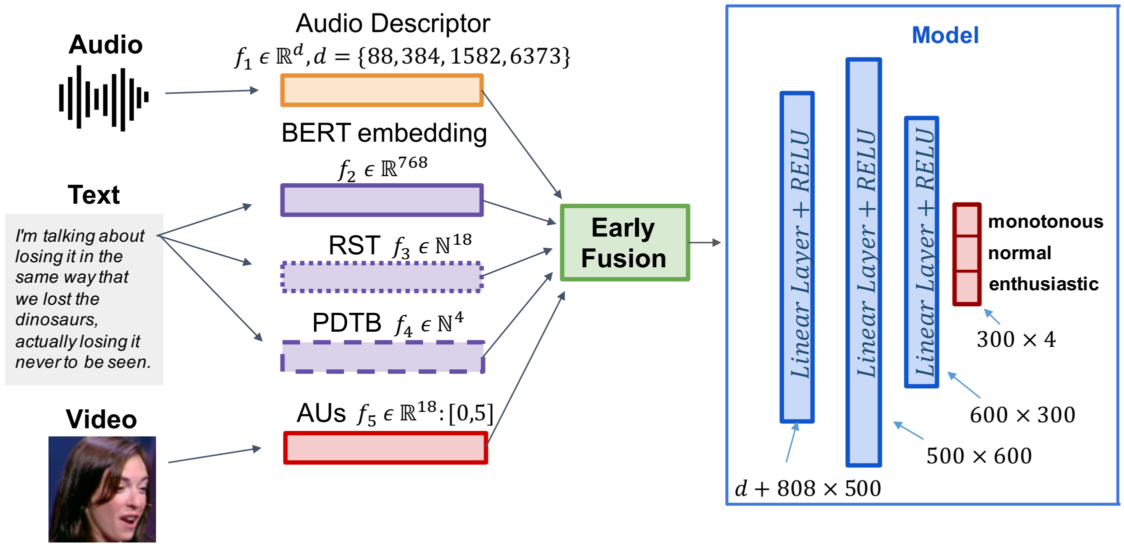

Audio
Audio Descriptor
$f_1 \in \mathbb{R}^d, d = \{88, 384, 1582, 6373\}$
BERT embedding
$f_2 \in \mathbb{R}^{768}$
Text
I'm talking about losing it in the same way that we lost the dinosaurs, actually losing it never to be seen.
RST $f_3 \in \mathbb{N}^{18}$
PDTB $f_4 \in \mathbb{N}^4$
Video
AUs $f_5 \in \mathbb{R}^{18}$: [0,5]
Early Fusion
Model
Linear Layer + RELU
Linear Layer + RELU
Linear Layer + RELU
monotonous
normal
enthusiastic
300 × 4
600 × 300
500 × 600
d + 808 × 500

Figure 5.4: An overview of our proposed multimodal dataset and model for predicting levels of enthusiasm using different features extracted from video, audio, and text.

the method presented by Wang et al. [300]. For both parsers, samples can have more than one relation or none at all. The annotations were converted into a bag-of-words representation, obtaining features of 18 dimensions for RST and 4 for PDTB.

5.3.2 Model Architecture

Our model (Fig. 5.4) is composed of four fully connected layers with ReLU activation functions in between. We use concatenation to combine different features in the multimodal setting. Given our imbalanced dataset, we compute class weights, which represent the relation of samples per label and the total sample number. The class weights are then passed to our loss function (cross-entropy loss) to give more weight to samples of the underrepresented classes. We use the Adam optimizer [135] and during training, we perform early stopping to avoid overfitting. We train the model for a three-class problem using all enthusiasm levels and also in a binary manner, combining "monotonous" and "normal" labels to the category called "non-enthusiastic".

5.4 Results

5.4.1 Predicting Enthusiasm level

For each feature combination, we performed a hyperparameter search with 10-fold cross-validation. The best hyperparameter combination was used to train the model with the entire training set. We evaluated the performance of the models on our test set. In Table 5.5, the weighted average results for precision, recall, and F1-score are shown. We see that in the unimodal case, BERT embeddings perform the best in the binary classification as well as in the three-class problem. Although PDTB has a higher F1 score in the binary

Table 5.5: Weighted average precision, recall, and F1-score for binary (**B**) and multiclass (**M**) classification. The same model architecture was used to train different feature combinations. BERT embeddings performed best in the unimodal setting. Combining acoustic with text features performed best in the multimodal setting.

Features	Precision [B/M]	Recall [B/M]	F1-Score [B/M]
RST	0.67/0.55	0.64/0.47	0.65/0.50
PDTB	0.70/0.68	0.70/0.29	0.70/0.32
BERT	**0.77/0.66**	**0.81/0.56**	**0.75/0.60**
EGEMAPS	0.80/0.59	0.71/0.47	0.74/0.50
IS09	0.70/0.60	0.76/0.57	0.72/0.55
IS10	0.68/0.56	0.70/0.44	0.69/0.48
IS13	0.65/0.68	0.69/0.37	0.67/0.43
AU	0.67/0.77	0.76/0.50	0.70/0.57
BERT + PDTB	0.77/0.66	0.80/0.57	0.77/0.61
BERT + RST	0.79/0.66	0.81/0.56	0.77/0.60
EGEMAPS + BERT	0.81/0.62	0.60/0.58	0.64/0.59
EGEMAPS + PDTB	0.75/0.69	0.75/0.52	0.75/0.54
EGEMAPS + RST	0.77/0.71	0.7/0.61	0.73/0.64
EGEMAPS + BERT + PDTB	0.74/0.72	0.77/0.65	0.75/0.67
EGEMAPS + BERT + RST	0.77/0.71	0.81/0.58	0.75/0.61
EGEMAPS + RST + PDTB + BERT	**0.83/0.63**	**0.84/0.65**	**0.83/0.64**
EGEMAPS + RST + PDTB + BERT + AU	0.81/0.65	0.65/0.58	0.68/0.60

case, RST performs better in the multi-class problem. Out of the different audio features, eGEMAPS performs slightly better than the other acoustic features. In the multi-class case, IS09 features are the best-performing acoustic features.

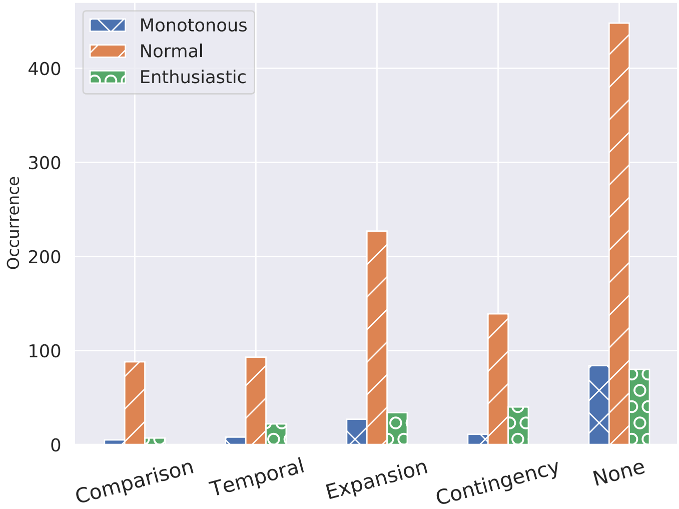
When all features except AUs are combined, we reach the highest F1-score for the binary problem, improving the best unimodal performance by 0.08. We also see that combining both discourse relation features with eGEMAPS and BERT improves the F1-score by 0.08 compared to using only one of them. In the multi-class problem, the best-performing feature combination shows only a slight improvement of 0.04 compared to the unimodal case. Although manually annotating the entire resource was beyond the scope of this paper, we believe that it is necessary to understand the weaknesses and strengths of automatic parsers when used in spoken monologues. With current efforts being made in the field of creating discourse parsers for speech, the role of discourse parsers for enthusiasm detection will be better understood.

5.4.2 Evaluating the Effect of Discourse Features

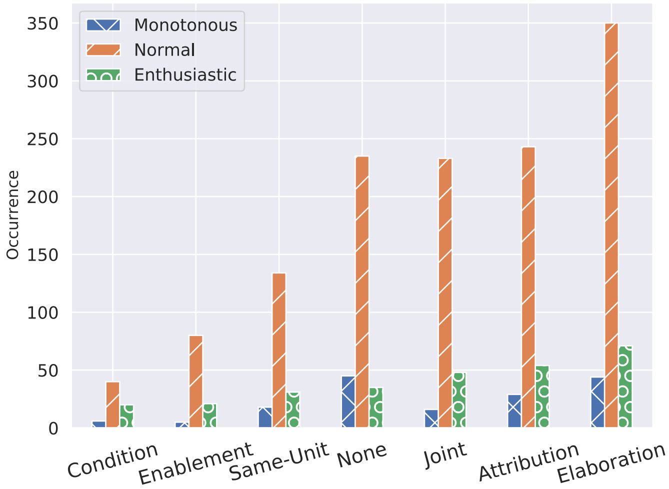
We see in Table 5.5 that discourse relations help the model achieve the highest F1 score. However, we obtained the discourse relations by using discourse parsers that are trained on Wall Street Journal data[6], which is different from monologues.

To evaluate the performance of the parsers, 40 samples of our data were manually annotated with RST and PDTB relations by two annotators. The annotation protocol was approved by our institution's human subject research center. The inter-rater agreement was $\kappa = 0.88$. The accuracy of the RST parser on our data sample was 46.7 and for the

[6] https://catalog.ldc.upenn.edu/LDC93S6A

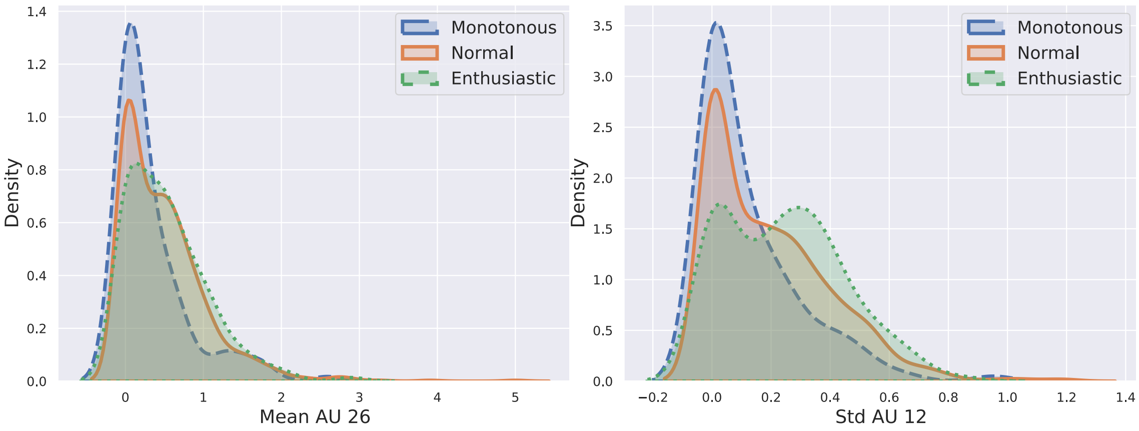


(a)



(b)

Figure 5.5: Label distribution of different enthusiasm levels in relation to discourse relations. In (a), most samples had no PDTB relation, however, there is a visible difference between monotonous and enthusiastic samples in the occurrence of temporal and contingency relations. In (b), RST relations show that enthusiastic samples compared to monotonous samples use more elaboration, attribution, and joint relations.



(a)

(b)

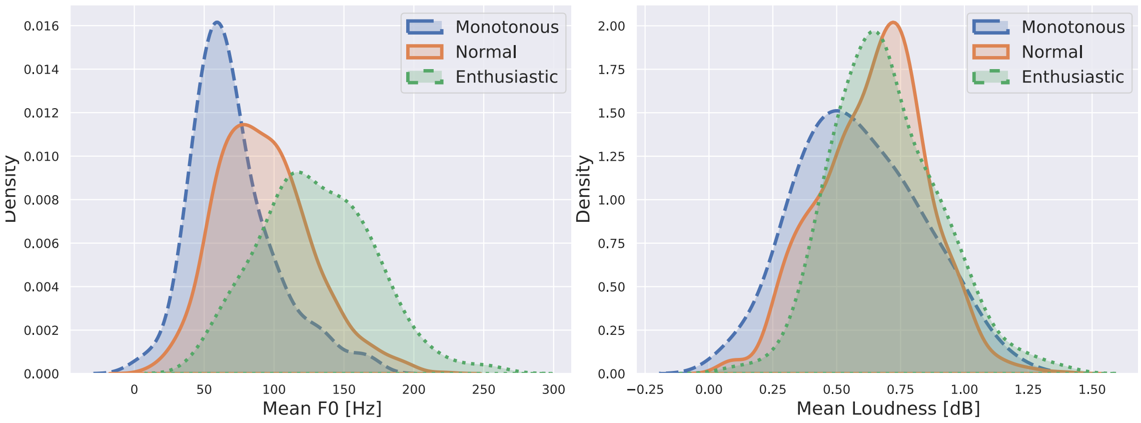
Figure 5.6: Label distribution of different enthusiasm levels in relation to facial action units. In (a), we can see that monotonous samples have more often low intensities for AU26 (jaw drop) than enthusiastic samples. (b) shows that monotonous samples have mostly very low standard deviation for AU12 (lip corner puller), but enthusiastic samples have a higher standard deviation.

PDTB parser 60.0. Although the accuracy of the parsers is low using our data, we have seen that concatenating both discourse relation features to BERT and eGEMAPS improved our model's performance from an F1-score of 0.64 to 0.83 in the binary classification.

In Figure 5.5 we evaluated the relative occurrence of each enthusiasm level for RST and PDTB relations in ascending order of enthusiastic samples. In Figure 5.5a we can see that most samples do not have any discourse relation. However, there is a clear difference in the number of monotonous and enthusiastic samples that show *contingency*, as well as *temporal* relations. In Figure 5.5b we see that enthusiastic samples compared to monotonous samples use more elaboration, attribution, and joint relations. We performed the Pearson Chi-Square test to verify our null hypothesis that discourse relations and enthusiasm levels are independent from each other. We obtained a p-value of 0.0001 for PDTB and a p-value of 0.008 for RST, which permits us to reject our null hypothesis, meaning that the discourse relations influence the level of enthusiasm.

5.4.3 Investigating Visual Features

Given that AUs have not helped our model improve, we evaluated their dependence on our labels. We performed two separate one-way ANOVAs to evaluate the dependence of the mean of the 18 AUs with our labels, as well as the standard deviation of the AUs with our labels. The AUs with p-value < 0.05 are AU 12 (lip corner puller), AU 15 (lip corner depressor), AU 17 (chin raiser), and AU 26 (jaw drop). In Figure 5.6 the label distribution for the mean of AU 26 and standard deviation of AU 12 is shown. In both cases, we can observe that monotonous samples have a mean and standard deviation of zero more frequently than enthusiastic samples. We can also see in Figure 5.6b that enthusiastic samples have more frequently a standard deviation of AU 12 > 0.02.



(a)

(b)

Figure 5.7: Label distribution of different enthusiasm levels in relation to acoustic features. In (a), we can see that enthusiastic samples have a higher mean F0 (pitch) compared to monotonous samples. (b) shows that monotonous speech tends to have lower mean loudness compared to enthusiastic speech.

5.4.4 Investigating Acoustic Features

We have seen that acoustic features are important in improving our model's performance. In this section, we want to evaluate if pitch (F0) and loudness are independent of enthusiasm level. We perform a one-way ANOVA for the mean F0 per sample and its enthusiasm level, as well as for the mean loudness. Both p-values are < 0.05, meaning that the enthusiasm labels depend on the acoustic features. In Figure 5.7a, we can see that monotonous samples have a lower mean F0 than that of enthusiastic samples. We can also see in Figure 5.7b that monotonous samples have lower mean loudness than of enthusiasm. These observations agree with the intuition that enthusiastic speakers speak louder and increase their pitch.

5.4.5 AU Statistical Tests

In order to understand which AU influence the enthusiasm level, we performed two different statistical tests: ANOVA for the three levels of enthusiasm (monotonous, normal, enthusiastic), and T-test for two levels of enthusiasm (enthusiastic, non-enthusiastic). In Table 5.6 on the left we can see the results of the ANOVA, analyzing the mean value of the different AUs per sample with the three levels of enthusiasm. All mean AUs that show p-value < 0.05 are highlighted. As AU 26 has the lowest p-value, the label distribution is shown in Figure 5.6a.

In Table 5.6 on the right we can see the results of the ANOVA, analyzing the standard deviation of the different AUs per sample with the three levels of enthusiasm. As AU 12 has the lowest p-value, the label distribution is shown in Figure 5.6b.

We also performed T-tests for the binary case using the labels enthusiastic and non-enthusiastic. Table 5.7 on the left shows that AU 17 (chin raiser) is the only AU with a

Table 5.6: ANOVA significance test for three levels of enthusiasm and AU mean values on the left and standard deviation of AU on the right. AUs with the lowest p-value are highlighted.

Mean Action Unit	F-Statistic	P-value
AU 01	0.1393	0.87
AU 02	0.4541	0.6351
AU 04	1.415	0.2434
AU 05	0.385	0.6805
AU 06	1.1288	0.3238
AU 07	0.3578	0.6993
AU 09	2.4968	0.0828
AU 10	2.4397	0.0877
AU 12	**4.7553**	**0.0088**
AU 14	1.1253	0.3249
AU 15	**5.1991**	**0.0057**
AU 17	**4.672**	**0.0095**
AU 20	0.8012	0.449
AU 23	1.1192	0.3269
AU 25	0.9896	0.3721
AU 26	**6.0058**	**0.0025**
AU 45	1.0887	0.337

Std Action Unit	F-statistic	P-value
AU 01	**5.614114**	**0.003749**
AU 02	1.052148	0.349531
AU 04	0.905609	0.404591
AU 05	1.778094	0.169435
AU 06	**5.960989**	**0.002660**
AU 07	1.948772	0.142930
AU 09	**10.337395**	**0.000036**
AU 10	**8.383927**	**0.000243**
AU 12	**12.263390**	**0.000005**
AU 14	**5.483568**	**0.004266**
AU 15	**4.347689**	**0.013155**
AU 17	**12.201065**	**0.000006**
AU 20	**3.262334**	**0.038662**
AU 23	**8.411563**	**0.000237**
AU 25	**6.328837**	**0.001848**
AU 26	**11.989375**	**0.000007**
AU 45	**4.585977**	**0.010385**

Table 5.7: T-test for two levels of enthusiasm and AU mean values on the left and standard deviation of AU on the right. AUs with lowest p-value are highlighted.

