Tampere University

Raisa Mattila

AUTOMATIC RATING SYSTEM FOR UNILATERAL FACIAL PALSY

Faculty of Information Technology and Communication Sciences
Master of Science Thesis
October 2019

ABSTRACT

TAMPERE UNIVERSITY
Degree Programme in Information Technology, MSc (Tech)
RAISA MATTILA: Automatic Rating System for Unilateral Facial Palsy
Master of Science Thesis, 134 pages, 52 Appendix pages
October 2019
Major: Data Engineering and Machine Learning
Examiners: Assoc.Prof. Heikki Huttunen, Asst.Prof. Antti Vehkaoja, and University Lecturer Terhi Kilamo
Keywords: Automatic Rating System, Facial Palsy, Capacitive Measurement, Advanced Analytics, Validated Learning

The purpose of this thesis is to investigate if an existing capacitive prototype could provide an automatic rating system to grade facial palsy level. The need for an objective rating system for facial palsy severity grading has been acknowledged for decades. The main problem in the existing solutions is the subjectivity, and the usage of a myriad of grading systems. These two factors hinder the scientific development as the comparison of results is difficult.

The evaluated parameters are determined based on a literature review on how humans perceive and detect facial asymmetry. Validated learning is applied to improve and steer the approach as moving from measuring 20 healthy test participants to study 17 facial palsy patients. The analysis compares the corresponding contralateral channels, and concentrates on investigating the waveform similarity and delay. Two movements are used; smile and eyebrow lift. The results gained from the patients are then compared against a reference scale, Sunnybrook, that is commonly used in clinics. The measured data correlated with the Sunnybrook values in waveform from weak to moderate (smile) and moderate (eyebrow lift) extent, and in delay from non-existent (smile) and weak to moderate (eyebrow lift) level. These results alone do not provide a strong signal to continue with the approach as such.

The system is applicable in principle. However, the prototype has a constant limitation of not being capable to assess static symmetry; the sensors measure the change in capacitance and thus only movement is detected. Additionally, the sensor adjustment is manual and thus introduces a subjective error to the measurement. Also, the prototype limits the measurement to a scarce amount of points. These problems could be potentially solved by an improved prototype version that would have head-support, automated adjustment, and more sensors. Also other approaches should be considered. Those involve convolutional neural networks (CNNs), automated facial action coding system (FACS), and infrared (IR) cameras. The next step should be carefully considered; the field is already suffering from a myriad of methods, and creating another compromised system does not solve the fundamental need. In the contrary, it adds to the issue.

TIIVISTELMÄ

TAMPEREEN YLIOPISTO
Tietotekniikan DI-tutkinto-ohjelma
RAISA MATTILA: Automaattinen normijärjestelmä unilateraaliseen kasvohalvaukseen
Diplomityö, 134 sivua, 52 liitesivua
Lokakuu 2019
Pääaine: Data Engineering and Machine Learning
Tarkastajat: Assoc.Prof. Heikki Huttunen, Asst.Prof. Antti Vehkaoja ja Yliopistonlehtori Terhi Kilamo
Avainsanat: Automaattinen normijärjestelmä, kasvohermohalvaus, kapasitiivinen mittaus, edistynyt analytiikka, vahvistettu oppiminen

Tässä diplomityössä tutkitaan, soveltuuko olemassa oleva kapasitiiviseen mittaukseen perustuva prototyyppi kasvohalvauksen asteen mittaamiseen. Tarve kasvohalvauksen asteen objektiiviseen määrittämiseen on tiedostettu vuosikymmeniä sitten. Tällä hetkellä olemassa olevien ratkaisujen keskeisin ongelma on subjektiivisuus ja eri järjestelmien suuri lukumäärä. Nämä kaksi tekijää hidastavat lääketieteellisen tutkimuksen etenemistä, koska tulosten vertailu on hankalaa.

Analyysissä käytettävät parametrit on valittu kirjallisuuteen pohjautuen; kuinka ihmiset havaitsevat kasvojen epäsymmetrian. Vahvistettua oppimista käytetään tutkittavan menetelmän parantamiseen. Aluksi tutkitaan 20 tervettä vapaaehtoista, minkä jälkeen mitataan 17 kasvohalvauspotilasta. Analyysi vertaa vastaavia kontralateraalisia kanavia toisiinsa tutkien aaltomuotoja ja viivettä. Tutkimuksessa käytetään kahta eri kasvojen liikettä: kulmien nostoa ja hymyä. Potilasmittausten tuloksia verrataan yleisesti käytössä olevaan Sunnybrookin asteikkoon. Mitattu potilasdata korreloi aaltomuodon suhteen verrokki-Sunnybrook-arvojen kanssa heikosti tai kohtalaisesti (hymy), ja kohtalaisesti (kulmien nosto). Viiveen suhteen korrelaatio on olematonta (hymy) ja heikkoa tai kohtalaista (kulmien nosto). Tulokset eivät anna vahvaa perustetta jatkaa menetelmän tutkimista sellaisenaan.

Järjestelmä on ainakin periaatteessa käyttökelpoinen, mutta rajoitteita on. Prototyypin pysyvä rajoite on staattisen symmetrian arviointi; anturit mittaavat kapasitanssin muutosta eli vain liike havaitaan. Lisäksi antureiden manuaalinen asettaminen kohdilleen tuo mittaukseen subjektiivisen komponentin ja virhettä tuloksiin. Prototyyppi myös rajoittaa mittauksen muutamaan pisteeseen. Parannettu prototyyppi voisi mahdollisesti ratkaista nämä ongelmat. Esimerkiksi voitaisiin hyödyntää optikon mikroskoopin päätuen kaltaista rakennetta, automatisoida antureiden paikalleen asettaminen ja lisätä antureiden lukumäärää. Myös muita ratkaisuja tulee harkita. Konvoluutioneuroverkot (convolutional neural networks, CNNs), automatisoitu kasvojen yksikköpohjainen luokittelu (facial action coding system, FACS) ja infrapunakamerat (infrared, IR, cameras) tarjoavat vaihtoehtoisia kehityspolkuja kasvohalvauksen asteen automaattiseen ja objektiiviseen määrittämiseen. Seuraava tutkimussuunta tulee harkita tarkkaan, sillä tutkimusalue kärsii lukuisten eri mittaussysteemien käytöstä. Uuden kompromissiratkaisun kehittäminen ei ratkaise perimmäistä tarvetta - päinvastoin, se kasvattaa ongelmaa.

PREFACE

This Master of Science Thesis has been written based on the research done and results gained during employment in the Sensor Technology and Biomeasurements research group in the Faculty of Medicine and Health Technology of Tampere University. The research done is related to a wider MIMEFACE project that also involves doctors from the Hospital District of Helsinki and Uusimaa (HUS) Plastic Surgery Ward.

I wish to thank Emeritus Professor Jukka Lekkala for giving me an opportunity to work in his research group. Also, I am grateful to Doctor of Technology Ville Rantanen for giving me a meaningful topic from his project and guiding a major part of this work. I wish to thank my examiners University Lecturer Terhi Kilamo, Assistant Professor Antti Vehkaoja, and Associate Professor Heikki Huttunen for feedback and useful conversations. Plastic Surgeons Tuija Ylä-Kotola and Sinikka Suominen from HUS deserve special thanks for enabling patient measurements. My current colleague, Doctor of Philosophy Henry Joutsijoki deserves recognition for his feedback. Finally, I am thankful and truly privileged to have supportive and caring family, grandparents, and dear friends as my safety net.