Mean Action Unit	F-Statistic	P-value
AU 01	-0.3995	0.6896
AU 02	0.0357	0.9715
AU 04	1.5205	0.1287
AU 05	0.4318	0.666
AU 06	-0.0535	0.9573
AU 07	0.7846	0.4328
AU 09	-1.8848	0.0597
AU 10	-0.9503	0.3422
AU 12	-1.1706	0.242
AU 14	0.5841	0.5592
AU 15	-0.9274	0.3539
AU 17	**-2.9922**	**0.0028**
AU 20	-1.2633	0.2067
AU 23	-1.4888	0.1368
AU 25	-0.586	0.558
AU 26	-1.2643	0.2064
AU 45	-0.3449	0.7303

Std Action Unit	F-statistic	P-value
AU 01	**-2.290794**	**0.022160**
AU 02	-1.451265	0.146985
AU 04	-0.909680	0.363186
AU 05	-1.067368	0.286035
AU 06	**-2.203459**	**0.027765**
AU 07	-1.574206	0.115721
AU 09	**-4.374609**	**0.000013**
AU 10	**-3.239400**	**0.001233**
AU 12	**-3.181258**	**0.001507**
AU 14	-1.460543	0.144420
AU 15	**-2.571532**	**0.010253**
AU 17	**-4.600027**	**0.000005**
AU 20	**-2.514758**	**0.012050**
AU 23	**-2.810554**	**0.005031**
AU 25	**-1.972713**	**0.048773**
AU 26	**-2.491479**	**0.012865**
AU 45	-1.378754	0.168245

p-value < 0.05. The distribution of the average values of AU 17 are shown in Figure 5.8(a). For comparison, the distribution of the average AU 02 (outer brow raiser) with highest p-value is shown in Figure 5.8(b). For both analysis, ANOVA and T-test, the differences of standard deviations among the enthusiasm levels are statistically significant for almost all AUs. This is not the case when analyzing the average values of AUs.

Table 5.8: Significance test for mean and standard deviation of F0 and loudness to evaluate the dependence with the different enthusiasm levels. Left: ANOVA significance test results three enthusiasm levels shows that all p-value< 0.05, which means that all variables influence the enthusiasm level. Right: T-test significance test for two levels of enthusiasm also shows that all variables influence the enthusiasm level.

	F-statistic	**P-value**
Mean F0	113.4309	0.0000
Mean Loudness	8.2467	0.0003
Std F0	146.9639	0.0000
Std Loudness	16.9411	0.0000

	F-statistic	**P-value**
Mean F0	-13.1960	0.0000
Mean Loudness	-2.9355	0.0034
Std F0	-13.9376	0.0000
Std Loudness	-4.508	0.0000

5.4.6 Prosody Statistical Tests

We performed statistical significance tests using the mean and standard deviation for F0 (pitch) and loudness. In Table 5.8(left), the ANOVA analysis results are shown and in Table 5.8(right), the results of the T-test. In both significance tests all variables have a p-value< 0.05, which means that all of them influence the enthusiasm level. Figure 5.8(c-f) show the label distribution for different values of the variables used in the significance test.

5.5 Discussion and Conclusion

We present the first multimodal dataset for enthusiasm detection called Entheos[7] and discuss several baseline models. In addition, we present qualitative and quantitative analyses for studying and predicting enthusiasm using the three modalities of text, acoustic, and visual.

Our work has several limitations. TED talks are a very specific form of monologues as they are well-rehearsed and prepared. However, it is more likely that we can find enthusiastic speakers or well-structured sentences in TED talks. To understand enthusiastic behaviors in daily conversations, more data from other domains need to be annotated and studied. We hope that our annotation protocol will help other researchers in the future.

Further theoretical and empirical research is needed for better studying enthusiastic behaviors in general. In our experiments, we used statistical information of each AU instead of the raw signal, which may dilute useful information. Nevertheless, we obtained an F1 score of 0.7 in our baseline model using solely information from facial expressions.

We hope our resources provide opportunities for multidisciplinary research in this area. Given the difficulties of annotating multimodal datasets in this domain, future work needs to investigate weakly supervised approaches for labeling multimodal data.

[7] https://github.com/clviegas/Entheos-Dataset

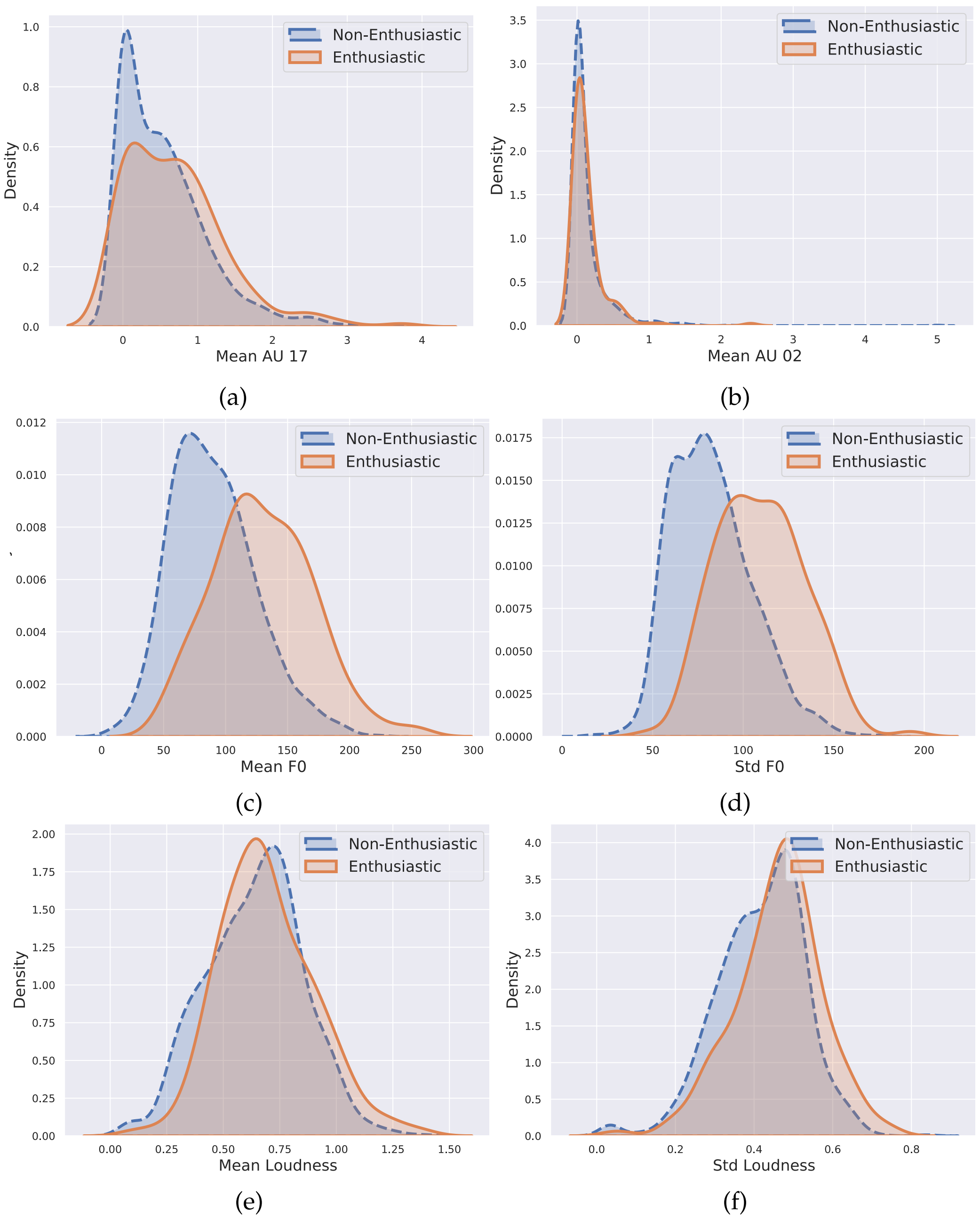


Figure 5.8: Label distribution of enthusiastic and non-enthusiastic samples in relation to the (a) mean AU 17 (p-value = 0.0028), mean AU 17 (p-value = 0.9715), (c) mean F0 (p-value = 0.0), (d) std F0 (p-value = 0.0), (e) mean loudness (p-value = 0.0034), (f) std loudness (p-value = 0.00).

6

Facial Expressions in Sign Language

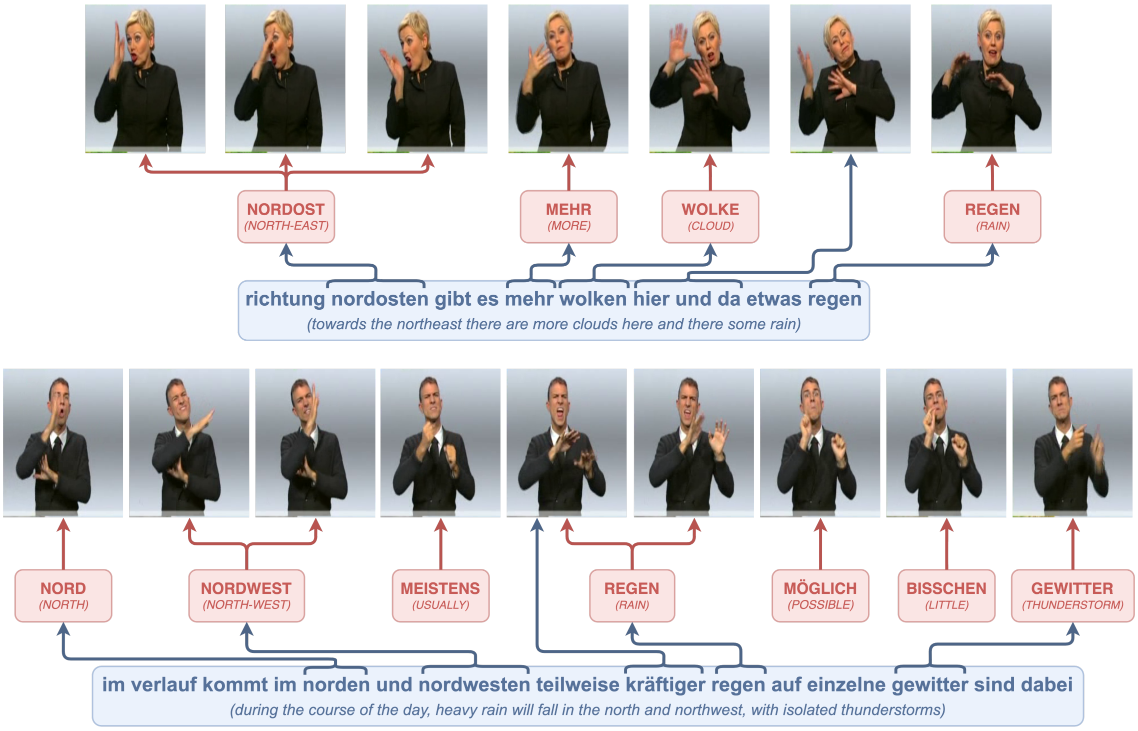

NORDOST
(NORTH-EAST)
MEHR
(MORE)
WOLKE
(CLOUD)
REGEN
(RAIN)
richtung nordosten gibt es mehr wolken hier und da etwas regen
(towards the northeast there are more clouds here and there some rain)
NORD
(NORTH)
NORDWEST
(NORTH-WEST)
MEISTENS
(USUALLY)
REGEN
(RAIN)
MÖGLICH
(POSSIBLE)
BISSCHEN
(LITTLE)
GEWITTER
(THUNDERSTORM)
im verlauf kommt im norden und nordwesten teilweise kräftiger regen auf einzelne gewitter sind dabei
(during the course of the day, heavy rain will fall in the north and northwest, with isolated thunderstorms)

Figure 6.1: Sign Language uses multiple channels, such as hands, body, and facial expressions to convey information. Although gloss annotation is often used to transcribe sign language, the above examples show that meaning encoded through facial expressions is not captured. Both examples from the German Sign Language weather corpus PHOENIX14T show signs for RAIN, however, the lower example shows lowered brows and a wrinkled nose to add the meaning of `kräftiger` *(heavy)* (present in blue text) to the manual signRAIN (shown in red). The example shows that the translation from text (blue) to gloss (red) is lossy even though sign languages have the capability to express the complete meaning of the spoken language.

6.1 Introduction

In Chapter 2.5 we elaborated on the importance of facial expressions in sign language and how they have been overlooked until recently in automatic sign language translation. Although, sign language linguists have established the role of facial expressions as grammatical and lexical components, a mapping of facial expressions to different lexical meanings or an analysis of such into phonological components is still nonexistent.

Despite the fact that previous work has used facial information through CNN features in sign language translation, we propose to use facial features that have an anatomical and functional definition that can be easily interpreted by humans, namely FACS (more details in Chapter 2.1.2). In Chapter 6.2, we present a novel framework that captures information from text and gloss annotation, as well as their relationship to generate continuous 3D sign pose sequences, facial landmarks, and facial action units. We show how including facial information improves sign language generation.

Nevertheless, data is of extreme importance to train more accurate models and unfortunately, large sign language datasets with expensive gloss annotation and non-anonymized faces are not available. For that reason, we started to take a closer a look at the role of facial expressions in sign language. We analyzed the PHOENIX14T datasets on the occurrence of part-of-speech in spoken text transcripts and in gloss annotation and found a significantly lower occurrence of adverbs and adjectives in gloss annotation than in spoken text (example in Fig. 6.1). In Chapter 6.3 we take advantage of this disalignment to develop a semi-automatic method to annotate facial expressions that communicate modifiers. We use FACS to analyze facial expressions with the meaning of "strong" in DGS and "a lot" in LGP and show that there are similarities of the facial expressions in both sign languages.

6.2 Facial Action Units for Sign Language Generation

State-of-the-art sign language generation frameworks lack expressivity and naturalness which is the result of only focusing manual signs, neglecting the affective, grammatical and semantic functions of facial expressions. The purpose of this work is to augment semantic representation of sign language through grounding facial expressions. We study the effect of modeling the relationship between text, gloss, and facial expressions on the performance of the sign generation systems. In particular, we propose a Dual Encoder Transformer able to generate manual signs as well as facial expressions by capturing the similarities and differences found in text and sign gloss annotation. We take into consideration the role of facial muscle activity to express intensities of manual signs by being the first to employ facial action units in sign language generation. We perform a series of experiments showing that our proposed model improves the quality of automatically generated sign language.

In summary, our main contributions are the following:

- Novel Dual Encoder Transformer for SLG which captures information from text and gloss, as well as their relationship to generate continuous 3D sign pose sequences, facial landmarks, and facial action units.
- Use of facial action units to ground semantic representation in sign language.

6.2.1 RWTH PHOENIX14T Sign Language Dataset

In this work we use the publicly available PHOENIX14T dataset [36], frequently used as benchmark dataset for SLR and SLG tasks. The dataset comprises a collection of weather forecast videos in German Sign Language (DGS), segmented into sentences and accompanied by German transcripts from the news anchor and sign-gloss annotations. PHOENIX14T contains videos of 9 different signers with 1066 different sign glosses and 2887 different German words. The video resolution is 210 by 260 pixels per frame and 30 frames per second. The dataset is partitioned into training, validation, and test set with respectively 7,096, 519, and 642 sentences.

6.2.2 Dual Encoder Transformer for Sign Language Generation

In this section, we present our proposed model, the Dual Encoder Transformer for Sign Language Generation. Given the loss of information that occurs when translating from text to gloss annotation described previously, our novel architecture takes into account the information from text and gloss as well as their similarities and differences to generate sign language in the form of skeleton poses and facial landmarks shown in Figure **??**. For that purpose, we learn the conditional probability $p = (Y|X, Z)$ of producing a sequence of signs $Y = (y_1, \ldots, y_T)$ with T frames, given the text of a spoken language sentence $X_T = (x_1, \ldots, x_N)$ with N words and the corresponding glosses $Z = (z_1, \ldots, z_U)$ with U glosses.

Our work is inspired by the Progressive Transformer [234] which allows translation from a symbolic representation (words or glosses) to a continuous domain (joint and face landmark coordinates), by employing positional encoding to permit the processing of inputs with varied lengths. In contrast to the Progressive Transformer which uses one encoder to use either text or glosses to generate skeleton poses, we employ two encoders, one for text and one for glosses, to capture information from both sources, and create a combined representation from the encoder outputs to represent correlations between text and glosses. In the following we will describe the different components of the dual encoder transformer.

6.2.2.1 Embeddings

As our input sources are words, we need to convert them into numerical representations. Similar as in transformers used for text-to-text translations, we use word embeddings based on the vocabulary present in the training set. As we are using two encoders to represent similarities and differences between text and glosses we use one word embedding based on the vocabulary of the text and one using the vocabulary of the glosses. We also experiment by using the text word embedding for both encoders. Given that our target is a sequence of skeleton joint coordinates, facial landmark coordinates, and continuous values of facial AUs with varying length we use counter encoding [234]. The counter c varies between [0,1] with intervals proportional to the sequence length. It allows the generation of frames without an end token. The target joints are then defined as:

$$m_t = [y_t, c_t] \text{ with } y_t = [y_{hands+body}, y_{face}, y_{facialAUs}] \tag{6.1}$$

The target joints m_t are then passed to a continuous embedding which is a linear layer.

6.2.2.2 Dual Encoders

We use two encoders, one for text and one for gloss annotations. Both encoders have the same architecture. They are composed by L layers each with one Multi-head Attention (MHA) and a feed-forward layer. Residual connections [110] around each of the two sublayers with subsequent layer normalization [18]. MHA uses multiple projections of scaled dot-products which permits the model to associate each word of the input with each other. The scaled dot-product attention outputs a vector of values, V, which is weighted by queries, Q, keys, K, and dimensionality, d_k:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}}) \tag{6.2}$$

Different self-attention heads are used in MHA, which allows to generate parallel mappings of the Q, V, and K with different learnt parameters.