Tampere, 15th October 2019

Raisa Mattila

CONTENTS

1 INTRODUCTION 1
2 MEDICAL BACKGROUND 3
2.1 Facial Palsy 3
2.2 Grading Scales: the Need to Measure 7
2.3 Subjective Grading Scales 8
2.3.1 House-Brackmann Facial Grading Scale 9
2.3.2 Facial Nerve Grading Scale 2.0 12
2.3.3 Sunnybrook Facial Grading System 15
2.3.4 Summary of Subjective Grading Scales 17
2.4 Detection and Perception of Asymmetry 18
2.4.1 Static Threshold 20
2.4.2 Dynamic Threshold 23
2.4.3 Central Triangle 24
3 TECHNICAL BACKGROUND 25
3.1 Excluding 2D Systems 25
3.2 Landmark-based 3D Systems 27
3.2.1 Manual Landmark Systems 29
3.2.2 Automatic Landmark Detection 31
3.2.3 Joint Landmark Criticism 34
3.3 Surface-based 3D Methods 34
3.3.1 Laser Scanning 34
3.3.2 Stereophotogrammetry 37
3.3.3 Infrared Cameras 41
3.4 Other Methods 44
3.4.1 Convolutional Neural Networks to Classify Facial Palsy 44
3.4.2 Facial Action Coding System 47
3.5 Demands for Facial Grading Scale 50
3.6 Face Hugger: Prototype and Software 51
4 MEASUREMENTS AND DATA ANALYSIS 54
4.1 Fundamental Problems with Facial Grading Systems 54
4.2 Our Approach: Overview and Limitations 56
4.3 Using Build-Measure-Learn Feedback Loop 58
4.4 Starting Point 59
4.4.1 Set-up and Protocol 59
4.4.2 Preprocessing 61
4.4.3 Quantities and Parameters 64
4.5 Approach I: Preliminary Experiments 66
4.5.1 Data Collection 66

4.5.2 Data Analysis . . . 68
4.5.3 Preliminary Results . . . 71
4.5.4 Perseverence Point . . . 77
4.6 Approach II: Patient Measurements and Software Development . . . 82
4.6.1 Data Collection II . . . 82
4.6.2 Technology Choices . . . 84
4.6.3 Test-Driven Development and Analysis . . . 86
4.6.4 Pivot Point . . . 89
4.7 Approach III: Patient Data and Python Scripting . . . 90
5 RESULTS . . . 92
5.1 What to Visualize . . . 92
5.2 Vector of Excursion: Waveform vs Sunnybrook . . . 98
5.3 Temporal Difference: Delay vs Sunnybrook . . . 104
6 DISCUSSION . . . 110
6.1 Findings of Results . . . 110
6.2 Limitations and Sources of Error . . . 114
6.3 Future Aspects . . . 119
7 CONCLUSIONS . . . 124
REFERENCES . . . 126

APPENDIX A SUMMARY OF FACIAL ASYMMETRY THRESHOLD STUDIES
APPENDIX B PRELIMINARY STUDY'S TEST PARTICIPANTS
APPENDIX C PRELIMINARY STUDY'S RESULTS
C.1 Smile Results . . . 140
C.2 Eyebrow Lift Results . . . 143
C.3 Results with Medium Intensity . . . 147
APPENDIX D PATIENT STUDY'S PARTICIPANTS
APPENDIX E SOFTWARE SPECIFICATIONS
APPENDIX F USER INTERFACE
F.1 Usability . . . 159
F.2 Graphic User Interface Wireframe Mock-Up . . . 160
F.3 Usability Testing . . . 164
APPENDIX G CLASS DIAGRAM
APPENDIX H ADDITIONAL RESULTS
H.1 Smile . . . 168
H.2 Eyebrow Lift . . . 177

LIST OF TABLES

2.1 The commonness of facial palsy etiology categories from literature review [14]. Several sources have listed idiopathic and infectious palsys to be the number one and two causes for palsy, respectively. The remaining etiology order varies depending on the source. The main purpose of this table is to layout the trend and order of magnitude instead of providing exact widely accepted and proven percentages. Thus, the table should be read for overview and general guidance. 6
2.2 The measurement method of the House-Brackmann grading scale. Table is recreated from [3]. The Measurement-column shows the amount of points available (denominator) and how many points are required for a certain grade (nominator). The Function-column displays the facial performance in percentages. 10
2.3 The description of the House-Brackmann grading scale. Table is recreated from [2]. Each grade is given a description and definition. In the Defined by-column, there is at first a general specification, and explanations for appearance at rest and motion behavior follow. Finally, the secondary symptoms related to the specific grade in question are explained. 11
2.4 Regional assessment of facial function in FNGS 2.0. Each region is assessed independently, and an equivalent score is given. NLF stands for nasolabial fold. This table is reproduced from [20]. However, the description of "Cannot close completely" in Eye-column and on score 4-row was in the Brow-column in the original table. This is believed to be a mistake and thus fixed here. 13
2.5 Global assessment of secondary movement in FNGS 2.0. No points of secondary movement will be addressed if all the regions score 6 in the regional assessment. Table is reproduced from [20]. 14
2.6 Conversion of the regional facial movement and secondary movement scores to final grade in FNGS 2.0. Table is reproduced from [20]. 14
2.7 The summary of introduced scales: the most typical criticism towards them, and their strengths. 18
2.8 The facial subunits' static asymmetry perception thresholds. Table is loosely adapted from [33]. 21
2.9 The facial subunits' dynamic asymmetry thresholds. Table is loosely adapted from [33]. 23
3.1 A summary of the methods of detecting 3D coordinates from 2D facial data such as images or video. 32

4.1 A summary of evaluation methods of each facial grading system demand. The demands marked to be the limitations of the research or method are outside the scope of this thesis. 57
4.2 A summary for preliminary phase's data analysis methods. 70
5.1 A summary for patient data analysis phase's data analysis methods. . . . 94
A.1 A summary of facial subunit asymmetry studies used in this thesis. Table is loosely adapted from [33]. 136
B.1 Necessary details and anonymization of the healthy test participants 1-11. 138
B.2 Necessary details and anonymization of the healthy test participants 12-20. 139
D.1 Notes on prototype fitting on patient test participants 21-37. 153
D.2 Necessary details and anonymization of the patient test participants 21-29. 154
D.3 Necessary details and anonymization of the patient test participants 30-37. 155

LIST OF FIGURES

2.1 The main branches of the facial nerve that innervate the facial muscles are: temporal branch, zygomatic branch, buccal branch, marginal mandibular branch and cervical branch. These major branches are marked to the image and are responsible for specific movements. These branches split further into smaller branches. Figure is from Mackinnon's book [12] in which it is reprinted from www.dnaillustrations.com. 4

2.2 The difference between central (on the left) and peripheral (on the right) facial palsy. The central palsy has not affected forehead or eyes. Instead, the wrinkles of the cheek, nose and the corner of the mouth have straightened out. The person's right mouth corner is hanging, too. The peripheral palsy has affected the person's entire right side and the symptoms can be seen also above eyes. Figure is taken from [1]. 5

2.3 The Sunnybrook facial grading system taken from [30]. Originally represented in [4]. 16

2.4 The three categories of asymmetry and their borderlines, or thresholds, illustrated. There are perceptive and intervention thresholds for facial features at rest (static threshold) and in motion (dynamic threshold). Figure is adapted from [33]. 19

2.5 The relationship between facial progressive asymmetry and the asymmetry detection rate is positively correlated, and has the shape of a sigmoid. The perception threshold is the point of statistically increased rate of perception. Figure is not mathematically accurate: the trend of the behavior is the key. Figure is taken from [33]. 19

3.1 This image shows the anatomical planes on a face: the coronal/frontal XY plane, the tranverse XZ plane, and the sagittal YZ plane [51]. 2D assessment of facial symmetry ignores other planes than the coronal XY plane, and thus the anteriorposterior component is projected to the XY plane and information gets lost [48, 51]. Figure is from [51]. 26

3.2 Some landmarks illustrated; midline and paired landmarks identified on soft tissue and marked with their typical abbreviations. The midline landmarks are: tr (trichion), g (glabella), n (nasion), prn (pronasale), c (columella), sn (subnasale), ls (labiale superius), sto (stomion), li (labiale inferius), sl (sublabiale), pg (pogonion), me (menton). The remaining landmarks are paired, and thus found on both sides, left and right, of the face. The list of landmarks and the image is from [52]. 28

3.3 The laser scanner stereo set-up illustrated. The patient was positioned on a chair and the mirror helped finding the correct spot. The two laser scanners then captured the facial 3D surface simultaneously as a stereo pair resulting in two raw data sets; one from each side, right and left. The merging was completed computationally. Figure is taken from [77]. . 35
3.4 A consecutive laser scanning process illustrated. There were three steps; at first (the uppermost image) the right lateral profile was captured following by frontal profile (the midlle image) recording and finally the left lateral profile (the lowermost image) was scanned. The laser scanner position was untouched during these three consecutive steps and an examiner rotated the chair instead. Figure is taken from [76]. 36
3.5 Passive stereophotogrammetry principle illustrated. The same point of the object being recorded is represented in two images taken from different viewpoints. This enables 3D image formation. Figure is taken from [80]. 37
3.6 The generic deformable mesh model on the left, collected 3D frame in the middle, and the conformed mesh on the right. The conforming was performed for the first frames of each expression sequence. The figure is taken from [81]. 39
3.7 A result of surface symmetry analysis illustrated as a heatmap. The face was divided into three areas that were defined by soft tissue landmarks. The colors represent the difference of corresponding points between the original and mirrored image. The scale ranged from 2.5 mm (blue) difference via 0.0 mm (green) to -2.5 mm (red) difference. The figure is taken from [82]. 39
3.8 Figure a) shows the used structured light pattern and image b) visualizes the pattern projected on a human face target. The distortion of the pattern is visible. Figure is taken from [83]. 40
3.9 The working principle of a TOF sensor illustrated. An IR pulse is emitted by IR projector (E), reflected from a surface point (P), and the reflected IR wave's phase shift in comparison to the emitted signal is measured by sensor (C). Thus, the distance h between the sensor and object point can be solved and a depth image formed. Figure is taken from [87]. 42
3.10 An illustration of the facial palsy classification procedure's major steps. Image is taken from [67]. It is worthy of comment that Sajid et al. [67] used range nulla - V whereas House-Brackmann intended their system to range from I to VI. This systematic shift should be kept in mind when reading further. 45
3.11 The CNN structures responsible for feature learning (C1 and C2) and palsy grading (C3) illustrated. Image is taken from [67]. 46

3.12 The main steps and substeps of 2D automised FACS solutions reviewed in [103]. It is worthy to mention that most feature extraction methods listed in this figure are so called hand-crafted feature extraction methods, when comparing to CNN-based feature extraction (see Subsection 3.4.1). Figure is taken from [103]. 48
3.13 An example result of the AUNets that shows the detected AU-name and its code, and also the probability of its presence. The AUNets was trained to recognize 12 different AUs. Figure is taken from [92]. 49
3.14 The prototype for capacitive distance measurement illustrated. There are 22 channels in total; 11 on both sides. The colored numbers mark the sensor spots. Figure is taken from [7]. 51
4.1 The build-measure-learn feedback loop illustrated. The image is adapted from [106]. 58
4.2 The flow of preprocessing the capacitive data illustrated. The image is adapted from [8]. 62
4.3 Contralateral spatial difference for repetitions of symmetrical smile, test person 13. 72
4.4 Contralateral spatial difference for repetitions of asymmetrical smile, test person 13. 73
4.5 Contralateral spatial difference for repetitions of symmetrical eyebrow lift, test person 4. 73
4.6 Contralateral spatial difference for repetitions of asymmetrical eyebrow lift, test person 4. 74
4.7 Contralateral spatial difference for repetitions of symmetrical eyebrow lift, test person 13. 74
4.8 Contralateral spatial difference for repetitions of supposedly asymmetrical eyebrow lift, test person 13. 75
4.9 Contralateral spatial difference for repetitions of symmetrical eyebrow lift, test person 20. 75
4.10 Contralateral spatial difference for repetitions of inconclusive asymmetrical eyebrow lift, test person 20. 76
4.11 Arithmetic average and its standard error for contralateral spatial difference in percentages computed over 18 healthy test subjects. 77
5.1 Different channel-pair behavior types illustrated. The green background signals healthy side, and the red one the paralyzed side. The smile movement and channel 4 have been chosen to provide an example. 93
5.2 Best smile cross-correlation coefficient averages per channel over all patients. 95
5.3 Best EBL cross-correlation coefficient averages per channel over all patients. 95
5.4 Best smile delay averages per channel over all patients. 96
5.5 Best EBL delay averages per channel over all patients. 96

5.6 Highest cross-correlation coefficients for smile movement over all patients. 99
5.7 Highest cross-correlation coefficients for eyebrow lift movement over all patients. 100
5.8 Pair plot of highest smile cross-correlation coefficients against Sunnybrook values. 101
5.9 Heatmap showing the correlation of prototype results and the Sunnybrook values for smile movement. 102
5.10 Pair plot of highest eyebrow lift cross-correlation coefficients against Sunnybrook values. 103
5.11 Heatmap showing the correlation of prototype results and the Sunnybrook values for eyebrow lift movement. 103
5.12 Contralateral delays for smile movement over all patients. 105
5.13 Contralateral delays for eyebrow lift movement over all patients. 105
5.14 Pair plot of smile delays against Sunnybrook values. 106
5.15 Heatmap showing the correlation of prototype results and the Sunnybrook values for smile lift delay. 107
5.16 Pair plot of eyebrow lift delays against Sunnybrook values. 108
5.17 Heatmap showing the correlation of prototype results and the Sunnybrook values for eyebrow lift delay. 109
C.1 Contralateral spatial difference for repetitions of symmetrical smile, test person 10. 140
C.2 Contralateral spatial difference for repetitions of asymmetrical smile, test person 10. 141
C.3 Contralateral spatial difference for repetitions of symmetrical smile, test person 15. 141
C.4 Contralateral spatial difference for repetitions of asymmetrical smile, test person 15. 142
C.5 Contralateral spatial difference for repetitions of symmetrical smile, test person 19. 142
C.6 Contralateral spatial difference for repetitions of asymmetrical smile, test person 19. 143
C.7 Contralateral spatial difference for repetitions of symmetrical eyebrow lift, test person 1. 144
C.8 Contralateral spatial difference for repetitions of asymmetrical eyebrow lift, test person 1. 144
C.9 Contralateral spatial difference for repetitions of symmetrical eyebrow lift, test person 18. 145
C.10 Contralateral spatial difference for repetitions of supposedly asymmetrical eyebrow lift, test person 18. 145

C.11 Contralateral spatial difference for repetitions of symmetrical eyebrow lift, test person 10. 146
C.12 Contralateral spatial difference for repetitions of inconclusive asymmetrical eyebrow lift, test person 10. 146
C.13 Contralateral spatial difference for repetitions of symmetrical smile at medium intensity, test person 3. 147
C.14 Contralateral spatial difference for repetitions of asymmetrical smile at medium intensity, test person 3. 148
C.15 Contralateral spatial difference for repetitions of symmetrical smile at medium intensity, test person 10. 148
C.16 Contralateral spatial difference for repetitions of asymmetrical smile at medium intensity, test person 10. 149
C.17 Contralateral spatial difference for repetitions of symmetrical eyebrow lift at medium intensity, test person 4. 149
C.18 Contralateral spatial difference for repetitions of successful asymmetrical eyebrow lift at medium intensity, test person 4. 150
C.19 Contralateral spatial difference for repetitions of symmetrical eyebrow lift at medium intensity, test person 6. 150
C.20 Contralateral spatial difference for repetitions of unsuccessful asymmetrical eyebrow lift at medium intensity, test person 6. 151
C.21 Contralateral spatial difference for repetitions of symmetrical eyebrow lift at medium intensity, test person 7. 151
C.22 Contralateral spatial difference for repetitions of unsuccessful asymmetrical eyebrow lift at medium intensity, test person 7. 152
F.1 GUI wireframe image: mode selection. 160
F.2 GUI wireframe image: configuring measurement settings. Channel choosing and Sunnybrook result filling open their own pop-up windows. 161
F.3 GUI wireframe image: connecting to the prototype (step 1), testing channel functioning (step 2), and measuring offset data (optional step 3). The buttons are enabled once the previous ones are pressed and successfully completed. 161
F.4 GUI wireframe image: measuring. There are two views; one for the person conducting the measurement, and one for the patient. 162
F.5 GUI wireframe image: results. The details are not illustrated. 162
F.6 GUI wireframe image: choosing whether to analyze clinical correlation or a single patient's development. 163
F.7 GUI wireframe image: importing the data into the app for clinical correlation analysis. 163
F.8 GUI wireframe image: importing the data into the app for single patient's development analysis. 164