The outputs of MHA are then fed into a non-linear feed-forward projection. In our case, where we employ two different encoders, their outputs can be formulated as:

$$\begin{aligned} H_n &= E_{text}(\hat{w}_n, \hat{w}_{1:N}) \\ H_u &= E_{gloss}(\hat{w}_u, \hat{w}_{1:U}) \end{aligned} \tag{6.3}$$

with h_n being the contextual representation of the source sequence, N the number of words, and U the number of glosses in the source sequence.

As we want to not only use the information encoded in text and gloss, but also their relationship, we combine the output of both encoders with a Hadamard multiplication. As the $N \neq U$, we stack h_n vertically for U times and stack h_u vertically for N times in order to have two matrices with the same dimensions. Then we multiply both matrices with the Hadamard multiplication.

$$H_{text,gloss} = \begin{bmatrix} H_{n0} \\ H_{n1} \\ \vdots \\ H_{nU} \end{bmatrix} \odot \begin{bmatrix} H_{u0} \\ H_{u1} \\ \vdots \\ H_{uN} \end{bmatrix} \tag{6.4}$$

6.2.2.3 Decoder

Our decoder is based on the progressive transformer decoder (DPT), an auto-regressive model that produces continuous sequences of sign pose and the previously described counter value [234]. In addition to producing sign poses and facial landmarks, our decoder also produces 17 facial AUs. The counter-concatenated joint embeddings which include manual and facial features (facial landmarks and AUs), $\hat{j}_u$, are used to represent the sign pose of each frame. Firstly, an initial MHA sub-layer is applied to the joint embeddings, similar to the encoder but with an extra masking operation. The masking of future frames is necessary to prevent the model from attending to future time steps. A further MHA mechanism is then used to map the sym- bolic representations from the encoder to the continuous domain of the decoder. A final feed forward sub-layer follows, with each sub-layer followed by a residual connection and layer normalisation as in the encoder. The output of the progressive decoder can be formulated as:

$$[\hat{y}_u, \hat{c}_u] = D(\hat{j}_{1:u-1}, h_{1:T}) \tag{6.5}$$

where $\hat{y}_u$ corresponds to the 3D joint positions, facial landmarks, and AUs, representing the produced sign pose of frame u and $\hat{c}_u$ is the respective counter value. The decoder learns to generate one frame at a time until the predicted counter value, $\hat{c}_u$, reaches 1. The model is trained using the mean squared error (MSE) loss between the predicted sequence, $\hat{y}_{1:U}$, and the ground truth, $y^*_{1:U}$:

$$L_{MSE} = \frac{1}{U}(y^*_{1:U} - \hat{y}_{1:U})^2 \tag{6.6}$$

6.2.3 Computational Experiments

6.2.3.1 Features

We extract three different types of features from the PHOENIX14T dataset: skeleton joint coordinates, facial landmark coordinates, and facial action unit intensities. We use OpenPose [39] to extract skeleton poses from each frame and use for our experiments the coordinates of 50 joints which represent the upper body, arms, and hands, which we will start referring to as “manual features”. We also use OpenFace [21] to extract 68 facial landmarks as well as 17 facial action units (AUs) shown in Figure 2.4 to describe “facial features”.

6.2.3.2 Baseline Models

We will compare the performance of our proposed model (TG2S) with two Progressive Transformers [234], one using gloss only to produce sign poses (G2S), and one that uses text only (T2S). We train each model only with manual features and also with the combination of manual and facial features through concatenation.

Components	**Dev Set**					**Test Set**				
	Bleu$_1$	Bleu$_2$	Bleu$_3$	Bleu$_4$	ROUGE	Bleu$_1$	Bleu$_2$	Bleu$_3$	Bleu$_4$	ROUGE
Manual	30.15	20.58	15.41	12.22	30.41	27.76	18.86	14.11	11.32	27.44
Manual + Facial	29.46	20.30	15.31	12.10	29.25	26.75	17.88	13.29	10.61	26.54

Table 6.1: Translation results of the SLT model [37] used for backtranslation when trained and evaluated with ground truth hand and body skeleton joints (manual) and facial landmarks and AUs (facial).

Model	**Dev Set**					**Test Set**				
	Bleu$_1$	Bleu$_2$	Bleu$_3$	Bleu$_4$	ROUGE	Bleu$_1$	Bleu$_2$	Bleu$_3$	Bleu$_4$	ROUGE
G2S	24.51	15.71	11.19	8.70	**24.84**	**23.26**	14.54	10.21	7.84	22.89
T2S	22.90	14.55	10.42	8.14	23.42	22.14	13.88	9.85	7.56	22.50
TG2S (Ours)	**24.60**	**16.20**	**11.68**	**8.97**	24.82	22.97	**14.71**	**10.59**	**8.19**	**23.45**

Table 6.2: Back translation results obtained from the generative models when using only manual features. Our proposed model has the highest scores in almost all metrics compared to the models using only gloss or text.

Model	**Dev Set**					**Test Set**				
	Bleu$_1$	Bleu$_2$	Bleu$_3$	Bleu$_4$	ROUGE	Bleu$_1$	Bleu$_2$	Bleu$_3$	Bleu$_4$	ROUGE
G2S	16.11	8.77	5.97	4.49	16.19	16.29	9.20	6.37	4.93	16.73
T2S	15.65	8.35	5.76	4.44	15.65	14.12	7.76	5.53	4.39	14.82
TG2S	**17.25**	**10.17**	**7.04**	**5.32**	**17.85**	**17.18**	**10.39**	**7.39**	**5.76**	**17.64**

Table 6.3: Back translation results obtained from the generative models when using manual features and facial landmarks and AUs. Our proposed model has the highest scores in all metrics compared to the models using only gloss or text.

6.2.3.3 Evaluation Methods

In order to automatically evaluate the performance of our model and the baseline models, we use back translation suggested by [234]. For that purpose we use the Sign Language Transformer (SLT) [37] which translates sign poses into text and computes BLEU and ROUGE scores between the translated text and the original text. As the original SLT was designed to receive video frames as input, we modified the architecture to enable the processing of skeleton poses and facial features.

6.2.4 Results

6.2.4.1 Quantitative Results

Table 6.1 shows how well the SLT model performs translation from ground truth sign poses to text when trained and evaluated with the PHOENIX14T dataset. The results show the highest BLEU scores are achieved when training the SLT model only with skeleton joints from hands and upper body, presenting a BLEU-4 score of 11.32 for test set. When facial AUs are added to the hands, body, and face features, the difference to using manual data only is slightly lower, being BLEU-4 of 10.61.

In Table 6.2 the results of using hands and body joint skeleton as sole input to the baseline models and our proposed model are shown. We can see that our proposed model TG2S shows the highest BLEU-4 scores of 8.19 in test set, compared to 7.84 for G2S and 7.56 for T2S.

Table 6.3 presents the results of including facial landmarks as well as facial AUs with body and hands skeleton joints as input. Also here we can see that our proposed model outperforms the baseline models showing BLEU-4 score of 5.76 in test set. G2S obtained BLUE-4 score of 6.37 and T2S 5.53.

We see in Tables 6.2 and 6.3 that G2S obtained higher scores than T2S. Given that gloss annotations fail to encode the richness of meaning in signs, it appears the smaller vocabulary helps the model achieve higher scores by neglecting information otherwise described in text. Our proposed model is able to obtain better results than G2S by makes a compromise of using information from gloss, text, and their similarities and differences. We also can see in both tables that the inclusion of facial information reduces the overall scores. We believe that this might be the case due to the diverse range of facial expressions possible. Also we cannot directly compare the results of Table 6.2 and 6.3 as two different SLT models were used to compute the BLEU scores.

6.2.4.2 Qualitative Results

Figure 6.2 shows the visual quality of our models prediction when using manual and facial information. Both examples show that the predictions captured the hand shape, orientation and movement from ground truth. In the lower example for RAIN, the predictions were even able to capture the repetitive hand movement symbolizing falling rain. What can also be noted is that the ground truth is not perfect. In both examples unnatural finger and head postures can be seen. In addition, ground truth is not displaying movements of the eyebrows and mouth in the expected intensities.

Figure 6.3 shows situations in which the predictions failed to represent the correct phonology of signs. In the first example we see that hand shape, orientation, and position are not correct. The predictions of our models also fail to capture pointing hand shapes as shown in example 2.

6.2.5 Discussion and Conclusion

Sign language generation is a multimodal problem which needs to take into account the phonology of manual signs as well as facial expressions. One of the main challenges is the loss of information that occurs when translating spoken text to gloss, as well from sign language to gloss. Although our proposed model helped bridge the loss of information by taking into account text, gloss, and their similarities and differences, there are still several challenges to be tackled by a multidisciplinary scientific community.

Complex hand shapes with pointing fingers are very challenging to generate. The first step to improve the generation of the fingers is in improving methods to recognize finger

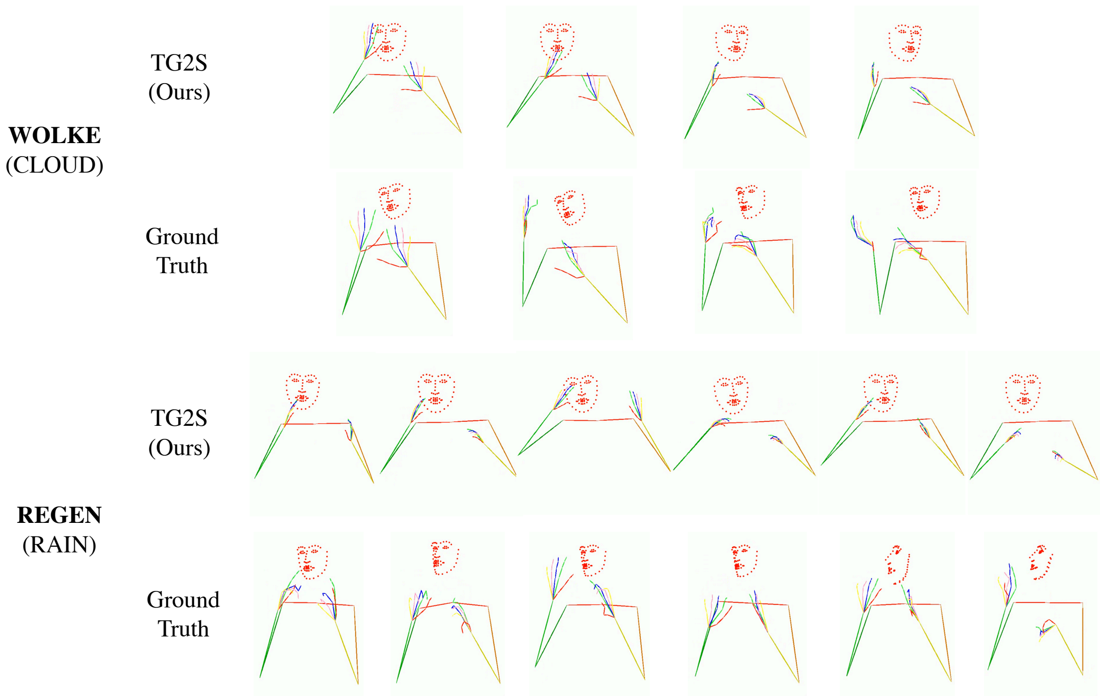


Figure 6.2: Comparison of the ground truth and the generated poses with our proposed dual encoder model for the gloss annotations CLOUD and RAIN. The upper example shows that the predictions captured the correct hand shape, orientation, and movement of the sign CLOUD. In the lower example it is visible that the predictions captured the repeating hand movement meaning RAIN. Although at first glance the hand orientation seems not correct, it is a slight variation which still is correct.

movements more accurately. Similarly, we need tools that are more robust in detecting facial expressions even in situations of occlusion. We also realize that SLG models are overfitting specific sign languages instead of learning a generalized representations of signs. In our case, our model, like many recent models work with the German Sign Language, due to the ease of use, and access to gloss annotations. How2Sign is a feasible dataset for ASL, yet it does not allow any model to extract facial landmarks, facial action units and any facial expression from original video frames as the faces are blurred. Hence we firmly believe that more datasets with better and diverse annotations for different sign languages that allow the design of natural and trustworthy models is necessary. Sign language research is an interdisciplinary field of rich grammar: it has cognitive, linguistic and social aspects. It requires researchers from different areas which also requires machine learning models that are trustworthy and multifaceted. This will allow the output of reasonable usage and application for the community.

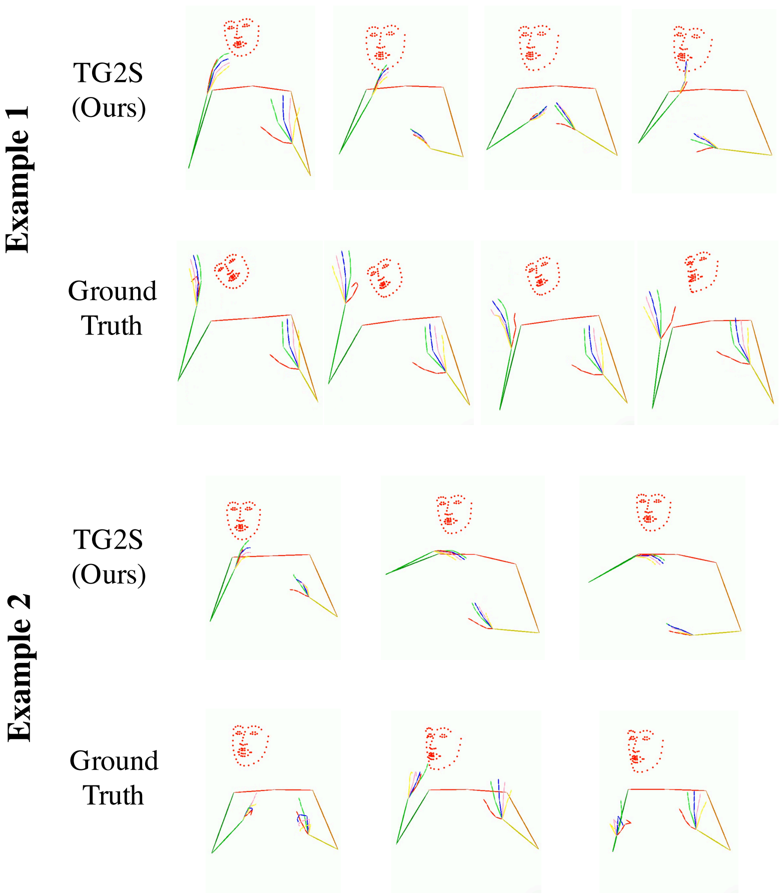


Figure 6.3: Examples in which our model failed to generate the correct phonology of signs. Example 1 depicts inaccuracies in hand shape, orientation, and movement. Example 2 shows the difficulty of the model to capture pointing hand shapes.

6.3 Facial Action Units during Sign Language Modifiers

With our work, we want to propose novel methods that allow computer scientists and linguists to shed light on grammatical facial expressions in sign language. Given that sign language data annotation requires rare expert knowledge and is very time-consuming, we propose a semi-automatic methodology that uses natural language processing (NLP) methods on gloss annotations and transcriptions to automatically detect potential facial expressions with semantic meaning. With our method, we were able to create a subsample of over 200 excerpts from the RWTH-PHOENIX-2014T Weather Forecast dataset in German Sign Language (DGS) [36], containing FEs for the modifiers "light" and "strong". We also introduce a manually annotated dataset in Portuguese Sign Language (LGP) with FEs carrying the meaning "a lot". For a systematic analysis of the linguistic FEs, we detect automatically Facial Action Units (FAUs) [74, 21] and present preliminary definitions of linguistic facial expressions confirmed by sign language linguists. Our contributions can be summarized to the following:

1. Novel semi-automatic method to detect potential FEs with linguistic function.

Table 6.4: Occurrence of different Part-of-Speech (POS) in the sign gloss annotation and the German transcripts computed with Spacy [116]. Although gloss annotations show fewer samples for all POS, the difference in the occurrence of adjectives and adverbs is statistically significant with $p < 0.05$.

	NOUN	VERB	ADV	ADJ
gloss	20927	6407	17718	648
TEXT	25952	7638	24755	5628

2. **DGS Modifier Facial Expression Dataset** containing FEs with the meaning "light" and "strong".

3. **LGP Modifier Facial Expression Dataset** containing FEs meaning "a lot".

4. Preliminary analysis of active Facial Action Units in our datasets.

With our work, we hope to empower other researchers with methodologies to study grammatical FEs in sign languages. Although FEs are often seen as major channels for emotions, they have critical roles in transmitting semantic meaning in sign languages [76, 270, 57, 206]. There is an ocean of words being communicated solely with FEs in sign language and it is time to start fishing them to accelerate the development of dictionaries and grammars as well as to enable faithful automatic sign language translation.

6.3.1 Sign Language Facial Expression Datasets

Unfortunately, at the time of writing, publicly available Sign Language datasets with annotated grammatical FEs do not exist. In this work, we propose a semi-automatic method to accelerate the creation of sign language datasets with grammatical facial expressions from existing SL datasets containing gloss annotations, and text transcriptions. We apply this method on the RWTH-PHOENIX-2014T dataset [36] in DGS which is frequently used as a benchmark dataset for sign language recognition and sign language generation tasks [196, 189, 220]. We also create a subsample of a manually annotated dataset in LGP, focusing on the adverb MUITO (A LOT) that is communicated solely with FEs. In the following, we describe the methods used to create both dataset subsamples, containing video excerpts of grammatical FEs of intensity.

6.3.1.1 Semi-automatically Annotated FEs in DGS

The RWTH-PHOENIX14T dataset comprises a collection of weather forecast videos in German Sign Language (DGS), segmented into sentences and accompanied by German transcripts from the news anchor and sign-gloss annotations. The original dataset contains videos of 9 different signers with 1066 different sign glosses and 2887 different German words.