G.1 The class diagram of the developed application. 166
H.1 All smile repetitions of patient number 12. 169
H.2 All smile repetitions of patient number 2. 169
H.3 All smile repetitions of patient number 7. 170
H.4 All smile repetitions of patient number 11. 170
H.5 All smile repetitions of patient number 8. 171
H.6 All smile repetitions of patient number 5. 171
H.7 All smile repetitions of patient number 10. 172
H.8 All smile repetitions of patient number 9. 172
H.9 All smile repetitions of patient number 3. 173
H.10 All smile repetitions of patient number 16. 173
H.11 All smile repetitions of patient number 13. 174
H.12 All smile repetitions of patient number 1. 174
H.13 All smile repetitions of patient number 6. 175
H.14 All smile repetitions of patient number 4. 175
H.15 All smile repetitions of patient number 14. 176
H.16 All smile repetitions of patient number 15. 176
H.17 All smile repetitions of patient number 17. 177
H.18 All eyebrow lift repetitions of patient number 12. 178
H.19 All eyebrow lift repetitions of patient number 2. 178
H.20 All eyebrow lift repetitions of patient number 7. 179
H.21 All eyebrow lift repetitions of patient number 11. 179
H.22 All eyebrow lift repetitions of patient number 8. 180
H.23 All eyebrow lift repetitions of patient number 5. 180
H.24 All eyebrow lift repetitions of patient number 10. 181
H.25 All eyebrow lift repetitions of patient number 9. 181
H.26 All eyebrow lift repetitions of patient number 3. 182
H.27 All eyebrow lift repetitions of patient number 16. 182
H.28 All eyebrow lift repetitions of patient number 13. 183
H.29 All eyebrow lift repetitions of patient number 1. 183
H.30 All eyebrow lift repetitions of patient number 6. 184
H.31 All eyebrow lift repetitions of patient number 4. 184
H.32 All eyebrow lift repetitions of patient number 14. 185
H.33 All eyebrow lift repetitions of patient number 15. 185
H.34 All eyebrow lift repetitions of patient number 17. 186

LIST OF SYMBOLS AND ABBREVIATIONS

Anterior	An anatomical term of location, refers a position of a structure being 'in front of' the body, opposite of posterior
AU	Action Unit
Bilateral	Affecting both sides of the body from the median plane, opposite of unilateral
BML	Build-Measure-Learn, -feedback loop
CNN	Convolutional Neural Network
Distal	An anatomical term of location, farther from a structure's origin, opposite of proximal
DoF	Degree of Freedom
EBL	Eye Brow Lift
FACS	Facial Action Coding System
FGS	Facial Grading Scale
FND	Facial Nerve Disorders
FNGS 2.0	Facial Nerve Grading Scale 2.0, an updated version of House-Brackmann Scale
GUI	Graphical User Interface; instead of using the program from a command line, a graphical interface may be created.
HBGS	House-Brackmann Grading Scale, House-Brackmann Scale
MVP	Minimum Viable Product
PoC	Proof of Concept
Posterior	An anatomical term of location, refers a position of a structure being at 'back' of the body, opposite of anterior
Proximal	An anatomical term of location, closer to a structure's origin, opposite of distal
SBFGS	Sunnybrook Facial Grading Scale, Sunnybrook Scale
Unilateral	Affecting only one side of the body from the median plane, opposite of bilateral

1. INTRODUCTION

Unilateral facial palsy is a condition that affects the seventh cranial nerve, the facial nerve. Its symptoms include one sided and varying level inability to perform facial expressions. Facial palsy has multiple causes and affects 20 to 30 people per 100 000 citizens in Finland annually [1]. The measuring of the facial palsy level is important in monitoring the progress of healing, assessing the need for an intervention and the intervention's influence. Currently, various rating systems are used clinically to evaluate the facial appearance and function; the palsy level. However, the current methods are subjective by nature; they are based on human opinion. This evokes inter alia interobserver variability within the results. Additionally, there is a myriad of scales. These two factors result in hindering the scientific advancements to treat facial palsy as comparison of results and decision making is not straightforward.

This **need for an objective universal facial palsy rating system** has been acknowledged also earlier. The early facial palsy researchers used verbal descriptions and since then the currently clinically used subjective grading scales such as House-Brackmann [2, 3] and Sunnybrook [4] have been developed. In principle, a subjective grading scale offers the examiner a framework to evaluate the palsy level; for example in brief, House-Brackmann scale consists of six levels to choose from. These aforementioned subjective facial grading scales (FGS) are the most common current clinical choices [5]. Another approach to study the human face and its expression is facial action coding system (FACS) that is based on units of muscles. However, there is no method to objectively rate the intensities and thus grade the palsy level, and FACS is labour intensive.

There has been true movement towards objective systems that started with two-dimensional (2D) approaches. The 2D systems have been shown to exclude important facial information in the anteriorposterior direction and thus the current focus is on three-dimensional (3D) and four-dimensional (4D) - time being the fourth dimension - system development to grade the facial palsy. Different data acquisition methods such as landmark- and surface-based approaches have been reported. Both parametrized and neural network based analysis approaches have been described. This thesis utilizes a parametrized method.

Even though the universal objective grading system has been seen as a necessity for decades, the issue remains unsolved. There is no system yet that fullfils the demands set in literature. Thus, the **research question** of this thesis is to study if a capacitive prototype developed in dissertation [6] and specifically in its publications [7–9], could be applied for the facial palsy severity rating system purpose. Methods to answer the research question include measuring healthy test participants in a demo, conducting patient measurements, and analyzing the data. Additionally, there is an **objective** to determine how to develop a software to enable further studies. To achieve the objective, the software's functionality should be specified to be based on the measurement needs, the graphic user interface (GUI) on consulting doctors, and designing careful testing.

This thesis is divided into seven chapters including the introduction. Chapter 2 lays out the medical background to detail facial palsy and the current clinical solutions. In Chapter 3, the technical theoretical context and the basis for the experimental approach are given. Also, the demands for an optimal grading scale from literature are listed. Chapter 4 is the methodology chapter and concentrates on the research methods to answer the research question. The objective is also addressed in Chapter 4. Chapter 5 contains the analysis results. Those results, their relation to the theoretical context and future aspects are discussed in Chapter 6. Finally, conclusions are drawn in Chapter 7.

2. MEDICAL BACKGROUND

The necessary medical background for this thesis is covered in the current chapter. At first, to provide basis for understanding the topic, insight is given on facial palsy in Section 2.1. Section 2.2 overviews the clinical estimation process, and motivates the following sections by laying out the need to measure. Then, Section 2.3 details the current clinical facial palsy evaluation procedures. These are the systems to be replaced. Finally, Section 2.4 describes the fundamental factors behind asymmetry perception to provide the parameters to be measured.

2.1 Facial Palsy

Humans have 12 pairs of cranial nerves that emerge directly from the brain, not from the spinal cord such as spinal nerves. The seventh cranial nerve is the **facial nerve** and is responsible for facial muscle movements. [10] **Facial palsy** is the general name that encompasses its **paresis** (mild) and **paralysis** (complete) forms [10, 11]. The facial nerve has several branches on the face that nerve together all the muscles required in facial expressions [11, 12]. The main branches are visualized in Figure 2.1 [12].

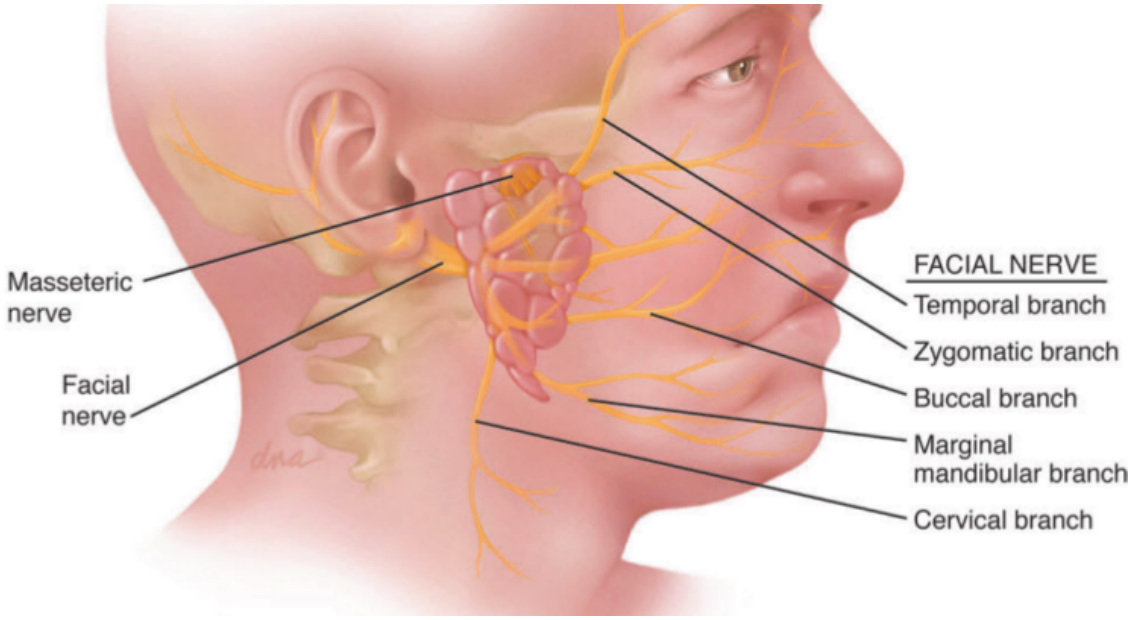

Figure 2.1 *The main branches of the facial nerve that innervate the facial muscles are: temporal branch, zygomatic branch, buccal branch, marginal mandibular branch and cervical branch. These major branches are marked to the image and are responsible for specific movements. These branches split further into smaller branches. Figure is from Mackinnon's book [12] in which it is reprinted from www.dnaillustrations.com.*

Facial palsy covers a wide range of varieties and thus narrowing down for the purposes of this thesis is a necessity. The first distinction is between peripheral and central facial palsy. The type of the palsy is determined by the location of the problem. If the palsy origin is in the brain, the result is called **central facial palsy** whereas **peripheral facial palsy** emanates from a problem in the peripheral facial nerve. [10] The distinction between the two palsy types can be observed by naked eye as the symptoms differ from each other. The border line between the paralyzed and healthy facial areas in peripheral palsy is the vertical mid axis: the paralysis may be either on the right or on the left side but it affects the side's all mimic muscles. In contrast, central facial palsy divides the face also by a horizontal axis: the symptoms are visible only below the eye level on the affected lateral side. [1, 10] This difference between central and peripheral palsy is illustrated in Figure 2.2.

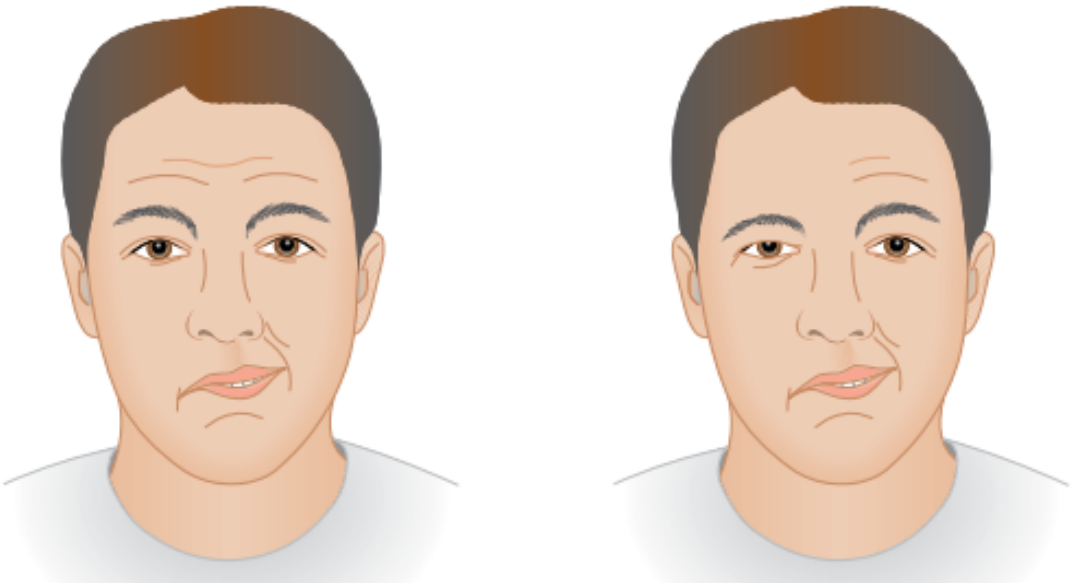

Figure 2.2 *The difference between central (on the left) and peripheral (on the right) facial palsy. The central palsy has not affected forehead or eyes. Instead, the wrinkles of the cheek, nose and the corner of the mouth have straightened out. The person's right mouth corner is hanging, too. The peripheral palsy has affected the person's entire right side and the symptoms can be seen also above eyes. Figure is taken from [1].*

There are also other features to be used to further specify the facial palsy. The peripheral facial palsy may affect only one side of the face and is then called **unilateral** palsy. If the palsy is on both sides, then it is referred to as **bilateral**. [11] However, bilateral palsy is rare [13]. The level of palsy also varies from **partial palsy** (some movement) to **complete palsy** (no movement) between patients and within a patient along time. The peripheral facial palsy may also recur, and would then be called **recurrent** facial palsy, **nonrecurrent** palsy being the alternative. Finally, facial palsy may be described as **acute** or **chronic** depending on the duration of the condition. The typical palsy case is acute, peripheral and unilateral. [11]

In Finland there are annually 20 to 30 facial palsy cases per 100 000 citizens for different reasons; peripheral unilateral facial palsy has several underlying causes, or **etiologies** [1]. Bleicher and Hamiel conducted a literature review and divided the etiologies into five major classifications in [14]. They reported that "*this information was previously untabulated*". Kanerva et al. [1] have gathered etiology information from multiple sources as well, and also books such as [11, 12] listed etiologies. However, those listings are unnecessarily detailed for this thesis, and thus the incidence percentages of facial paralysis based on the study by Bleicher and Hamiel are given in Table 2.1.

Table 2.1 *The commonness of facial palsy etiology categories from literature review [14]. Several sources have listed idiopathic and infectious palsys to be the number one and two causes for palsy, respectively. The remaining etiology order varies depending on the source. The main purpose of this table is to layout the trend and order of magnitude instead of providing exact widely accepted and proven percentages. Thus, the table should be read for overview and general guidance.*

Etiology	Idiopathic	Infectious	Neoplastic	Neurologic	Traumatic
Percentage	49.5 %	15.3 %	13.5 %	13.5 %	8.2 %

The most common etiology is **idiopathic** [1, 11, 12, 14] that refers to unknown origin [1]. Idiopathic facial palsy is commonly called **Bell's palsy** after its describer Sir Charles Bell [15]. The second most common etiology is **infection** [1, 12, 14] such as Lyme disease or varieties of Herpes [1, 14]. The third most frequent etiology category is **neoplastic**, or tumor related palsys. The facial palsys of neoplastic origin are caused in two ways: the tumor itself or the surgical removal of it may cause the palsy. Such tumors include parotid tumor and some ear area tumors. The next most common etiology is the **neurologic etiology** such as cerebrovascular (arteries, vessels, and blood flow in the brain) accidents and several kind of syndromes. Finally, a **traumatic injury** is the fifth most frequent etiology. An injury may come into existence for example during childbirth. Another example is a motor vehicle accident where the temporal bone is fractured. [14]

The origin of peripheral facial palsy may be located in different stages of the facial nerve; four stages with increasing number when moving to more distal parts of the nerve have been listed in [10]. The symptoms vary based on the location and in the most mild form of stage 4 based palsy, the symptoms are bound to the diminished or lost function of the facial mimic muscles [10]. The loss of muscle control can be observed on the affected side also as disappeared wrinkles and folds, drooping mouth corner, and as side effects such as eye irritation due to lack of lubrication as the eye lid can not be closed properly [15].

The more proximally originated palsys of stages 1-3 include other symptoms in addition to mimic muscle palsy [10] often referred to as **secondary symptoms** [1]. Those symptoms include inter alia effects on lacrimal glands [10]. Facial palsy may decrease the tear production but as the eye lid control is diminished, tears may flow freely to the cheek [15]. Also, a syndrome of **crocodile tears** have been reported when involuntary homolateral tearing is related to eating [16]. Some salivary glands' saliva production may decline and the anterior two thirds of the affected side's tongue may lose the sense of taste [10]. The hearing may also become overly sensitive on the

palsy side. That symptom is called hyperacusia. [1] These mentioned symptoms are a consequence of the facial nerve containing parasympathetic fibers to lacrimal and salivary glands [10], being partially responsible for the sense of taste, and branching out to innervate some middle ear muscles [1].