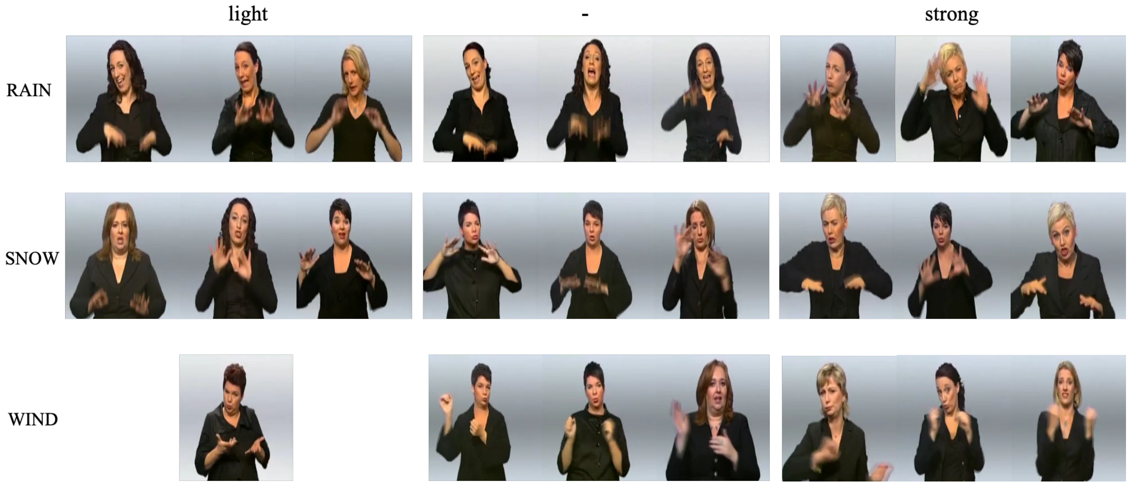


Figure 6.4: Different samples from the DGS Modifier Facial Expression Dataset for different weather nouns and modifiers. Samples for the category "light" show frequently raised eyebrows compared to samples without modifiers (center). In the category "strong" similar eyebrow and mouth movements can be seen to describe strong rain. However, in the other weather noun groups a variety of FEs occurred.

Table 6.5: DGS Modifier Facial Expression Dataset Statistics. The majority of the samples were found for facial expressions indicating "strong" weather. Fewer samples were found for facial expressions indicating "light" weather as signers preferred to use the manual sign LIGHT or LITTLE instead. For that purpose samples without intensity adjectives were also annotated, being shown in the column "-".

		Adjective			
		light	-	strong	Total
Weather	rain	8	31	92	131
	snow	7	3	8	18
	wind	1	3	52	56
	Total	16	37	152	205

Table 6.4 provides an overview of the Part-of-speech (POS) distribution in the original dataset. We used Spacy [116] for automatic detection of nouns, verbs, adverbs, and adjectives in the text transcripts and gloss annotations over the entire dataset. We can see in Table 6.4 that although gloss annotations have lower occurrence for all POS, the difference is statistically significant for adverbs and adjectives with $p < 0.05$. We used this characteristic to search samples that contained adjectives in the text transcriptions but not in the gloss annotations to find signs accompanied by modifying facial expressions. In the following, we will describe in more detail the process of selecting potential samples, the annotation procedure as well as our current preliminary dataset statistics which we call **DGS Modifier Facial Expression Dataset**.

Excerpt Selection Criteria: In a preliminary analysis of the dataset, we observed that modifying facial expressions occur frequently with manual signs describing different weather nouns such as "rain", "sun", "snow", "wind", etc. The most frequently modifying facial expressions have been shown to add meaning to the strength of the weather noun they accompany. For that reason, we decided to annotate samples that have adjectives with the meaning "light" and "strong" only in the text transcription and not in the gloss annotations. We used the following synonyms in German for the meaning "strong": "kräftige", "heftige", "kräftigen", "starken", "starke", "starker", "kräftiges", "heftigen", "harscher", "kräftiger", "′stärkere′. Similarly we used the following synonyms for "light": "milde", "leichten", "leichtem", "leichter".

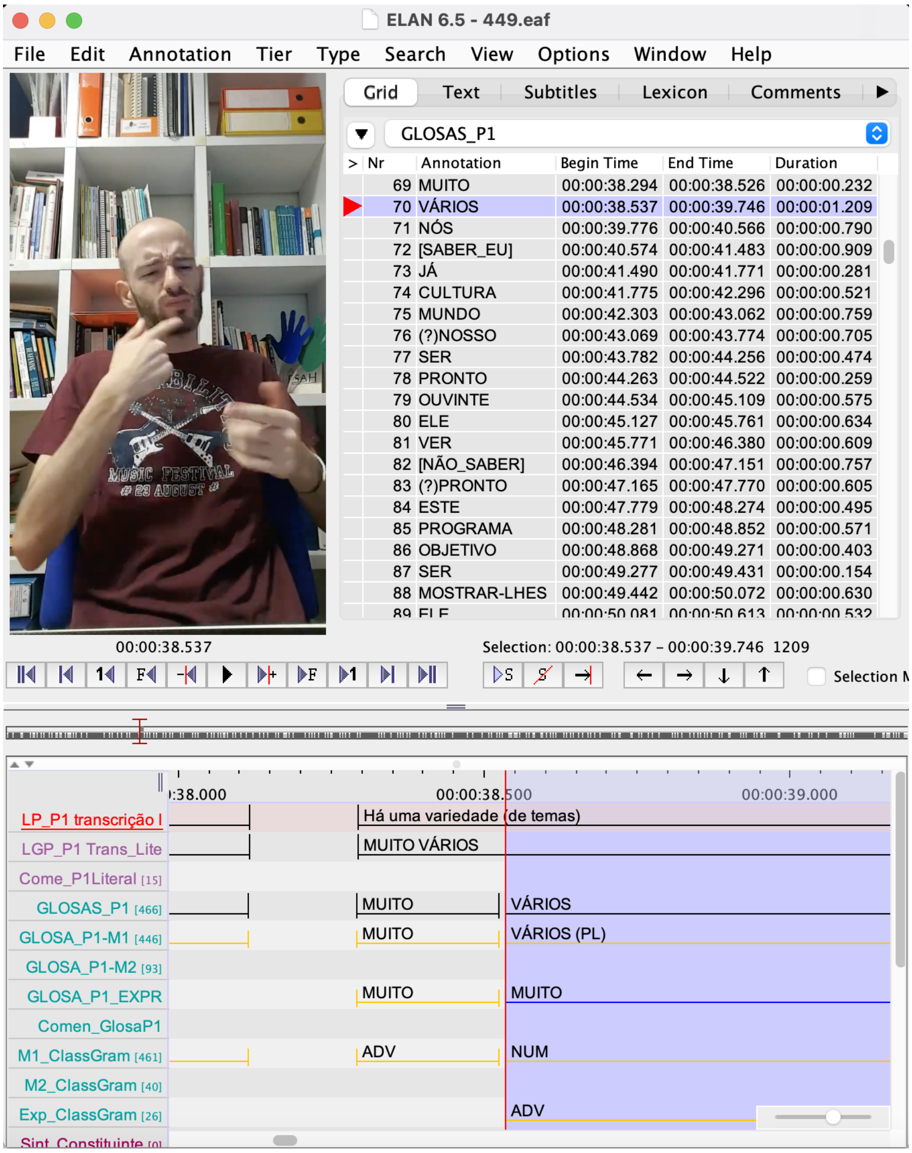
We decided to select samples for the weather nouns "rain", "snow", and "wind" as they have shown to have the most samples with modifying facial expressions for "light" and "strong". Also here we created extensive lists with synonyms such as "Gewitter", "Regen", "Schauer", "Niederschlag" for "rain", "Schneefälle" and "Schneeschauer" for "snow", and "Westwind", "Böen", "Luftmasse" for "wind".

Annotation Procedure: For the selection of excerpts showing the chosen weather nouns with the modifying facial expressions, we chose an annotator who is fluent in German to verify the presence of adjectives with the correct meaning in the text transcript and the lack of them in the gloss annotation. Furthermore, the annotator has basic knowledge of German Sign Language (DGS) and received additional training to recognize the manual signs "rain", "snow", and "wind". During annotation, the original sample and the selected excerpts could be viewed multiple times to ensure that transitioning movements or facial expressions between signs were not included. To distinguish between the modifying FEs and potential FEs of the weather sign, we also selected samples without any adjective describing the intensity of the weather. Fig. 6.4 shows exemplary frames from different weather conditions and modifiers.

Dataset Statistics: Table 6.5 shows our preliminary dataset statistics. We annotated a total number of 205 excerpts, most of which showed the sign REGEN (RAIN). Our extracted excerpts show eight different signers (7 female and 2 male). The number of excerpts per signer varied from a minimum of two to a maximum of 50 excerpts. Although all data splits (train, dev, test) were considered during annotation, we obtained only 16 excerpts for the modifying FE corresponding to "light". We observed that a manual sign for WENIG (LITTLE) was more frequently used with the weather nouns than manual signs such as ENORM (HUGE) or STARK (STRONG). For the following FAU analysis, we decided to proceed with the samples showing the modifier "strong".

6.3.1.2 Manually Annotated FEs in LGP

The FEs from LGP were extracted from an annotated corpus, the Reference Corpus of LGP [41]. This corpus consists of 113 hours of LGP data of video recordings, of which 20 are annotated at different linguistic levels (morphonological, lexical, syntactic, and semantic


ELAN 6.5 - 449.eaf
File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Lexicon Comments
GLOSAS_P1
Selection Mode

>	Nr	Annotation	Begin Time	End Time	Duration
	69	MUITO	00:00:38.294	00:00:38.526	00:00:00.232
▶	70	VÁRIOS	00:00:38.537	00:00:39.746	00:00:01.209
	71	NÓS	00:00:39.776	00:00:40.566	00:00:00.790
	72	[SABER_EU]	00:00:40.574	00:00:41.483	00:00:00.909
	73	JÁ	00:00:41.490	00:00:41.771	00:00:00.281
	74	CULTURA	00:00:41.775	00:00:42.296	00:00:00.521
	75	MUNDO	00:00:42.303	00:00:43.062	00:00:00.759
	76	(?)NOSSO	00:00:43.069	00:00:43.774	00:00:00.705
	77	SER	00:00:43.782	00:00:44.256	00:00:00.474
	78	PRONTO	00:00:44.263	00:00:44.522	00:00:00.259
	79	OUVINTE	00:00:44.534	00:00:45.109	00:00:00.575
	80	ELE	00:00:45.127	00:00:45.761	00:00:00.634
	81	VER	00:00:45.771	00:00:46.380	00:00:00.609
	82	[NÃO_SABER]	00:00:46.394	00:00:47.151	00:00:00.757
	83	(?)PRONTO	00:00:47.165	00:00:47.770	00:00:00.605
	84	ESTE	00:00:47.779	00:00:48.274	00:00:00.495
	85	PROGRAMA	00:00:48.281	00:00:48.852	00:00:00.571
	86	OBJETIVO	00:00:48.868	00:00:49.271	00:00:00.403
	87	SER	00:00:49.277	00:00:49.431	00:00:00.154
	88	MOSTRAR-LHES	00:00:49.442	00:00:50.072	00:00:00.630
	89	ELE	00:00:50.081	00:00:50.613	00:00:00.532

00:00:38.537

Selection: 00:00:38.537 – 00:00:39.746 1209

:38.000 00:00:38.500 00:00:39.000

LP_P1 transcrição I	Há uma variedade (de temas)	
LGP_P1 Trans_Lite	MUITO VÁRIOS	
Come_P1Literal [15]		
GLOSAS_P1 [466]	MUITO	VÁRIOS
GLOSA_P1-M1 [446]	MUITO	VÁRIOS (PL)
GLOSA_P1-M2 [93]		
GLOSA_P1_EXPR	MUITO	MUITO
Comen_GlosaP1		
M1_ClassGram [461]	ADV	NUM
M2_ClassGram [40]		
Exp_ClassGram [26]		ADV
Sint_Constituinte [0]		

Figure 6.5: Annotated sample of simultaneous co-articulation of two grammatical classes (NUM: determiner, ADV: adverb) in the sentence. The hand articulator produces the sign VARIOS (SEVERAL) while the FEs MUITO (A LOT). The figure shows the different tiers for signed annotation: translation into Portuguese (LP_P1 transcricao), literal translation into Portuguese respecting the LGP syntactic order (LGP_P1 Tran_Lite), the written corresponded meaning of the sign (gloss) (GLOSA_P1), gloss produced by one hand articulator (GLOSA_P1-M1), gloss produced by FEs (GLOSA_P1_EXPR), grammatical class of the sign produced by one hand articulator (M1_ClassGram), and the grammatical class of the sign produced by FEs (Exp_ClassGram).

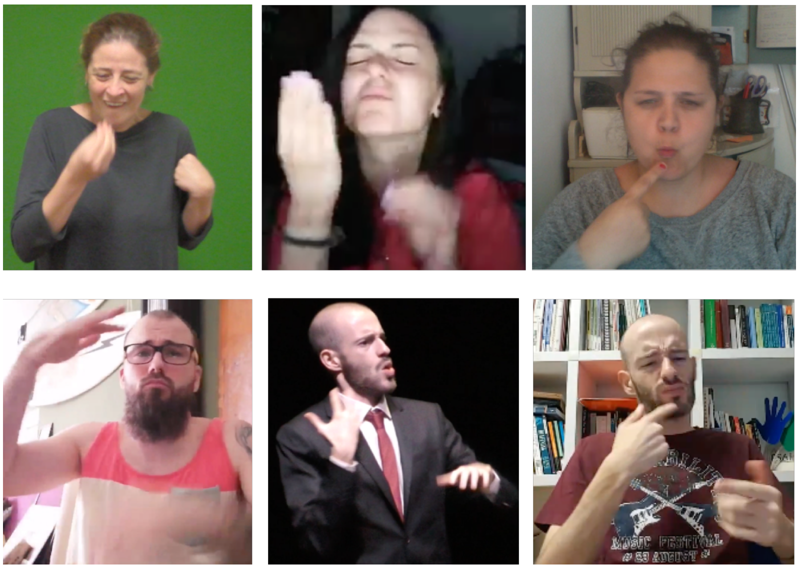


Figure 6.6: Samples from the LGP data showing facial expressions producing the sign MUITO (A LOT).

levels) using the ELAN tool [256]. The videos of the corpus were recorded between 1992 and 2019, showing DHH signers ranging from 4 to 89 years old from different regions in Portugal. We created a subsample of the original LGP corpus, the **LGP Modifier Facial Expression Dataset**, including only excerpts that show FEs with the grammatical function of modifiers. In the following we will describe how the original corpus was annotated and our procedure to create our subsample dataset.

Annotation Procedure: A team of LGP signers annotated the signed data based on glosses and a predefined taxonomy for linguistic phenomena, such as the simultaneous manual and non-manual co-articulation of grammatical categories. The annotation method was: one signer annotator performed the first annotation, a second signer annotator made a revision of the first annotation and, in weekly meetings, the team of annotators resolved the found inconsistencies between the first and second annotator. The identification of the signed grammatical categories was established by the role that the manual and non-manual signs play in the sentence. In Fig. 6.5 an example is shown where a sign with the grammatical class adverb, MUITO (A LOT), is performed by two articulators (one hand: tier *GLOSA_P1-M1*, and FEs: tier *GLOSA_P1-EXPR*). The sign is followed by a non-manual articulator (FEs), MUITO (A LOT), occurring simultaneously during the determiner sign VARIOS (SEVERAL) performed by one hand.

Excerpt Selection Criteria and Dataset Statistics: To find grammatical facial expressions, we searched the corpus for *Exp_ClassGram* annotations. We restricted our search to annotated videos from 2017 to 2019 to obtain excerpts with higher pixel resolution to ensure better quality in the automatic FAU detection. The most frequent grammatical classes with their occurrences in parenthesis were: adjectives (356), nouns (292), transitive verbs (234), verbs (219), adverbs (208), and negation (166). After consulting with LGP linguists, we were recommended to search for modifiers such as "a lot", "big", "a little", "strong", and "intense". We decided to continue working with the most occurring adverb MUITO (A LOT). We obtained 26 samples performed by five different signers (3 female, 2 male). Fig. 6.6 shows some of the detected samples.

6.3.2 Facial Action Unit Analysis

In the following, we will describe the automatic detection of FAUs and the statistical analysis of the modifiers "strong" and "a lot".

6.3.2.1 AU Detection

To analyze different facial expressions, we used the Facial Action Coding System (FACS) [74] which has been used to analyze basic emotions and other expressions such as enthusiasm [289] and stress [294, 293]. We extracted the intensity of 16 different Action Units (AU 1, AU 2, AU 4, AU 6, AU 7, AU 9, AU 10, AU 12, AU 14, AU 15, AU17, AU 20, AU 23, AU 25, AU26) using the tool OpenFace [21]. The automatic FAU detection provides intensity values ranging from zero to five.

6.3.2.2 Statistical Analysis

As a preliminary analysis, we decided to visualize the data for the two modifiers "strong" and "a lot" in two separate boxplots. Boxplots permit us to view how the intensities of the different AUs are distributed in different quantiles, and whether they are normally distributed or have skewed distribution. Additionally, outliers that are numerically distant from the rest of the data are shown. As expressiveness can vary between individuals, we use the PyFeat tool [45] to visualize the active facial muscle group for the different quantiles of the data.