There are also other secondary symptoms worthy of mention. **Synkinesis** is involuntary movement related to voluntary action and may appear also even years after the palsy onset [1]. **Contractures** change for example the eyelid or nasolabial fold depth, or affect the mouth corner position [1]. Finally, **hemifacial spasm** may include twitching, spasms or weakness in the muscles innervated by the seventh cranial nerve. It is a possible sequela to Bell's palsy and may be painful. [10]

The prognosis and treatment vary based on the etiology of the facial palsy. In cases of partial Bell's palsy, 95 % of the patients have been reported to fully recover from the paresis. Even after complete Bell's palsy half of the patients reach full recovery from the paralysis. Most of the Bell's palsy patients get well in three to six weeks, or at least in three to six months. [1] Depending on the etiology, different kind of medication, intervention [1] or finally reconstructive surgeries may be performed [12].

In this thesis, the attention will be on the unilateral peripheral facial nerve palsy independent on the etiology or possible recurrence and will be referred simply as facial palsy.

2.2 Grading Scales: the Need to Measure

The **subjective clinical assessment** of facial palsy consists of two main parts [1]:

1. The facial symmetry and appearance at rest is evaluated [1, 17]. For example, the level of drooping of the corner of the mouth is under assessment [1].
2. The second assessment is to investigate the facial nerve's main branches' differential movement [1, 17]. The movements used to perform the evaluation of the function are: frowning, eyebrow lift, closing the eyes, wrinkling the nose, smiling, and lip puckering [1].

In addition, in the follow-up clinics secondary features such as synkinesis, contractures, spasms, crocodile tears, and other findings related to the condition are assessed. [1]

Clinicians need a tool to conduct the assessment of facial palsy patients' facial status. In more detail, the tool should enable 1) the description of the current functional level, 2) monitoring the development of the palsy over time, and 3) assessing the effects of the treatment. [4] In addition, clinicians require means to exchange information peer-to-peer in order to conduct more effective diagnosis, and evaluation and choosing

of treatments [5, 18]. **Facial Grading Scales (FGS)** have been developed for these purposes [4, 5].

There are numerous different facial grading scales and working principles behind them. The division of the grading systems into categories may also be done in several ways. In [18], the broad categorizing is done into **subjective and objective grading scales**. Moreover, there is further division resulting in four categories, patient centered and multifactorial being the complementing classes. Another taxonomy for classifying the grading scales is given in the same review: division into observational, mathematical, and computer-graphical measurements.

In this thesis, the main division between grading systems is into subjective (Section 2.3) and objective (Chapter 3), and the finer categories are discussed under those mentioned main headings. However, in this thesis, the line between objective and subjective is kept more strict compared to [18]. In [18], for example, a grading scale called Sunnybrook was discussed under clinical objective methods due to application of mathematics despite of the individual scores being based on a clinician's judgement. In this thesis, objective method refers to machine/algorithm based grading where the grading is independent of the human opinion after the system has been implemented for usage. The division used in this thesis is supported, for example, by Mishima and Sugahara in [19].

2.3 Subjective Grading Scales

The fundamental dissection of subjective facial grading scales into finer categories was done in 1983 in a review article by [2]. In the article, the grading scales were divided into three main types: **gross scales**, **regional scales**, and **specific scales**.

The gross, or general, scales postulate that the observer evaluates the facial function and secondary symptoms simultaneously with a single grade. In other words, the face is observed as an entity. Overall, the gross grades form an ordinal scale variable. Therefore the main the purpose of the gross scales is description. [2]

The regional scales require the observer to consider different facial areas separately, and then to combine the separate grades to get the total score. Hence the name; each facial region is evaluated individually. The scores may be either based on verbal description or percentages of normal function. There are two types of regional scales: weighted and unweighted. With a **weighted regional scale**, some areas are considered more important than others. For example, the forehead may be given less weight due to its smaller cosmetic importance. An **unweighted regional scale** on the other hand, does not differentiate the significance of the facial areas. [2]

For the sake of completeness, the specific scale -type is also introduced briefly even though they are not discussed further in this thesis. With specific scales, the observer answers questions about different facial areas with either 'yes' (a point given) or 'no' (no point). The motoric functions and secondary symptoms are evaluated with separate questions. A specific scale may also be weighted. Finally, the separate answers are totaled. [2]

The following Subsections 2.3.1-2.3.3 detail three grading scales: House-Brackmann, Facial Nerve Grading Scale 2.0, and Sunnybrook scale. Subsection 2.3.4 summarizes the considered subjective grading scales.

2.3.1 House-Brackmann Facial Grading Scale

The **Facial Nerve Grading Scale (FNGS)**, often referred as **House-Brackmann grading scale (HBGS)**, was initially introduced by House in 1983 [2], and modified to its final form by House and Brackmann in 1985 [3]. To put the adjustment briefly, a method for measurement was added to the descriptive grading scale. [3]

The House-Brackmann scale was endorsed by Facial Nerve Disorders Committee (FND Committee) of the American Academy of Otolaryngology - Head and Neck Surgery Foundation, in 1985 to be used as the standard grading scale for reporting facial nerve function [3]. In 2009 Vrabec et al. [20] reported that the House-Brackmann scale has indeed become the standard scale. This statement was supported by a survey conducted in 2014 which revealed that by then the House-Brackmann scale was the most commonly used grading scale [5].

The House-Brackmann scale is a six-point gross scale. The scale ranges from Grade 1 (normal) to Grade 6 (total paralysis) and includes the secondary features within. [2] Table 2.2 illustrates the measurement system of the House-Brackmann scale whereas Table 2.3 details the grade descriptions.

The method for measurement is represented in Table 2.2. The observer measures the movement of the patient's paralyzed side. The upwards (superior) movement of the midpoint of the eyebrow, and the outwards (lateral) movement of the oral commissure are measured. One point is assigned for each 0.25 cm movement, 1.0 cm being the maximum displacement for both the eyebrow and oral commissure. Thus, both structures may gain four points in total producing the maximum score of eight. The Measurement-column shows the score, or scores, corresponding the grade in question in the nominator. The denominator represents the maximum score. In the Function-column, the range for the amount of function is given in percentages, and the Estimated function-column displays the percentage of facial nerve function that the paralyzed side should have in theory for the specific grade. [3]

Table 2.2 *The measurement method of the House-Brackmann grading scale. Table is recreated from [3]. The Measurement-column shows the amount of points available (denominator) and how many points are required for a certain grade (nominator). The Function-column displays the facial performance in percentages.*

Grade	**Description**	**Measurement**	**Function %**	**Estimated function %**
I	Normal	8/8	100	100
II	Mild dysfunction	7/8	76-99	80
III	Moderate dysfunction	5/8 - 6/8	51-75	60
IV	Moderately severe dysfunction	3/8 - 4/8	26-50	40
V	Severe dysfunction	1/8 - 2/8	1-25	20
VI	Total paralysis	0/8	0	0

The description of the House-Brackmann scale is given in Table 2.3. The two left-hand columns contain the grades from one to six, and the function level's descriptive name. The right-hand column holds the explanation for the grade in question. The definition of each grade begins with a general description of the grade; for example, for Grade V that is *"Only barely perceptile motion."* [3]. The finer elucidation of the grades then contains information for the observer about rest, motion, and secondary symptoms of the grade, respectively. The motion characteristics are given separately for forehead, eye, and mouth.

The House-Brackmann scale is widely used due to its position as the standard reporting scale, and its simple usage [20]. However, the House-Brackmann scale is also widely criticized on several matters.

Firstly, there are issues related to the measurement method of the House-Brackmann scale: it seems to be completely ignored in practice. Thus the grading in clinics is done simply based on the descriptions of the different grades, in other words based on Table 2.3. [20] Additionally the measuring method has not been verified adequately [20] and an analysis done by Lewis and Adour concluded that the measurements could not be generalized to the descriptive scale [21]. In the same analysis, problematic situations were reported to be arisen, if the healthy side's movement was above the maximum displacement of 1cm. Hence, the paralyzed side could move less than the healthy side, but still gain the maximum score. There is no procedure given by House and Brackmann for such patients [3].

Table 2.3 *The description of the House-Brackmann grading scale. Table is recreated from [2]. Each grade is given a description and definition. In the Defined by-column, there is at first a general specification, and explanations for appearance at rest and motion behavior follow. Finally, the secondary symptoms related to the specific grade in question are explained.*

Grade	Description	Defined by
I	Normal	Normal facial function in all areas
II	Mild dysfunction	Slight weakness noticeable only on close inspection. At rest: normal symmetry and tone. Motion: some to normal movement of forehead; ability to close eye with minimal effort and slight asymmetry; ability to move corners of mouth with maximal effort and slight asymmetry. No synkinesis, contracture, or hemifacial spasm.
III	Moderate dysfunction	Obvious but not disfiguring difference between two sides; no functional impairment; noticeable but not severe synkinesis, contracture, and/or hemifacial spasm. At rest: normal symmetry and tone. Motion: slight to no movement of forehead; ability to close eye with maximum effort and obvious asymmetry. Patients with obvious but not disfiguring synkinesis, contracture, and/or hemifacial spasm are Grade III regardless of degree of motor activity.
IV	Moderately severe dysfunction	Obvious weakness and/or disfiguring asymmetry. At rest: normal symmetry and tone. Motion: no movement of forehead; inability to close eye completely with maximal effort; asymmetrical movement of corners of mouth with maximal effort. Patients with synkinesis, mass action, and/or hemifacial spasm severe enough to interfere with function are Grade 4 regardless of degree of motor activity.
V	Severe dysfunction	Only barely perceptile motion. At rest: possible asymmetry with droop of corner of mouth and decreased or absence of nasal labial fold. Motion: no movement of forehead, incomplete closure of eye and only slight movement of lid with maximal effort; slight movement of corner of mouth. Synkinesis, contracture, and hemifacial spasm usually absent.
VI	Total paralysis	Loss of tone; asymmetry; no motion; no synkinesis, contracture, or hemifacial spasm.

Secondly, Coulson et al. [22] and Murty et al. [23] have reported notable interobserver variability in House-Brackmann grading between trained observers. Further, Murty et al. [23] suggested that the reasons behind the interobserver variability would inter alia include inadequate recording of secondary defects. That brings up the third issue: the House-Brackmann scale admits of several interpretations for secondary movement grading [20].

Finally, the House-Brackmann scale has been criticized due to its gross scale nature. A single global score has been found to be inadequate to describe differential facial function [24–26]. The reason for insufficiency has been explained further: other than grades I or VI (normal and total paralysis) may be difficult to assess if the patients have different levels of function in the upper and lower facial parts [24]. However, it is controversial that which part of the face the global score primarily correlates in unclear cases; in [25] the score was found to mostly reflect the function of the eye whereas in [26] the function of the midface.

Another element of the gross scale criticism is the deficient sensitivity of the House-Brackmann scale. Ross et al. [4] argued that the House-Brackmann scale is too coarse to detect clinically important change. An example of such situation is reported by Gidley et al. [27] who stated that the House-Brackmann scale is inappropriate to be used as a measure of the change after surgical repair of nerves.

This criticism has led to an updated version of the original House-Brackmann scale. The Facial Nerve Grading Scale 2.0 is considered in Subsection 2.3.2.

2.3.2 Facial Nerve Grading Scale 2.0

At the end of the 2000s, the FND committee had debated several years about the response to the criticism towards the recommended de facto grading scale, the House-Brackmann scale. There were three options: to update an existing scale, to adopt a newer scale, or to design a totally new scale. The committee decided that the existing standard House-Brackmann scale should be revised. That decision resulted in the **Facial Nerve Grading Scale 2.0 (FNGS 2.0)**, an unweighted regional scale. [20]

The grading scale revision process incorporated the analysis of newer scales and addressing the common criticism. Regional assessment was added to the updated House-Brackmann. [20] Already in 2003 Yen et al. [25] used a regional assessment of the House-Brackmann scale. Reitzen et al. [26] had then supported their conclusion later in 2009 by reporting that regional House-Brackmann indeed communicates better the state of the facial function of the paralyzed face than the original gross scale. Thus, the updated scale instructs to assess independently the different facial

regions [20]. Additionally, the secondary movement evaluation was separated from the movement analysis to avoid ambiguity. Finally, the measurement score was completely deleted as it neither had been widely adopted nor adequately verified. [20]

The revised scale was validated against the original House-Brackmann scale, and found comparable and to have a satisfactory amount of agreement. Also, the revised scale produces the results in the same format as the original. This is important, as conversion between scales is usually difficult. In brief, the revised scale produces additional information, reduces ambiguity, and maintains to be comparable when juxtaposed with the original scale. [20] Tables 2.4-2.6 represent the FNGS 2.0.

Table 2.4 *Regional assessment of facial function in FNGS 2.0. Each region is assessed independently, and an equivalent score is given. NLF stands for nasolabial fold. This table is reproduced from [20]. However, the description of "Cannot close completely" in Eye-column and on score 4-row was in the Brow-column in the original table. This is believed to be a mistake and thus fixed here.*

Score	**Brow**	**Eye**	**NLF**	**Oral**
1	Normal	Normal	Normal	Normal
2	Slight weakness $> 75\%$ of normal	Slight weakness $> 75\%$ of normal Complete closure with mild effort	Slight weakness $> 75\%$ of normal	Slight weakness $> 75\%$ of normal
3	Obvious weakness $> 50\%$ of normal Resting symmetry	Obvious weakness $> 50\%$ of normal Complete closure with maximal effort	Obvious weakness $> 50\%$ of normal Resting symmetry	Obvious weakness $> 50\%$ of normal Resting symmetry
4	Asymmetry at rest $< 50\%$ of normal	Asymmetry at rest $< 50\%$ of normal Cannot close completely	Asymmetry at rest $< 50\%$ of normal	Asymmetry at rest $< 50\%$ of normal
5	Trace movement	Trace movement	Trace movement	Trace movement
6	No movement	No movement	No movement	No movement

The regional assessment of facial function follows the qualitative scoring system shown in Table 2.4. The leftmost column gives the score, and the other columns give the specifications for that score region-wise. There is one column for each region being assessed: brow, eye, nasolabial fold, and oral commissure. The observer rates the movement of each region separately with a score ranging from 1 to 6. Score 1

corresponds to normal movement, and score 6 to total paralysis. The percentages written in the region columns define the amount of facial function or appearance that should be present when comparing to normal. For example if considering brow region, the movement that is above 50% but below 75% of normal, yields a score of 3. Once the brow region is assessed, the examiner then moves to evaluate the eye, et cetera. This scale is not weighted: each region has an equal effect to the total score. [20]

Table 2.5 *Global assessment of secondary movement in FNGS 2.0. No points of secondary movement will be addressed if all the regions score 6 in the regional assessment. Table is reproduced from [20].*

Score	**Degree of movement**
0	None
1	Slight synkinesis; minimal contracture
2	Obvious synkinesis; mild to moderate contracture
3	Disfiguring synkinesis; severe contracture

The global assessment of secondary movements abides the descriptions shown in Table 2.5. The right-hand column reads the definition, and the left column gives the matching score. However, it is to be noted, that no points of secondary movement will be addressed if all the regions score 6 in the regional assessment; total paralysis lacks synkinesis. Additionally, it is difficult to detect synkinesis in case all the regions have scored 5. [20]

Table 2.6 *Conversion of the regional facial movement and secondary movement scores to final grade in FNGS 2.0. Table is reproduced from [20].*

Grade	I	II	III	IV	V	VI
Total score	4	5-9	10-14	15-19	20-23	24

Finally, Table 2.6 presents the method to convert the four values of regional scores, and a score of secondary movements, into a grade. The five numbers are summed together, and the equivalent grade is read from the table. The left-hand column shows the grade, and the right-hand column the sum or its range. [20] To give a fabricated example, if a patient has obtained scores of 3, 3, 2, and 2 for the regions of brow, eye, nasolabial fold, and oral commissure respectively, and score 1 for synkinesis, the total score is 11 which corresponds to grade III.