6.3.3 Results

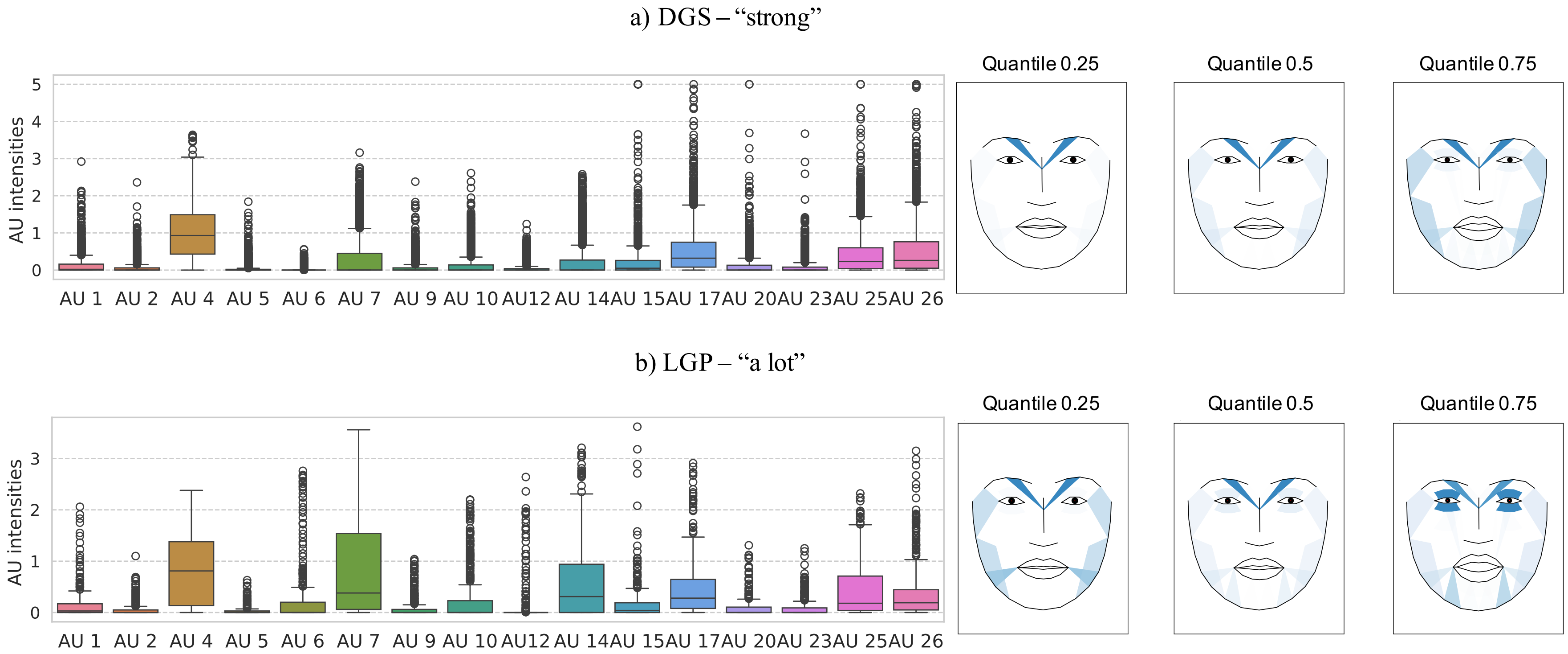
Fig. 6.7 shows the boxplot of the AU intensities over all samples for a) "strong" in DGS and b) "a lot" in LGP. Out of the 16 detected AUs, eight AUs are active in both sign languages. The AU with the highest median values in both, DGS and LGP, is AU 4 (brow lowered). Also, AU 7 (lid tightened) is active in both sign languages. Both AUs are a frequently occurring movement in both data samples (see Fig. 6.4 and 6.6). Both sign languages have active AU 25 (lips apart) and AU 26 (jaw drop) which indicate an open mouth due to mouthing the words "stark" and "muito" in German and Portuguese. In both sign languages AU 10 (upper lip raiser), AU 14 (dimpler), AU 15 (lip corner depressor), and AU 17 (chin raiser) are also active. However, in LGP AU 7, AU 14, and AU 17 show higher intensities than in DGS. In both boxplots, several data points are shown as outliers. However, in LGP the AU 4 and AU 7 do not present outliers. On the right of each boxplot, a visualization of the intensity of the activated AUs is shown for the different quantiles. Interestingly, although similar AUs are active, the AU intensities in LGP are stronger than in DGS.

6.3.4 Discussion

In this work, we present a preliminary FAU analysis of two modifiers expressed solely through FEs: "strong" in DGS and "a lot" in LGP. Although the words used to describe the modifiers are different, the meanings presented are similar in the context of the used samples. In DGS we used samples quantifying a strong intensity of different weather forms, and in LGP we used adverbs intensifying verbs in strength. Interestingly, we were able to find similarities in the active FAUs for both modifiers in DGS and in LGP. Especially in the upper face, lowered brows and tightened eyelids seem to be important to communicate intensity. The remaining active AUs describe the lower face movements which indicate the use of mouthings. The similarities found in the FEs to communicate similar meanings fit with linguistic theories that universalities exist among languages [13, 232, 32]. However, universal phonological components in FEs are still understudied and more data would be necessary to confirm universalities in grammatical facial expressions. Our analysis also shows that though similar AUs are active in both sign languages, the AU intensities are stronger in LGP. We hypothesize that these are due to cultural differences in emotional verbal and non-verbal reactions as described by Fernandez et al. [82]. However, also the use of facial expressions among different sign languages from different cultures is still understudied.

6.3.5 Limitations

In this work, we present two datasets of FEs behaving as modifiers in DGS with the meaning "light" and "strong", and "a lot" in LGP. Given that we are using sign language datasets with samples from weather forecasts, tv, and personal recordings, the amount of relevant samples showing FEs communicating modifiers is sparse. Even though we chose the most frequent modifiers, the data size is still limited. More data showing the modifiers we studied is necessary to perform statistical significance tests with more power. Additionally, more data with different adverbs and adjectives is required to understand how FEs are employed in different manners to communicate semantic meaning. Besides the limitations of samples per modifier, we were limited by the number of signers. We analyzed data of nine signers in DGS (7 female, 2 male) and five signers in LGP (3 female, 2 male). More data from different signers would be needed to understand for example if there are differences due to the sex or age of the signer. Although we found similarities in the FEs used in DGS and LGP to communicate similar meanings (strong rain can also be interpreted as a lot of rain) we can only hypothesize that certain words seem to contain universalities in the use of FE in sign language and that cultural differences could influence



a) DGS – “strong”

b) LGP – “a lot”

Figure 6.7: Boxplot analysis of all samples for a) "strong" in DGS and b) "a lot" in LGP. The AU with the highest median values in both sign languages is AU 4 (brow lowered). Both sign languages have active AU 25 (lips apart) and AU 26 (jaw drop) which indicate an open mouth. In both sign languages AU 7 (lid tightened), AU 10 (upper lip raiser), AU 14 (dimpler), AU 15 (lip corner depressor), and AU 17 (chin raiser) are also active. However, in LGP AU 7, AU 14, and AU 17 show higher intensities than in DGS. On the right of each boxplot, a visualization of the intensity of the activated AUs is shown for the different quantiles.

the intensity of active AUs. Given the limited amount of data, we cannot claim to have found the final definition of active AUs for the modifiers we studied.

6.3.6 Future Work

Our presented results are only the tip of the iceberg of knowledge of linguistic facial expressions in sign language. In this work, we focused only on studying two modifiers in two different sign languages. For future work, we intend to extend our datasets with more adjectives and adverbs. Especially, adverbs of quantity such as "enough", "too much", "a little", "a lot", and "very" are frequently described through FEs. Modifications of our semi-automatic method could also be used to detect grammatical facial expressions of sentences by detecting different types of questions and clauses in the transcripts of spoken language. We also want to increase the sample size of the modifiers we studied to be able to make stronger claims. Other sign languages could also be included to understand if certain semantics have a universal behavior in facial expressions, similar to the basic emotions defined by Ekman et al.[74].

6.3.7 Conclusion

Although facial expressions have essential linguistic functions in sign language, they are still understudied due to the lack of annotated data. With this work, we propose a semi-automatic method to accelerate the annotation of grammatical facial expressions in sign language. We present the DGS Modifier Facial Expression Dataset created with our proposed method as well as the LGP Modifier Facial Expression Dataset based on a manually annotated corpus. We also presented preliminary results using facial action units to quantify and compare the facial expressions of "strong" in DGS and "a lot" in LGP. We found similarities pointing out to a universal character for certain words expressed with FEs but also differences in intensity which can be due to cultural differences.

7

Conclusion

Facial expressions are an essential component of our daily lives. Facial expressions illustrate speech, regulate conversations, serve as emblematic gestures, reveal cognitive processes, signal and regulate emotions, and also serve us during talking and eating [170]. Nevertheless, most of the facial expression datasets publicly available study the basic emotions defined by Ekman et al. [71]. To extend the existing knowledge of facial expressions the research aim of this thesis is to deepen the understanding of unexplored facial expressions. In this thesis in particular, we chose to study facial expressions during cognitive stress, enthusiasm, and sign language.

Our first objective was to evaluate whether it is possible to distinguish stress, enthusiasm, and linguistic facial expressions in sign language by using only the Facial Action Coding System (FACS) as features with different computational algorithms. We were the first to show that stress can be detected with machine learning algorithms using solely FACS features. In addition, our results suggested that stress is not expressed with universal facial expressions which supports outcomes from previous and later publications [150, 59, 31]. We also were the first to study enthusiasm using a multimodal model which was trained on a dataset created by us. Our baseline model solely trained on statistical information from AUs classified enthusiastic samples from non-enthusiastic with an F1 score of 0.7 and our best-preforming multimodal model achieved an F1 score of 0.83, suggesting that enthusiasm is an expression that requires multimodal context to be recognized. In our research on facial expressions in sign language, we trained a sign language generation model that takes into consideration AUs together with manual hand poses. Although the quantitative results in terms of BLEU scores were lower, phrases in sign language exist (although scarce in our dataset) where facial expressions play an essential role in transferring meaning.

Our second objective was to evaluate if characteristic facial expressions exist during stress, enthusiasm, and linguistic facial expressions using FACS. We proposed a novel semi-automatic method that allows to obtain relevant facial patterns of stress from a large data pool with high variability between individuals. By combining clustering, statistical analysis, and human annotation we were able to obtain seven facial activity patterns of stress. To understand facial expressions during enthusiasm, we performed statistical analysis of the action units during enthusiastic and non-enthusiastic samples. AU26 (jaw drop) and AU12 (lip corner puller) have been shown to be relevant in detecting enthusiasm. We also analyzed in more detail modifiers in German and Portuguese Sign language, proposing a semi-automatic annotation method to find facial expressions with grammatical roles. In our FACS analysis, we were able to show how modifiers with similar meanings ("strong" and "a lot") have a similar facial activity pattern.

Our last objective was to create facial expression databases for stress, enthusiasm, and linguistic facial expressions with samples from the studied datasets, by performing user studies. Starting from the stress dataset created by Lau et al. [145] we created a subsample dataset, using our proposed semi-automatic method, with static frames that have been recognized to show stress through a user study. Our multimodal enthusiasm dataset (Entheos) resulted from the manual annotation of TED talk sentences. We also created the German Sign Language Modifier Facial Expression Dataset with our semi-automatic annotation method. All of the mentioned datasets are the first of their kind and we hope that they will support other researchers in building better recognition and generation models.

Although I was able to obtain novel insights into how facial expressions are active during stress, enthusiasm, and sign language, there is still much to be explored. As Paul Ekman said [70] "Every student who examines expression itself, not its recognition, must be impressed with individual differences in the speed, magnitude, and duration of expression as well as variations in which facial expression of emotion occurs in response to a particular event." I believe that this phrase can be transferred to facial expressions of conversational expressions as well as to facial expressions with grammatical role in sign language. With this work, I hope to contribute valuable knowledge and resources to the community of researchers in different fields such as perceptual and cognitive sciences, affective computing, linguistics, as well as computer vision. I believe that our research on facial expressions during stress, enthusiasm, and sign language will enable researchers in academia and industry to create more engaging human-computer interactions through extended facial expression recognizers as well as more expressive virtual agents.

Bibliography

[1] August 2023. URL: https://wfdeaf.org/news/the-legal-recognition-of-national-sign-languages/ (cit. on p. 41).

[2] August 2023. URL: https://manchester.unh.edu/blog/2023/04/spotlight-aslenglish-interpreting (cit. on p. 41).

[3] February 2024. URL: https://imotions.com/applications/consumer-insights/?utm_source=google&utm_medium=cpc&utm_campaign=Human_Behavior_Software&utm_content=Consumer_Insights&utm_term=imotions&gad_source=1&gclid=CjwKCAjw1K-zBhBIEiwAWeCOFz7OggyS0mbDffCfiTC3drF5pUPcZlXbmXUwkOtbiU1rdoZCMh9TkBoCMfIQAvD_BwE (cit. on p. 3).

[4] A. H. A. N. *Chronic stress can cause heart trouble*. www.heart.org, 2020-02. URL: https://www.heart.org/en/news/2020/02/04/chronic-stress-can-cause-heart-trouble#:~:text=Heart%20disease%20is%20another%20potential (visited on 2023-10-20) (cit. on p. 1).

[5] *About Face: Human Expression on Paper*. URL: https://www.metmuseum.org/exhibitions/listings/2015/about-face (cit. on pp. 2, 15).

[6] S. C. Agrawal, A. S. Jalal, and R. K. Tripathi. “A survey on manual and non-manual sign language recognition for isolated and continuous sign”. In: *International Journal of Applied Pattern Recognition* 3.2 (2016), pp. 99–134 (cit. on p. 41).

[7] F. de Almeida Freitas et al. “Grammatical facial expressions recognition with machine learning”. In: *The Twenty-seventh international flairs conference*. 2014 (cit. on p. 42).

[8] Z. Ambadar, J. F. Cohn, and L. I. Reed. “All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous”. In: *Journal of nonverbal behavior* 33 (2009), pp. 17–34 (cit. on p. 66).

[9] E. André et al. “Integrating models of personality and emotions into lifelike characters”. In: *Lecture notes in computer science* 1814 (2000), pp. 150–165 (cit. on p. 1).

[10] J. Antonakis et al. “Just words? Just speeches? On the economic value of charismatic leadership”. In: *NBER Rep. 4* (2019) (cit. on pp. 8, 35, 37).

[11] R. S. Arkushin, A. Moryossef, and O. Fried. “Ham2pose: Animating sign language notation into pose sequences”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 21046–21056 (cit. on p. 43).

[12] C. Armon-Jones. “The social functions of emotion”. In: *The social construction of emotions* (1986), pp. 57–82 (cit. on p. 24).

[13] M. Aronoff et al. “Morphological universals and the sign language type”. In: *Yearbook of morphology 2004*. Springer, 2004, pp. 19–39 (cit. on p. 98).

[14] P. T. Assistivas. *ProDeaf Translator (Version 3.6)*. 2023-10. URL: https://prodeaf-libras-translator.soft112.com/ (visited on 2023-10-21) (cit. on p. 4).

[15] A. P. Association. *Understanding Chronic Stress*. URL: https://www.apa.org/helpcenter/understanding-chronic-stress.aspx (visited on 2018-05-16) (cit. on p. 8).

[16] J. R. Averill. *Anger and aggression: An essay on emotion*. Springer Science & Business Media, 2012 (cit. on p. 24).

[17] J. R. Averill. “Illusions of anger.” In: (1993) (cit. on p. 24).

[18] J. L. Ba, J. R. Kiros, and G. E. Hinton. “Layer normalization”. In: *arXiv preprint arXiv:1607.06450* (2016) (cit. on p. 86).

[19] A. Baker et al. *The linguistics of sign languages: An introduction*. John Benjamins Publishing Company, 2016 (cit. on p. 8).

[20] T. Baltrušaitis, P. Robinson, and L.-P. Morency. “Openface: an open source facial behavior analysis toolkit”. In: *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE. 2016, pp. 1–10 (cit. on pp. 33, 52, 60).

[21] T. Baltrusaitis et al. “Openface 2.0: Facial behavior analysis toolkit”. In: *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE. 2018, pp. 59–66 (cit. on pp. 74, 87, 91, 97).

[22] L. F. Barrett. “The theory of constructed emotion: an active inference account of interoception and categorization”. In: *Social cognitive and affective neuroscience* 12.1 (2017), pp. 1–23 (cit. on pp. 26, 27).

[23] L. F. Barrett and A. B. Satpute. “Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain”. In: *Current opinion in neurobiology* 23.3 (2013), pp. 361–372 (cit. on p. 26).

[24] J. B. Bavelas and N. Chovil. “Nonverbal and Verbal Communication: Hand Gestures and Facial Displays as Part of Language Use in Face-to-face Dialogue.” In: (2006) (cit. on pp. 3, 35).

[25] C. Becker-Asano and I. Wachsmuth. “Affective computing with primary and secondary emotions in a virtual human”. In: *Autonomous Agents and Multi-Agent Systems* 20 (2010), pp. 32–49 (cit. on p. 1).

[26] A. G. Bell. “Fallacies concerning the deaf”. In: *American Annals of the Deaf and Dumb* 29.1 (1884), pp. 32–69 (cit. on p. 39).

[27] C. Bell. *The anatomy and philosophy of expression: as connected with the fine arts*. George Bell and Sons, York Street, Covent Garden, 1877 (cit. on pp. 2, 15).

[28] J. Belluso. *A Nervous Smile*. Dramatists Play Service Inc, 2006 (cit. on p. 66).

[29] C. F. Benitez-Quiroz et al. “Discriminant features and temporal structure of nonmanuals in American Sign Language”. In: *PloS one* 9.2 (2014), e86268 (cit. on pp. 41–43).

[30] E. M. Bettencourt et al. “Effects of teacher enthusiasm training on student on-task behavior and achievement”. In: *American educational research journal* 20.3 (1983), pp. 435–450 (cit. on pp. 1, 8, 35, 67).

[31] J. U. Blasberg et al. “You look stressed: A pilot study on facial action unit activity in the context of psychosocial stress”. In: *Comprehensive Psychoneuroendocrinology* (2023), p. 100187 (cit. on pp. 33, 34, 58, 66, 101).

[32] D. Brentari. “Sign language phonology: Issues of iconicity and universality”. In: *Empirical approaches to language typology* 36 (2007), p. 59 (cit. on p. 98).

[33] J. Bruner. “The perception of people”. In: *Handbook of social psychology* (1954) (cit. on pp. 3, 21).

[34] A. Burmania, S. Parthasarathy, and C. Busso. “Increasing the reliability of crowdsourcing evaluations using online quality assessment”. In: *IEEE Transactions on Affective Computing* 7.4 (2015), pp. 374–388 (cit. on pp. 62, 71).

[35] N. C. Camgöz et al. “BosphorusSign: a Turkish sign language recognition corpus in health and finance domains”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 1383–1388 (cit. on p. 43).

[36] N. C. Camgoz et al. “Neural Sign Language Translation”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018-06, pp. 7784–7793. DOI: 10.1109/CVPR.2018.00812 (cit. on pp. 85, 91, 92).

[37] N. C. Camgoz et al. “Sign language transformers: Joint end-to-end sign language recognition and translation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10023–10033 (cit. on p. 88).

[38] W. B. Cannon. “The James-Lange theory of emotions: a critical examination and an alternative theory”. In: *The American journal of psychology* 100.3/4 (1987), pp. 567–586 (cit. on p. 23).

[39] Z. Cao et al. “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019) (cit. on p. 87).

[40] D. Carneiro et al. “New methods for stress assessment and monitoring at the workplace”. In: *IEEE Transactions on Affective Computing* (2017) (cit. on p. 33).

[41] P. V. d. Carvalho et al. “From sign elicitation to the construction of a Linguistic Corpus: Designing a Reference Corpus of Portuguese Sign Language”. In: *International Journal of Corpus Linguistics* (2024) (cit. on p. 94).

[42] S. Castillo, K. Legde, and D. W. Cunningham. “The semantic space for motion-captured facial expressions”. In: *Computer Animation and Virtual Worlds* 29.3-4 (2018), e1823 (cit. on pp. 3, 4, 36).

[43] Y. Chen et al. “A simple multi-modality transfer learning baseline for sign language translation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5120–5130 (cit. on p. 44).

[44] Z. Chen, R. Ansari, and D. Wilkie. “Automated pain detection from facial expressions using facs: A review”. In: *arXiv preprint arXiv:1811.07988* (2018) (cit. on pp. 5, 46).

[45] J. H. Cheong et al. “Py-Feat: Python Facial Expression Analysis Toolbox”. In: *CoRR* abs/2104.03509 (2021). arXiv: 2104.03509. URL: https://arxiv.org/abs/2104.03509 (cit. on p. 97).

[46] S. Clarke, L. G. Jaimes, and M. A. Labrador. “mStress: A mobile recommender system for just-in-time interventions for stress”. In: *Consumer Communications & Networking Conference (CCNC), 2017 14th IEEE Annual*. IEEE. 2017, pp. 1–5 (cit. on p. 33).

[47] J. C. Coleman. “Facial expressions of emotion.” In: *Psychological Monographs: General and Applied* 63.1 (1949), p. i (cit. on p. 21).

[48] L. Cosmides and J. Tooby. “Evolutionary psychology and the emotions”. In: *Handbook of emotions* 2.2 (2000), pp. 91–115 (cit. on p. 24).

[49] S. Costa et al. “Emotional storytelling using virtual and robotic agents”. In: *International Journal of Humanoid Robotics* 15.03 (2018), p. 1850006 (cit. on p. 3).

[50] S. Cox et al. “Tessa, a System to Aid Communication with Deaf People”. In: *Proceedings of the Fifth International ACM Conference on Assistive Technologies*. Assets ’02. Edinburgh, Scotland: Association for Computing Machinery, 2002, pp. 205–212. ISBN: 1581134649. DOI: 10.1145/638249.638287. URL: https://doi.org/10.1145/638249.638287 (cit. on p. 44).

[51] D. W. Cunningham et al. “The components of conversational facial expressions”. In: *Proceedings of the 1st Symposium on Applied Perception in Graphics and Visualization*. 2004, pp. 143–150 (cit. on pp. 3, 5, 36).

[52] B. Cvetković et al. “Real-time physical activity and mental stress management with a wristband and a smartphone”. In: *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. ACM. 2017, pp. 225–228 (cit. on p. 32).

[53] C. Darwin. *The Expression Of The Emotions In Man And Animals*. Oxford University Press, 1998-04. ISBN: 9780195112719. DOI: 10.1093/oso/9780195112719.001.0001. URL: https://doi.org/10.1093/oso/9780195112719.001.0001 (cit. on pp. 2, 3, 15, 21, 35).

[54] R. Davis. “The specificity of facial expressions”. In: *The Journal of General Psychology* 10.1 (1934), pp. 42–58 (cit. on p. 21).

[55] J. R. Davitz. *The language of emotion*. Academic Press, 2016 (cit. on p. 24).

[56] J. P. De Jong and D. N. Den Hartog. “How leaders influence employees’ innovative behaviour”. In: *European Journal of innovation management* (2007) (cit. on p. 37).

[57] T. Denmark et al. “Signing with the face: Emotional expression in narrative production in deaf children with autism spectrum disorder”. In: *Journal of autism and developmental disorders* 49 (2019), pp. 294–306 (cit. on p. 92).

[59] D. F. Dinges et al. “Optical computer recognition of facial expressions associated with stress induced by performance demands”. In: *Aviation, space, and environmental medicine* 76.6 (2005), B172–B182 (cit. on p. 101).

[60] S. Du and A. M. Martinez. “Compound facial expressions of emotion: from basic research to clinical applications”. In: *Dialogues in clinical neuroscience* 17.4 (2015), pp. 443–455 (cit. on p. 63).

[61] S. Du, Y. Tao, and A. M. Martinez. “Compound facial expressions of emotion”. In: *Proceedings of the national academy of sciences* 111.15 (2014), E1454–E1462 (cit. on pp. 4, 29).

[62] A. Duarte et al. “How2Sign: a large-scale multimodal dataset for continuous American sign language”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 2735–2744 (cit. on pp. 5, 43, 44).

[63] G.-B. Duchenne. *Mécanisme de la physionomie humaine ou analyse électro-physiologique de l’expression des passions*. Librairie J.-B. Baillière et fils, 1876 (cit. on pp. 2, 15, 16).

[64] D. Dupré et al. “Emotion recognition in humans and machine using posed and spontaneous facial expression”. In: (2019) (cit. on p. 4).

[65] P. Ekman, W. V. Friesen, and J. C. Hager. “Facial action coding system [CD-Rom]”. In: *Salt Lake City, UT: Nexus* (2002) (cit. on p. 17).

[66] P. Ekman. *About brows: emotional and conversational signals; in von Cranach M, Foppa K, Lepenies W, Ploog D (eds): Human Ethology*. 1979 (cit. on pp. 3, 35).

[67] P. Ekman. "An argument for basic emotions". In: *Cognition & emotion* 6.3-4 (1992), pp. 169–200 (cit. on pp. 28, 29).

[68] P. Ekman. "Basic emotions". In: *Handbook of cognition and emotion* 98.45-60 (1999), p. 16 (cit. on pp. 1, 42).

[69] P. Ekman. "Biological and cultural contributions to body and facial movement". In: *The anthropology of the body* (1977) (cit. on p. 26).

[70] P. Ekman. *Facial expressions of emotion: New findings, new questions*. 1992 (cit. on p. 102).

[71] P. Ekman. "Universals and cultural differences in facial expressions of emotion." In: *Nebraska symposium on motivation*. University of Nebraska Press. 1971 (cit. on pp. 21, 47, 101).

[72] P. Ekman, W. V. Freisen, and S. Ancoli. "Facial signs of emotional experience." In: *Journal of personality and social psychology* 39.6 (1980), p. 1125 (cit. on pp. 13, 30).

[73] P. Ekman and W. V. Friesen. "Constants across cultures in the face and emotion." In: *Journal of personality and social psychology* 17.2 (1971), p. 124 (cit. on p. 3).

[74] P. Ekman and W. V. Friesen. "Facial action coding system". In: *Environmental Psychology & Nonverbal Behavior* (1978) (cit. on pp. 15, 17, 43, 91, 97, 100).

[75] P. Ekman and W. V. Friesen. *Facial Action Coding System: Investigatoris Guide*. Consulting Psychologists Press, 1978 (cit. on pp. 3, 5, 49).

[76] E. A. Elliott and A. M. Jacobs. "Facial expressions, emotions, and sign languages". In: *Frontiers in psychology* 4 (2013), p. 39013 (cit. on pp. 42, 92).

[77] *Emotion AI provider. Facial Emotion Recognition*. February 2023. URL: https://www.morphcast.com/ (cit. on p. 3).

[78] *Emotion Recognition Software*. September 2022. URL: https://www.affectiva.com/ (cit. on p. 3).

[79] F. Eyben, M. Wöllmer, and B. Schuller. "Opensmile: the munich versatile and fast open-source audio feature extractor". In: *Proceedings of the 18th ACM international conference on Multimedia*. 2010, pp. 1459–1462 (cit. on p. 74).

[80] F. Eyben et al. "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing". In: *IEEE transactions on affective computing* 7.2 (2015), pp. 190–202 (cit. on p. 74).

[81] J. Fenlon, K. Cormier, and D. Brentari. "The phonology of sign languages". In: *The Routledge handbook of phonological theory*. Routledge, 2017, pp. 453–475 (cit. on p. 40).

[82] I. Fernández Sedano et al. "Differences between cultures in emotional verbal and non-verbal reactions". In: *Psicothema* 12.1 (2000), pp. 83–92 (cit. on p. 98).

[83] G. Fink. “Stress: definition and history”. In: *Stress science: neuroendocrinology* (2010), pp. 3–9 (cit. on pp. 30, 31).

[84] M. G. Frank and J. Stennett. “The forced-choice paradigm and the perception of facial expressions of emotion.” In: *Journal of personality and social psychology* 80.1 (2001), p. 75 (cit. on p. 36).

[85] A. J. Fridlund. *Human facial expression: An evolutionary view*. Academic press, 1994 (cit. on pp. 3, 35).

[86] A. J. Fridlund et al. “Emotions and facial expression”. In: *Science* 230.4726 (1985), pp. 607–608 (cit. on pp. 3, 4, 35).

[87] E Friesen and P. Ekman. “Facial action coding system: a technique for the measurement of facial movement”. In: *Palo Alto* 3.2 (1978), p. 5 (cit. on pp. 18, 74).

[88] W. V. Friesen, P. Ekman, et al. “EMFACS-7: Emotional facial action coding system”. In: *Unpublished manuscript, University of California at San Francisco* 2.36 (1983), p. 1 (cit. on pp. 18, 63).

[89] J Frois-Wittman. “The judgment of facial expression.” In: *Journal of Experimental Psychology* 13.2 (1930), p. 113 (cit. on p. 21).

[90] E. M. Gallaudet. “The Milan Convention”. In: *American Annals of the Deaf and Dumb* 26.1 (1881), pp. 1–16 (cit. on pp. 39, 41).

[91] H. Gao, A. Yüce, and J.-P. Thiran. “Detecting emotional stress from facial expressions for driving safety”. In: *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE. 2014, pp. 5961–5965 (cit. on pp. 33, 51, 59, 66).

[92] E. Garcia-Ceja, V. Osmani, and O. Mayora. “Automatic stress detection in working environments from smartphones’ accelerometer data: a first step”. In: *IEEE journal of biomedical and health informatics* 20.4 (2016), pp. 1053–1060 (cit. on p. 33).

[93] J. M. Garcia-Garcia, V. M. Penichet, and M. D. Lozano. “Emotion detection: a technology review”. In: *Proceedings of the XVIII international conference on human computer interaction*. 2017, pp. 1–8 (cit. on p. 38).

[94] M. Ghafurian, N. Budnarain, and J. Hoey. “Improving Humanness of Virtual Agents and Users’ Cooperation through Emotions”. In: *IEEE Transactions on Affective Computing* (2021) (cit. on p. 1).

[95] G Giannakakis et al. “Stress and anxiety detection using facial cues from videos”. In: *Biomedical Signal Processing and Control* 31 (2017), pp. 89–101 (cit. on p. 58).

[96] G. Giannakakis et al. “Automatic stress analysis from facial videos based on deep facial action units recognition”. In: *Pattern Analysis and Applications* (2022), pp. 1–15 (cit. on pp. 33, 34, 58).

[97] G. Giannakakis et al. “Review on psychological stress detection using biosignals”. In: *IEEE Transactions on Affective Computing* (2019) (cit. on p. 58).

[98] J. M. Girard et al. "Social risk and depression: Evidence from manual and automatic facial expression analysis". In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE. 2013, pp. 1–8 (cit. on p. 5).

[99] N. Gokul et al. "Addressing Resource Scarcity across Sign Languages with Multilingual Pretraining and Unified-Vocabulary Datasets". In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2022 (cit. on pp. 4, 5).

[100] F. L. Goodenough and M. A. Tinker. "The relative potency of facial expression and verbal description of stimulus in the judgment of emotion." In: *Journal of Comparative Psychology* 12.4 (1931), p. 365 (cit. on p. 21).

[101] P. Griffiths. "Is emotion a natural kind?" In: (2002) (cit. on p. 24).

[102] P. E. Griffiths. *The Problem of Psychological Categories*. Chicago: University of Chicago Press, 2008. ISBN: 9780226308760. DOI: doi:10.7208/9780226308760. URL: https://doi.org/10.7208/9780226308760 (cit. on pp. 24, 26).

[103] P. E. Group. URL: https://www.paulekman.com/resources/universal-facial-expressions/ (cit. on p. 28).

[104] S. Hajera and M. M. Ali. "A Comparative analysis of psychological stress detection methods". In: *IJCEM* 21.2 (2018) (cit. on pp. 31, 32).

[105] M. L. Hall, W. C. Hall, and N. K. Caselli. "Deaf children need language, not (just) speech". In: *First Language* 39.4 (2019), pp. 367–395 (cit. on p. 39).

[106] T. Hanke. "HamNoSys-representing sign language data in language resources and language processing contexts". In: *LREC*. Vol. 4. 2004, pp. 1–6 (cit. on pp. 43, 44).

[107] T. Hanke et al. "Extending the Public DGS Corpus in size and depth". In: *sign-lang@ LREC 2020*. European Language Resources Association (ELRA). 2020, pp. 75–82 (cit. on p. 5).

[108] S. Happy et al. "The Indian spontaneous expression database for emotion recognition". In: *IEEE Transactions on Affective Computing* 8.1 (2015), pp. 131–142 (cit. on p. 4).

[109] J. Hassard et al. "The cost of work-related stress to society: A systematic review." In: *Journal of Occupational Health Psychology* 23.1 (2018), p. 1 (cit. on p. 31).

[110] K. He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on p. 86).

[111] M. Helmy et al. "Predicting Parkinson disease related genes based on PyFeat and gradient boosted decision tree". In: *Scientific Reports* 12.1 (2022), p. 10004 (cit. on p. 5).

[112] F. Hernandez et al. “TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation”. In: *International Conference on Speech and Computer*. Springer. 2018, pp. 198–208 (cit. on p. 68).

[113] J. Hernandez, R. R. Morris, and R. W. Picard. “Call center stress recognition with person-specific models”. In: *International Conference on Affective Computing and Intelligent Interaction*. Springer. 2011, pp. 125–134 (cit. on p. 31).

[114] J. Hernandez et al. “Under pressure: sensing stress of computer users”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2014, pp. 51–60 (cit. on p. 33).

[115] D. Hernando et al. “Validation of the apple watch for heart rate variability measurements during relax and mental stress in healthy subjects”. In: *Sensors* 18.8 (2018), p. 2619 (cit. on p. 32).

[116] M. Honnibal and I. Montani. “spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing”. To appear. 2017 (cit. on pp. 92, 93).

[117] M. Inaba, F. Toriumi, and K. Ishii. “Automatic detection of “enthusiasm” in non-task-oriented dialogues using word co-occurrence”. In: *2011 IEEE Workshop on Affective Computational Intelligence (WACI)*. IEEE. 2011, pp. 1–7 (cit. on pp. 36, 70).

[118] *Intelligent Virtual Assistant (Iva) Market - Growth, Trends, Covid-19 Impact, And Forecasts (2023 - 2028)*. URL: https://www.mordorintelligence.com/industry-reports/intelligent-virtual-assistant-market (cit. on p. 7).

[119] M. Ivanović et al. “Emotional intelligence and agents: Survey and possible applications”. In: *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*. 2014, pp. 1–7 (cit. on p. 3).

[120] C. E. Izard. “The face of emotion.” In: (1971) (cit. on p. 21).

[121] C. E. Izard and C. E. Izard. “Differential emotions theory”. In: *Human emotions* (1977), pp. 43–66 (cit. on p. 3).

[122] C. E. Izard and M. Weiss. *Maximally discriminative facial movement coding system*. University of Delaware, instructional resources Center, 1979 (cit. on pp. 3, 15).

[123] R. E. Jack and P. G. Schyns. “Toward a social psychophysics of face communication”. In: *Annual review of psychology* 68 (2017), pp. 269–297 (cit. on p. 5).

[124] S. E. Jackson and C. Maslach. “After-effects of job-related stress: Families as victims”. In: *Journal of organizational behavior* 3.1 (1982), pp. 63–77 (cit. on p. 31).

[125] E. Jahn et al. “Publishing DGS corpus data: Different formats for different needs”. In: *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*. Vol. 2. 2018 (cit. on p. 43).

[126] W. James. “The emotions.” In: (1922) (cit. on p. 23).

[127] G. Johnson. *Theories of Emotion*. URL: https://iep.utm.edu/theories-of-emotion/#H3 (cit. on p. 22).

[128] T. Johnston and A. Schembri. *Australian Sign Language (Auslan): An introduction to sign language linguistics*. Cambridge University Press, 2007 (cit. on p. 39).

[129] T. Kanade, J. F. Cohn, and Y. Tian. "Comprehensive database for facial expression analysis". In: *Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. No. PR00580)*. IEEE. 2000, pp. 46–53 (cit. on pp. 3, 4).

[130] K. Karpouzis et al. "Educational resources and implementation of a Greek sign language synthesis architecture". In: *Comput. Educ.* 49 (2007), pp. 54–74 (cit. on p. 44).

[131] K. Kaulard et al. "The MPI facial expression database—a validated database of emotional and conversational facial expressions". In: *PloS one* 7.3 (2012), e32321 (cit. on pp. 1, 3–5, 36, 37, 47).

[132] C. F. Keating. "Channelling charisma through face and body status cues". In: *Social psychological dynamics* (2011), pp. 93–111 (cit. on pp. 8, 35).

[133] M. M. Keller et al. "Feeling and showing: A new conceptualization of dispositional teacher enthusiasm and its relation to students' interest". In: *Learning and Instruction* 33 (2014), pp. 29–38 (cit. on p. 38).

[134] M. M. Keller et al. "Teacher enthusiasm: Reviewing and redefining a complex construct". In: *Educational Psychology Review* 28.4 (2016), pp. 743–769 (cit. on pp. 35, 37, 70).

[135] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations (ICLR)*. 2015. URL: http://arxiv.org/abs/1412.6980 (cit. on p. 75).

[136] R. E. Kleck et al. "Effects of being observed on expressive, subjective, and physiological responses to painful stimuli." In: *Journal of Personality and Social Psychology* 34.6 (1976), p. 1211 (cit. on p. 20).

[138] K. Kozik. *Without sign language, deaf people are not equal*. October 2020. URL: https://www.hrw.org/news/2019/09/23/without-sign-language-deaf-people-are-not-equal# (cit. on p. 39).

[139] M. Kunz and S. Lautenbacher. "The faces of pain: a cluster analysis of individual differences in facial activity patterns of pain". In: *European Journal of Pain* 18.6 (2014), pp. 813–823 (cit. on p. 60).

[140] P. Kuppens et al. "The relation between valence and arousal in subjective experience." In: *Psychological bulletin* 139.4 (2013), p. 917 (cit. on p. 6).

[141] H. Kurniawan, A. V. Maslov, and M. Pechenizkiy. "Stress detection from speech and galvanic skin response signals". In: *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*. IEEE. 2013, pp. 209–214 (cit. on p. 33).

[142] C. Landis. “Studies of Emotional Reactions. II. General Behavior and Facial Expression.” In: *Journal of Comparative Psychology* 4.5 (1924), p. 447 (cit. on pp. 2, 21).

[143] C. Landis. “The interpretation of facial expression in emotion”. In: *The Journal of General Psychology* 2.1 (1929), pp. 59–72 (cit. on pp. 2, 21).

[144] J. R. Landis and G. G. Koch. “The measurement of observer agreement for categorical data”. In: *Biometrics* (1977), pp. 159–174 (cit. on pp. 70, 71).

[145] S. hon Lau. *Stress Detection for Keystroke Dynamics*. Tech. rep. CMU-ML-18-104. Carnegie Mellon University, 2018-05 (cit. on pp. 50, 57, 102).

[146] R. S. Lazarus. *Emotion and adaptation*. Oxford University Press, 1991 (cit. on p. 25).

[147] R. S. Lazarus. “Psychological stress and the coping process.” In: (1966) (cit. on p. 30).

[148] R. S. Lazarus and S. Folkman. *Stress, appraisal, and coping*. Springer publishing company, 1984 (cit. on pp. 23, 30).

[149] C. Le Brun. *Méthode pour apprendre à dessiner les passions, proposée dans une conférence sur l’expression générale, et particulière*. Chez François van-der Plaats, 1702 (cit. on pp. 2, 15).

[150] J. S. Lerner et al. “Facial expressions of emotion reveal neuroendocrine and cardiovascular stress responses”. In: *Biological psychiatry* 61.2 (2007), pp. 253–260 (cit. on pp. 33, 51, 101).

[151] J. Li et al. “MEGC2022: ACM Multimedia 2022 Micro-Expression Grand Challenge”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 7170–7174 (cit. on p. 4).

[152] Y. Li, T. Zhao, and T. Kawahara. “Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning.” In: *Interspeech*. 2019, pp. 2803–2807 (cit. on p. 71).

[153] T. W. Liew, N. A. M. Zin, and N. Sahari. “Exploring the affective, motivational and cognitive effects of pedagogical agent enthusiasm in a multimedia learning environment”. In: *Human-centric Computing and Information Sciences* 7.1 (2017), p. 9 (cit. on p. 35).

[154] T. W. Liew et al. “Does speaker’s voice enthusiasm affect social cue, cognitive load and transfer in multimedia learning?” In: *Information and Learning Sciences* (2020) (cit. on pp. 35, 38).

[155] Z. Lin, H. T. Ng, and M.-Y. Kan. “A PDTB-styled end-to-end discourse parser”. In: *Natural Language Engineering* 20.2 (2014), pp. 151–184 (cit. on p. 74).

[156] K. A. Lindquist et al. “The brain basis of emotion: a meta-analytic review”. In: *The Behavioral and brain sciences* 35.3 (2012), p. 121 (cit. on p. 26).

[157] *Linkura Program*. https://www.linkura.com/linkuraprogrammet. Accessed: 2018-05-16 (cit. on p. 31).

[158] J. M. Lourenço. *The NOVAthesis LATEX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/master/template.pdf (cit. on p. i).

[159] P. Lucey et al. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression". In: *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE. 2010, pp. 94–101 (cit. on p. 54).

[160] W. Lyons. "The philosophy of cognition and emotion". In: *Handbook of cognition and emotion* (1999), pp. 21–44 (cit. on p. 2).

[161] Y. Ma et al. "A survey on empathetic dialogue systems". In: *Information Fusion* 64 (2020), pp. 50–70 (cit. on p. 3).

[162] L. v. d. Maaten and G. Hinton. "Visualizing data using t-SNE". In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605 (cit. on p. 54).

[163] W. C. Mann and S. A. Thompson. "Rhetorical structure theory: Toward a functional theory of text organization". In: *Text* 8.3 (1988), pp. 243–281 (cit. on p. 74).

[164] I. Mansutti et al. "Individuals with hearing impairment/deafness during the COVID-19 pandemic: a rapid review on communication challenges and strategies". In: *Journal of Clinical Nursing* 32.15-16 (2023), pp. 4454–4472 (cit. on p. 41).

[165] C. Marechal et al. "Survey on AI-Based Multimodal Methods for Emotion Detection". In: *High-Performance Modelling and Simulation for Big Data Applications: Selected Results of the COST Action IC1406 cHiPSet*. Cham: Springer International Publishing, 2019, pp. 307–324. ISBN: 978-3-030-16272-6. DOI: 10.1007/978-3-030-16272-6_11 (cit. on p. 38).

[166] M. Marschark and P. E. Spencer. *The Oxford handbook of deaf studies, language, and education, vol. 2*. Oxford University Press, 2010 (cit. on p. 4).

[167] A. A. Marsh, S. A. Rhoads, and R. M. Ryan. "A multi-semester classroom demonstration yields evidence in support of the facial feedback effect." In: *Emotion* 19.8 (2019), p. 1500 (cit. on p. 23).

[168] B. Martinez et al. "Automatic analysis of facial actions: A survey". In: *IEEE transactions on affective computing* 10.3 (2017), pp. 325–347 (cit. on p. 5).

[169] K. P. Masuku, N. Moroe, and D. van der Merwe. "'The world is not only for hearing people–It's for all people': The experiences of women who are deaf or hard of hearing in accessing healthcare services in Johannesburg, South Africa". In: *African Journal of Disability* 10 (2021) (cit. on p. 40).

[170] D. Matsumoto and P. Ekman. “Facial expression analysis”. In: *Scholarpedia* 3.5 (2008). revision #88993, p. 4237. DOI: 10.4249/scholarpedia.4237 (cit. on pp. 13, 15, 16, 101).

[171] M. Mavadati, P. Sanger, and M. H. Mahoor. “Extended disfa dataset: Investigating posed and spontaneous facial expressions”. In: *proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2016, pp. 1–8 (cit. on p. 3).

[172] S. M. Mavadati et al. “Disfa: A spontaneous facial action intensity database”. In: *IEEE Transactions on Affective Computing* 4.2 (2013), pp. 151–160 (cit. on p. 4).

[173] L. M. Mayo and M. Heilig. “In the face of stress: Interpreting individual differences in stress-induced facial expressions”. In: *Neurobiology of stress* 10 (2019), p. 100166 (cit. on p. 30).

[174] J. Mcdonald et al. “An Automated Technique for Real-Time Production of Lifelike Animations of American Sign Language”. In: *Univers. Access Inf. Soc.* 15.4 (2016), pp. 551–566. ISSN: 1615-5289. DOI: 10.1007/s10209-015-0407-2. URL: https://doi.org/10.1007/s10209-015-0407-2 (cit. on p. 44).

[175] M. L. McHugh. “Interrater reliability: the kappa statistic”. In: *Biochemia medica* 22.3 (2012), pp. 276–282 (cit. on pp. 63, 71).

[176] M. McKee, C. Moran, and P. Zazove. “Overcoming additional barriers to care for deaf and hard of hearing patients during COVID-19”. In: *JAMA Otolaryngology–Head & Neck Surgery* 146.9 (2020), pp. 781–782 (cit. on p. 40).

[177] G. McKeown et al. “The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent”. In: *IEEE transactions on affective computing* 3.1 (2011), pp. 5–17 (cit. on p. 5).

[178] A. Mehrabian. “Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament”. In: *Current Psychology* 14 (1996), pp. 261–292 (cit. on pp. 28, 29).

[179] E. Merdivan et al. “Dialogue systems for intelligent human computer interactions”. In: *Electronic Notes in Theoretical Computer Science* 343 (2019), pp. 57–71 (cit. on p. 3).

[180] *Metaverse market size, share, trends, analysis and forecasts by Vertical (BFSI, retail, media and entertainment, education, Aerospace and defense, manufacturing, others), component stack (hardware, software, services), region and segment 2022-2030*. February 2023. URL: https://www.globaldata.com/store/report/metaverse-market-analysis/ (cit. on p. 7).

[181] S. Michie. “Causes and management of stress at work”. In: *Occupational and environmental medicine* 59.1 (2002), pp. 67–72 (cit. on p. 31).

[182] J. A. Mikels et al. “Emotional category data on images from the International Affective Picture System”. In: *Behavior research methods* 37.4 (2005), pp. 626–630 (cit. on pp. 28, 29).

[183] E. Miltsakaki et al. "The Penn Discourse Treebank". In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* (2004) (cit. on p. 74).

[184] A. Mineiro, J. Jardim, and A. Morais. "Para além das mãos: elementos para o estudo da expressão facial (EF) em Língua Gestual Portuguesa (LGP)". In: *Cardernos de Saúde* 4 (2011), pp. 37–42 (cit. on p. 41).

[185] B. Mirela and C. Mădălina-Adriana. "Organizational stress and its impact on work performance". In: *Conference Proceedings, European Integration–New Challenges*. 2011, pp. 1622–1628 (cit. on p. 31).

[186] S. Mishra and J. Diesner. "Capturing Signals of Enthusiasm and Support Towards Social Issues from Twitter". In: *Proceedings of the 5th International Workshop on Social Media World Sensors*. 2019, pp. 19–24 (cit. on p. 37).

[187] A. Mollahosseini, B. Hasani, and M. H. Mahoor. "Affectnet: A database for facial expression, valence, and arousal computing in the wild". In: *IEEE Transactions on Affective Computing* 10.1 (2017), pp. 18–31 (cit. on pp. 4, 6).

[188] E Morales-Vargas, C. Reyes-García, and H. Peregrina-Barreto. "On the use of Action Units and fuzzy explanatory models for facial expression recognition". In: *PloS one* 14.10 (2019), e0223563 (cit. on p. 63).

[189] M. B. Narayanan et al. "Sign Language Translation Using Multi Context Transformer". In: *Advances in Soft Computing*. Ed. by I. Batyrshin, A. Gelbukh, and G. Sidorov. Cham: Springer International Publishing, 2021, pp. 311–324. ISBN: 978-3-030-89820-5 (cit. on p. 92).

[190] R. M. Nesse. "Evolutionary explanations of emotions". In: *Human nature* 1.3 (1990), pp. 261–289 (cit. on p. 24).

[191] J. L. Newton. "Familiar with the Deaf World: The Influence of Alexander Graham Bell and Oralism on the History of the North American Deaf Community". PhD thesis. Sam Houston State University, 2020 (cit. on p. 39).

[192] O. Niebuhr. "Space fighters on stage-How the F1 and F2 vowel-space dimensions contribute to perceived speaker charisma". In: *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2020* (2020), pp. 265–277 (cit. on p. 37).

[193] O. Niebuhr, J. Voße, and A. Brem. "What makes a charismatic speaker? A computer-based acoustic-prosodic analysis of Steve Jobs tone of voice". In: *Computers in Human Behavior* 64 (2016), pp. 366–382 (cit. on p. 37).

[194] M. Norden et al. "Automatic Detection of Subjective, Annotated and Physiological Stress Responses from Video Data". In: *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2022, pp. 1–8 (cit. on p. 33).

[195] A. Núñez-Marcos, O. Perez-de Viñaspre, and G. Labaka. "A survey on Sign Language machine translation". In: *Expert Systems with Applications* 213 (2023), p. 118993 (cit. on p. 44).

[196] F. Nunnari, C. España Bonet, and E. Avramidis. "A Data Augmentation Approach for Sign-Language-To-Text Translation In-The-Wild". In: *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Ed. by D. Gromann et al. Vol. 93. Open Access Series in Informatics (OASIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, 36:1–36:8. ISBN: 978-3-95977-199-3. DOI: 10.4230/OASIcs.LDK.2021.36. URL: https://drops.dagstuhl.de/opus/volltexte/2021/14572 (cit. on p. 92).

[197] V. Nyst. "Sign language fieldwork". In: *Research methods in sign language studies: A practical guide* (2015), pp. 105–122 (cit. on p. 41).

[198] OpenStax. *1106 front and side views of the muscles of facial expressions*. URL: https://commons.wikimedia.org/wiki/File:1106_Front_and_Side_Views_of_the_Muscles_of_Facial_Expressions.jpg (cit. on p. 17).

[199] A. Ortony, G. L. Clore, and A. Collins. "The cognitive structure of emotions." In: (1988) (cit. on p. 25).

[200] H. Oster. "Baby FACS: Facial Action Coding System for infants and young children". In: *Unpublished monograph and coding manual. New York University* (2006) (cit. on p. 18).

[201] B. Pan et al. "A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods". In: *Neurocomputing* (2023), p. 126866 (cit. on pp. 28, 29).

[202] B. Parkinson. "Emotions are social". In: *British journal of psychology* 87.4 (1996), pp. 663–683 (cit. on p. 24).

[203] W. G. Parrott. *Emotions in social psychology: Essential readings*. psychology press, 2001 (cit. on pp. 28, 29).

[204] M. Pasikowska-Schnass. September 2018. URL: https://www.europarl.europa.eu/RegData/etudes/ATAG/2018/625196/EPRS_ATA(2018)625196_EN.pdf (cit. on p. 40).

[205] C. Pelachaud. "Greta: A conversing socio-emotional agent". In: *Proceedings of the 1st acm sigchi international workshop on investigating social interactions with artificial agents*. 2017, pp. 9–10 (cit. on p. 3).

[206] E. Petrocchi. "Prosody of emotions: The relation between (prosodical) linguistic and affective functions of non-manual components in Italian Sign Language". In: *Ca'Foscari University of Venice* (2021) (cit. on p. 92).

[207] R. Pfau, M. Steinbach, and B. Woll. *Sign language: An international handbook*. De Gruyter Mouton, 2012 (cit. on p. 41).

[208] R. W. Picard. *Affective computing*. MIT press, 2000 (cit. on p. 1).

[209] N. Pimenta and R. M. d. Quadros. "Curso de LIBRAS 1". In: *Rio de Janeiro: LSB vídeo* (2006) (cit. on p. 42).

[210] R. Plutchik. "A general psychoevolutionary theory of emotion". In: *Theories of emotion*. Elsevier, 1980, pp. 3–33 (cit. on pp. 28, 29).

[211] R. Plutchik. "Emotions: A general psychoevolutionary theory". In: *Approaches to emotion* 1984.197-219 (1984), pp. 2–4 (cit. on p. 24).

[212] R. Prasad et al. "The Penn Discourse Treebank 2.0". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)* (2008) (cit. on p. 74).

[213] C. Price and E. A. Walle. "History and Beyond". In: *Emotion Researcher* 1.March (2018). URL: http://emotionresearcher.com/wp-content/uploads/2018/03/Emotion-Researcher-March-2018.pdf (cit. on p. 21).

[214] F. Principi et al. "The Florence 4D Facial Expression Dataset". In: *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE. 2023, pp. 1–6 (cit. on p. 4).

[215] E. Raniolo. "COVID-19 pandemic in the eyes of deaf people: sign language translation in Italy". In: *Antonio Lavieri* (), p. 71 (cit. on p. 41).

[216] R. Rastgoo, K. Kiani, and S. Escalera. "Real-time isolated hand sign language recognition using deep networks and SVD". In: *Journal of Ambient Intelligence and Humanized Computing* 13.1 (2022), pp. 591–611 (cit. on p. 41).

[217] R. Rastgoo, K. Kiani, and S. Escalera. "Sign language recognition: A deep survey". In: *Expert Systems with Applications* 164 (2021), p. 113794 (cit. on pp. 41, 42, 44).

[218] R. Rastgoo et al. "Sign Language Production: A Review". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2021, pp. 3451–3461 (cit. on p. 42).

[219] W. E. Rinn. "The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions." In: *Psychological bulletin* 95.1 (1984), p. 52 (cit. on p. 17).

[220] J. Rodriguez and F. Martínez. "How important is motion in sign language translation?" In: *IET Computer Vision* 15.3 (2021), pp. 224–234. DOI: https://doi.org/10.1049/cvi2.12037. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/cvi2.12037. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cvi2.12037 (cit. on p. 92).

[221] M. Z. Rosaldo. *Knowledge and passion*. Vol. 4. Cambridge University Press, 1980 (cit. on p. 24).

[222] R. Rosaldo. “Grief and a headhunter’s rage: On the cultural force of emotions”. In: *Text, play and story: The construction and reconstruction of self and society* (1984), pp. 178–195 (cit. on p. 24).

[223] I. Roseman and C. Smith. “Appraisal theory: Overview, assumptions, varieties, controversies (2001) Appraisal Processes in Emotion: Theory, Methods, Research”. In: *Oxford University Press New York* (), pp. 3–19 (cit. on p. 25).

[224] E. L. Rosenberg. “3 Introduction: The Study of Spontaneous Facial Expressions in Psychology”. In: *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, 2005-04. ISBN: 9780195179644. DOI: 10.1093/acprof:oso/9780195179644.003.0001. eprint: https://academic.oup.com/book/0/chapter/270563981/chapter-ag-pdf/44537684/book_32646_section_270563981.ag.pdf. URL: https://doi.org/10.1093/acprof:oso/9780195179644.003.0001 (cit. on pp. 17–19).

[225] E. L. Rosenberg and P. Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997 (cit. on p. 4).

[226] J. A. Russell and J. M. Fernandez-Dols. *The psychology of facial expression*. Cambridge university press, 1997 (cit. on pp. 3, 35).

[227] E. Saad et al. “Enthusiastic Robots Make Better Contact.” In: *IROS*. 2019, pp. 1094–1100 (cit. on p. 35).

[228] M. R. Salleh. “Life event, stress and illness”. In: *The Malaysian journal of medical sciences: MJMS* 15.4 (2008), p. 9 (cit. on pp. 1, 31).

[229] B. Sandberg. “Enthusiasm in the development of radical innovations”. In: *Creativity and Innovation Management* 16.3 (2007), pp. 265–273 (cit. on pp. 8, 35).

[230] W. Sandler. “Prosody and syntax in sign languages”. In: *Transactions of the philological society* 108.3 (2010), pp. 298–328 (cit. on p. 41).

[231] W. Sandler. “The phonological organization of sign languages”. In: *Language Linguistics Compass* 6 (3 2012), pp. 162–182 (cit. on p. 42).

[232] W. Sandler and D. Lillo-Martin. *Sign language and linguistic universals*. Cambridge University Press, 2006 (cit. on pp. 39, 98).

[233] B. Saunders, N. C. Camgoz, and R. Bowden. “Everybody Sign Now: Translating Spoken Language to Photo Realistic Sign Language Video”. In: *arXiv preprint arXiv:2011.09846* (2020) (cit. on pp. 41, 44).

[234] B. Saunders, N. C. Camgoz, and R. Bowden. “Progressive transformers for end-to-end sign language production”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 687–705 (cit. on pp. 85–88).

[235] B. Saunders, N. C. Camgoz, and R. Bowden. “Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5141–5151 (cit. on pp. 41, 44).

[236] B Savers, H. Beagley, and W. Henshall. “The mechanism of auditory evoked EEG responses”. In: *Nature* 247.5441 (1974), pp. 481–483 (cit. on p. 26).

[237] A. Scarantino and R. de Sousa. “Emotion”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University, 2021 (cit. on p. 22).

[238] S. Schachter and J. Singer. “Cognitive, social, and physiological determinants of emotional state.” In: *Psychological review* 69.5 (1962), p. 379 (cit. on p. 23).

[239] K. R. Scherer, A. Schorr, and T. Johnstone. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001 (cit. on p. 25).

[240] H. Schlosberg. “Three dimensions of emotion.” In: *Psychological review* 61.2 (1954), p. 81 (cit. on pp. 28, 29).

[241] L. M. Schreiber, G. D. Paul, and L. R. Shibley. “The development and test of the public speaking competence rubric”. In: *Communication Education* 61.3 (2012), pp. 205–233 (cit. on p. 69).

[242] B. Schuller, S. Steidl, and A. Batliner. “The interspeech 2009 emotion challenge”. In: *Tenth Annual Conference of the International Speech Communication Association*. 2009 (cit. on p. 74).

[243] B. Schuller et al. “The INTERSPEECH 2010 paralinguistic challenge”. In: *Eleventh Annual Conference of the International Speech Communication Association*. 2010 (cit. on p. 74).

[244] B. Schuller et al. “The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism”. In: *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*. 2013 (cit. on p. 74).

[245] H. Selye. “Stress without Distress”. In: *Psychopathology of Human Adaptation*. Ed. by G. Serban. Boston, MA: Springer US, 1976, pp. 137–146. ISBN: 978-1-4684-2238-2. DOI: 10.1007/978-1-4684-2238-2_9. URL: https://doi.org/10.1007/978-1-4684-2238-2_9 (cit. on p. 30).

[246] D. Seuss et al. “Emotion expression from different angles: A video database for facial expressions of actors shot by a camera array”. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2019, pp. 35–41 (cit. on p. 5).

[247] A. Seyeditabari, N. Tabari, and W. Zadrozny. “Emotion detection in text: a review”. In: *arXiv preprint arXiv:1806.00674* (2018) (cit. on p. 38).

[248] K. R. Shahapure and C. Nicholas. “Cluster quality analysis using silhouette score”. In: *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*. IEEE. 2020, pp. 747–748 (cit. on p. 61).

[249] S Sharma, R. Gupta, and A Kumar. “Continuous sign language recognition using isolated signs data and deep transfer learning”. In: *Journal of Ambient Intelligence and Humanized Computing* (2023), pp. 1–12 (cit. on p. 41).

[250] B. Shi et al. “Open-domain sign language translation learned from online video”. In: *arXiv preprint arXiv:2205.12870* (2022) (cit. on p. 5).

[251] E. P. da Silva et al. “Facial action unit detection methodology with application in Brazilian sign language recognition”. In: *Pattern Analysis and Applications* (2021), pp. 1–17 (cit. on p. 42).

[252] E. P. da Silva et al. “Recognition of affective and grammatical facial expressions: a study for Brazilian sign language”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 218–236 (cit. on pp. 3, 4, 8, 40, 42).

[253] E. P. da Silva et al. “Silfa: Sign language facial action database for the development of assistive technologies for the deaf”. In: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE. 2020, pp. 688–692 (cit. on pp. 4, 5, 41, 43).

[254] O. Simantiraki et al. “Stress Detection from Speech Using Spectral Slope Measurements”. In: *Pervasive Computing Paradigms for Mental Health*. Springer, 2016, pp. 41–50 (cit. on p. 33).

[255] V. Skobov and Y. Lepage. “Video-to-hamnosys automated annotation system”. In: *sign-lang@ LREC 2020*. European Language Resources Association (ELRA). 2020, pp. 209–216 (cit. on p. 43).

[256] H. Sloetjes and P. Wittenburg. “Annotation by category - ELAN and ISO DCR”. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*. 2008 (cit. on p. 96).

[257] *Social Robots Market is experience an immense growth in near future by 2029 with leading CAGR, top countries and regional data, regional status of key players, SWOT analysis*. February 2023. URL: `https://www.marketwatch.com/press-release/social-robots-market-is-experience-an-immense-growth-in-near-future-by-2029-with-leading-cagr-top-countries-and-regional-data-regional-status-of-key-players-swot-analysis-2023-02-02` (cit. on p. 7).

[258] R. C. Solomon. “The logic of emotion”. In: *Nous* (1977), pp. 41–49 (cit. on p. 25).

[259] R. C. Solomon. *The passions: Emotions and the meaning of life*. Hackett Publishing, 1993 (cit. on p. 25).

[260] H. Spencer. “The principles of psychology”. In: *New York: D. Appleton & Company* (1855) (cit. on p. 21).

[261] M. E. Spencer. “What is charisma?” In: *The British Journal of Sociology* 24.3 (1973), pp. 341–354 (cit. on p. 37).

[262] Spread the Sign. *About Us*. 2017. URL: https://www.spreadthesign.com/isl.intl/about/ (cit. on p. 39).

[263] W. C. Stokoe Jr. “Sign language structure: An outline of the visual communication systems of the American deaf”. In: *Journal of deaf studies and deaf education* 10.1 (2005), pp. 3–37 (cit. on pp. 39, 41).

[264] R. S. Sudhakar and M. C. Anil. “Analysis of speech features for emotion detection: a review”. In: *2015 International Conference on Computing Communication Control and Automation*. IEEE. 2015, pp. 661–664 (cit. on p. 38).

[265] F.-T. Sun et al. “Activity-aware mental stress detection using physiological sensors”. In: *International Conference on Mobile Computing, Applications, and Services*. Springer. 2010, pp. 282–301 (cit. on p. 31).

[266] Z. Sun et al. “Estimating Stress in Online Meetings by Remote Physiological Signal and Behavioral Features”. In: *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*. 2022, pp. 216–220 (cit. on p. 33).

[267] V. Sutton. *SignWriting*. 1974. URL: https://www.signwriting.org/about/what/what02.html (cit. on pp. 43, 44).

[268] M. Syakur et al. “Integration k-means clustering method and elbow method for identification of the best customer profile cluster”. In: *IOP conference series: materials science and engineering*. Vol. 336. IOP Publishing. 2018, p. 012017 (cit. on p. 61).

[269] M. Sysoev, A. Kos, and M. Pogačnik. “Noninvasive stress recognition considering the current activity”. In: *Personal and Ubiquitous Computing* 19.7 (2015), pp. 1045–1052 (cit. on p. 33).

[270] F. Sze. “From gestures to grammatical non-manuals in sign language: A case study of polar questions and negation in Hong Kong Sign Language”. In: *Lingua* 267 (2022), p. 103188 (cit. on pp. 40, 41, 92).

[271] H. Talk. *Hand Talk App (Version 4.2.0)*. 2023-10. URL: https://www.handtalk.me/en/app/ (visited on 2023-10-21) (cit. on p. 4).

[272] R. Tokuhisa and R. Terashima. “Relationship between utterances and “enthusiasm” in non-task-oriented conversational dialogue”. In: *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*. 2006, pp. 161–167 (cit. on p. 37).

[273] S. Tomkins. *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company, 1962 (cit. on p. 21).

[274] S. S. Tomkins and R. McCarter. “What and where are the primary affects? Some evidence for a theory”. In: *Perceptual and motor skills* 18.1 (1964), pp. 119–158 (cit. on p. 21).

[275] F. De la Torre and J. F. Cohn. "Facial expression analysis". In: *Visual analysis of humans* (2011), pp. 377–409 (cit. on p. 18).

[276] A. Touroutoglou et al. "Intrinsic connectivity in the human brain does not reveal networks for 'basic'emotions". In: *Social cognitive and affective neuroscience* 10.9 (2015), pp. 1257–1265 (cit. on p. 26).

[277] J. Trzeciak Huss and J. Huss. "Deaf, not invisible: sign language interpreting in a global pandemic". In: *AJOB neuroscience* 12.4 (2021), pp. 280–283 (cit. on p. 41).

[278] C. O. Tze et al. "Neural sign reenactor: Deep photorealistic sign language retargeting". In: *arXiv preprint arXiv:2209.01470* (2022) (cit. on pp. 41, 44).

[279] D. Uthus, G. Tanzer, and M. Georg. "Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus". In: *Advances in Neural Information Processing Systems* 36 (2024) (cit. on p. 43).

[280] D. Uthus, G. Tanzer, and M. Georg. "YouTube-ASL: A Large-Scale, Open-Domain American Sign Language-English Parallel Corpus". In: *arXiv preprint arXiv:2306.15162* (2023) (cit. on p. 5).

[281] E. Vahdani et al. "Recognizing american sign language nonmanual signal grammar errors in continuous videos". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 1–8 (cit. on pp. 41, 42).

[282] C. Valli and C. Lucas. *Linguistics of American sign language: An introduction*. Gallaudet University Press, 2000 (cit. on p. 39).

[283] A. Van Staden, G. Badenhorst, and E. Ridge. "The benefits of sign language for deaf learners with language challenges". In: *Per Linguam: a Journal of Language Learning= Per Linguam: Tydskrif vir Taalaanleer* 25.1 (2009), pp. 44–60 (cit. on p. 39).

[284] M. Vázquez-Enríquez et al. "Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 3462–3471 (cit. on p. 41).

[285] P. E. Velmovitsky et al. "Can heart rate variability data from the Apple Watch electrocardiogram quantify stress?" In: *Frontiers in Public Health* 11 (2023) (cit. on p. 32).

[286] P. E. Velmovitsky et al. "Using apple watch ECG data for heart rate variability monitoring and stress prediction: A pilot study". In: *Frontiers in Digital Health* 4 (2022), p. 1058826 (cit. on p. 32).

[287] M. Vernon and S. D. Koh. "Effects of oral preschool compared to early manual communication on education and communication in deaf children". In: *American Annals of the Deaf* (1971), pp. 569–574 (cit. on p. 39).

[288] S.-J. Vick et al. "A cross-species comparison of facial morphology and movement in humans and chimpanzees using the facial action coding system (FACS)". In: *Journal of nonverbal behavior* 31.1 (2007), pp. 1–20 (cit. on p. 18).

[289] C. Viegas and M. Alikhani. "Entheos: A Multimodal Dataset for Studying Enthusiasm". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, 2021-08, pp. 2047–2060. DOI: 10.18653/v1/2021.findings-acl.180. URL: https://aclanthology.org/2021.findings-acl.180 (cit. on pp. 8, 97).

[290] C. Viegas and M. Alikhani. "Including Enthusiasm in Human–AI Communication". In: *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents*. 2021 (cit. on p. 1).

[291] C. Viegas et al. "Including Facial Expressions in Contextual Embeddings for Sign Language Generation". In: *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*. Ed. by A. Palmer and J. Camacho-collados. Toronto, Canada: Association for Computational Linguistics, 2023-07, pp. 1–10. DOI: 10.18653/v1/2023.starsem-1.1. URL: https://aclanthology.org/2023.starsem-1.1 (cit. on p. 8).

[292] C. Viegas et al. "Spark Creativity by Speaking Enthusiastically: Communication Training Using an E-Coach". In: *Proceedings of the 2020 International Conference on Multimodal Interaction*. ICMI '20. Virtual Event, Netherlands: Association for Computing Machinery, 2020, pp. 764–765. ISBN: 9781450375818. DOI: 10.1145/3382507.3421164. URL: https://doi.org/10.1145/3382507.3421164 (cit. on pp. 1, 38).

[293] C. Viegas et al. "The Seven Faces of Stress: Understanding Facial Activity Patterns during Cognitive Stress". In: *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE. 2024 (cit. on p. 97).

[294] C. Viegas et al. "Towards Independent Stress Detection: A Dependent Model Using Facial Action Units". In: *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE. 2018, pp. 1–6 (cit. on pp. 33, 58, 97).

[295] E. Vildjiounaite et al. "Unsupervised Stress Detection Algorithm and Experiments with Real Life Data". In: *Portuguese Conference on Artificial Intelligence*. Springer. 2017, pp. 95–107 (cit. on p. 33).

[296] M. V. Villarejo, B. G. Zapirain, and A. M. Zorrilla. "A stress sensor based on Galvanic Skin Response (GSR) controlled by ZigBee". In: *Sensors* 12.5 (2012), pp. 6075–6101 (cit. on p. 31).

[297] K. Vytal and S. Hamann. "Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis". In: *Journal of cognitive neuroscience* 22.12 (2010), pp. 2864–2885 (cit. on p. 26).

[298] C. Wallraven, A. Shin, and F. Biessmann. "Valence and arousal underlie evaluation of emotional and conversational facial expressions across cultures". In: *Journal of Vision* 14.10 (2014), pp. 210–210 (cit. on p. 6).

[299] Y. Wang, S. Li, and H. Wang. "A two-stage parsing method for text-level discourse analysis". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2017, pp. 184–188 (cit. on p. 74).

[300] Y. Wang, S. Li, and J. Yang. "Toward Fast and Accurate Neural Discourse Segmentation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018-08, pp. 962–967. DOI: 10.18653/v1/D18-1116. URL: https://www.aclweb.org/anthology/D18-1116 (cit. on p. 75).

[301] J. Whitehouse et al. "Signal value of stress behaviour". In: *Evolution and human behavior* 43.4 (2022), pp. 325–333 (cit. on p. 30).

[302] K. Yin and J. Read. "Better Sign Language Translation with STMC-Transformer". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020-12, pp. 5975–5989. DOI: 10.18653/v1/2020.coling-main.525. URL: https://aclanthology.org/2020.coling-main.525 (cit. on p. 44).

[303] A. Yüce et al. "Action units and their cross-correlations for prediction of cognitive load during driving". In: *IEEE Transactions on Affective Computing* 8.2 (2017), pp. 161–175 (cit. on pp. 33, 57).

[304] Q. Zhang. "Assessing the effects of instructor enthusiasm on classroom engagement, learning goal orientation, and academic self-efficacy". In: *Communication Teacher* 28.1 (2014), pp. 44–56 (cit. on pp. 1, 35, 38).

[305] X. Zhang et al. "A high-resolution spontaneous 3d dynamic facial expression database". In: *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE. 2013, pp. 1–6 (cit. on p. 3).

[306] Z. Zhang et al. "Multimodal spontaneous emotion corpus for human behavior analysis". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3438–3446 (cit. on p. 5).

[307] H. Zhou et al. "Improving Sign Language Translation with Monolingual Data by Sign Back-Translation". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 1316–1325 (cit. on p. 44).

[308] H. Zhou et al. "Spatial-Temporal Multi-Cue Network for Sign Language Recognition and Translation". In: *IEEE Transactions on Multimedia* 24 (2022), pp. 768–779. DOI: 10.1109/TMM.2021.3059098 (cit. on p. 44).

A

Dataset Contributions

A.1 Stress Dataset

Data available on project website:
https://github.com/clviegas/SevenFacesOfStress

- 115 subjects (48 male and 67 female)
- original goal of the dataset was to determine if detecting stress through keystroke dynamics is possible
- frontal video recordings with Microsoft Life Studio Pro webcam with 1080p resolution, capturing at 30 frames per second
- stress phase approximately 15 minutes
- qualitative (ECG and blood pressure data) and quantitative (State-Trait Anxiety Inventory (STAI) and NASA Taskload Work Index (NASA-TLX) questionnaires) evidence of actual stress state during the stressor task

More details in Chapter 4.2.

A.2 Entheos Dataset

Data available on project website:
https://github.com/clviegas/Entheos-Dataset

- original data from TEDLIUM
- 113 different TED talk speeches (60 male and 53 female)
- 1,126 sentence samples
- Label distribution: 123 monotonous, 848 normal, and 155 enthusiasm samples

More details in Chapter 5.2

A.3 Sign Language Dataset

Data available on project website: https://github.com/clviegas/dgs-adjectives

- German Sign Language (DGS)
- original data from RWTH-PHOENIX-2014T Weather Forecast dataset
- 205 semi-automatically annotated excerpts
- data from 9 different signers (7 female and 2 male)

More details in Chapter 6.3.1.1.