Vrabec et al. [20] listed deficits of the revised House-Brackmann scale. For example, the case of surgical nerve repair was neither considered in the design nor did the testing include such patients. They estimated that most surgical interventions of nerve grafting resulted in grade III or IV. A suggestion of adding subcores to those grades was made. [20]

2.3.3 Sunnybrook Facial Grading System

Many facial grading scales have proliferated from the House-Brackmann scale criticism [28]. This section details the **Sunnybrook Facial Grading System (SBFGS)**, or the **Sunnybrook scale**, which is one of these proposed scales.

Ross et al. [4] introduced the Sunnybrook scale in 1996. The Sunnybrook scale is a **weighted regional scale** that consists of three domains: evaluating the resting symmetry, symmetry of voluntary movement, and synkinesis [4, 17]. Each domain is assessed separately either with five movements (evaluation of voluntary movement, evaluation of synkinesis) or by observing three facial regions and their details (resting symmetry examination). Three to five levels are used for description at each domain, and the domains' totals are weighted. Then, a composite score is computed to describe the dynamic facial function, and the static appearance. [4, 29] Figure 2.3 represents the Sunnybrook scale.

Figure 2.3 gives the method to assess the resting symmetry in the leftmost section. The idea is to compare the affected side to the healthy side, and evaluate the resting appearance of the eye, cheek, and mouth one by one. To be more precise, from the eye the palpebral fissure is assessed, from the cheek the nasolabial fold, and from the mouth the oral commissure. There are typical details related to facial palsy listed under every facial area to guide the examiner. From each facial region, a single score from 0 to 2 is chosen. Then, the gained three points, one from each region, are summed up. Finally, the result is multiplied by five which gives the **resting symmetry score**. [4]

The middle column of Figure 2.3 shows the system to examine the symmetry of voluntary motion. The muscle excursion of the paralyzed side is compared to the normal side. There are five standard movements that are examined one at a time: forehead wrinkle, gentle eye closure, open mouth smile, snarl, and lip pucker. These movements evaluate the functioning of the five peripheral facial nerve branches: in other words, the different facial regions. A score, in the range of 1-5, is given for each movement. The **voluntary movement score** is then gained by summing up the five scores, and multiplying the total by four. [4]

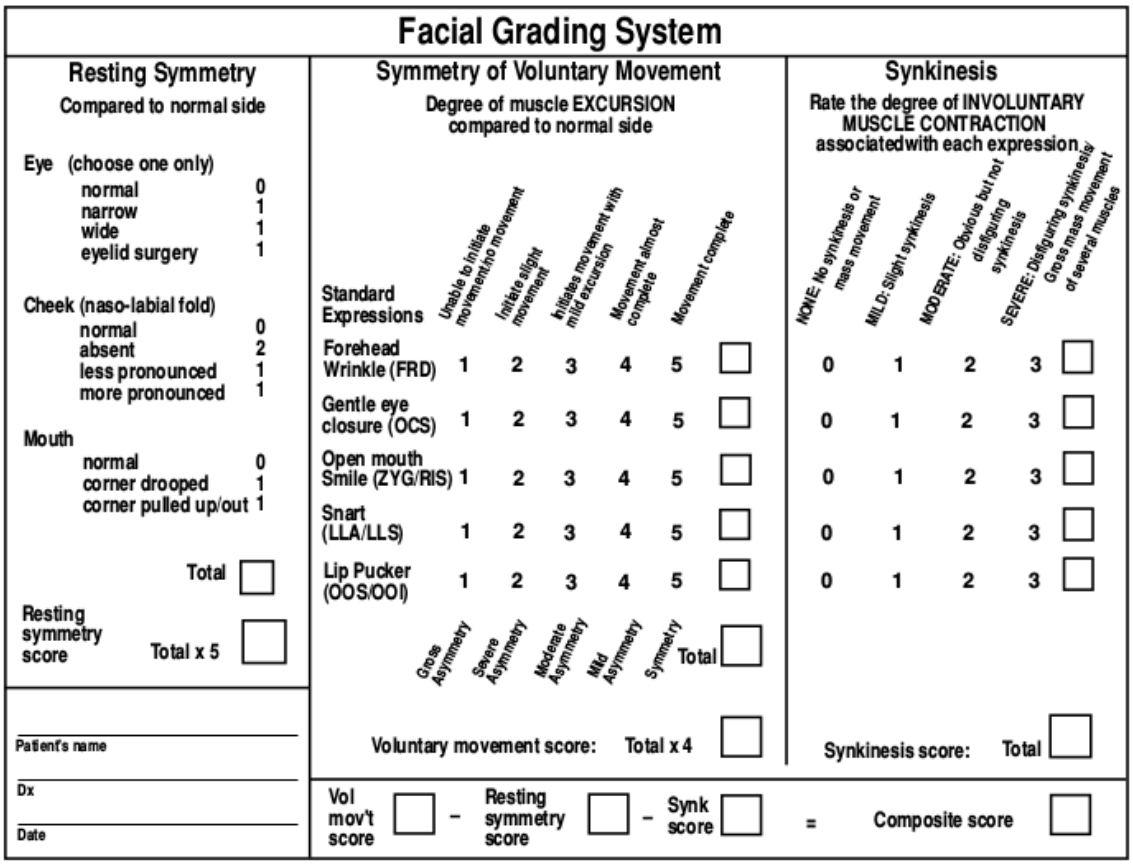

Figure 2.3 *The Sunnybrook facial grading system taken from [30]. Originally represented in [4].*

The evaluation of synkinesis is illustrated on the rightmost column of Figure 2.3. The procedure explained for the voluntary movement assessment is used, but this time the involuntary contraction related to the actual movement is observed and graded with a score from 0 to 3. Thus each facial movement's synkinesis is examined and graded separately. Finally, the **synkinesis score** is computed by adding up the five movements' scores. [4]

The bottom section of Figure 2.3 shows the formation of the **composite score** from each dimension, or the separate scores gained for resting symmetry, symmetry of voluntary motion, and synkinesis. The final score is calculated by subtracting the resting symmetry score and synkinesis score from the voluntary movement score. The individual scores are weighted in a manner that the composite score is between 0 (total paralysis) and 100 (normal). This kind of range mimics the common language. [4]

To conclude, Ross et al. [4] developed a scale that is sensitive to clinical changes; a statement that has been supported by Neely et al. in [29]. This characteristic rectifies a common deficit of the House-Brackmann scale. In more detail, the Sunnybrook scale gives results in more continuous manner and has a larger range than the 6-point

House-Brackmann. Also, the Sunnybrook scale is easy to apply as it does not require much time or any equipment. [4] In general, the Sunnybrook scale has been stated to be a reliable and validated tool for both clinical and research work [4, 29].

However, due to the human-based assessment, some interobserver variability has been reported also in the analysis of the Sunnybrook system by Neely et al. in [29]. In their study, the most significant source for the variability was narrowed down to the mid areas of the composite score range. Especially the voluntary movement evaluation with the expressions of brow raising and lip puckering were located to be the variability origin. Nevertheless, Neely et al. [29] chose the Sunnybrook scale for their multicenter clinical research. Another inconvenience related to the Sunnybrook scale is its inadequacy to measure the status of bilateral paralysis [4]. Fattah et al. [17] concluded that "*no instrument is truly suitable for bilateral facial palsy; a reference for the average normal smile, for example, does not exist*" and thus the insufficiency is not a major drawback.

There is research to support the argument that the Sunnybrook scale should be used instead of the House-Brackmann scale. Firstly, the typical issue of converting results from one scale to another [20] is not apparent with the Sunnybrook scale, as its wide adoption minimizes the need for conversion: Neely et al. [29] argued based on multiple sources such as [30–32] that the Sunnybrook scale is widely used. This is supported by a survey conducted in 2014 that verified the Sunnybrook scale to be the second most used scale after the House-Brackmann scale [5].

Secondly, Fattah et al. [17] and Kanerva et al. [31] suggested that the Sunnybrook system should be used instead of the House-Brackmann until an optimal system has been developed. This encouragement by peers is present also in [30, 32] where the authors conclude that the Sunnybrook scale should indeed be used when sensitive, reliable, and precise facial status documentation is needed.

2.3.4 Summary of Subjective Grading Scales

Three facial grading systems (the House-Brackmann scale, FNGS 2.0, and the Sunnybrook scale) are covered in this thesis in their own Subsections 2.3.1-2.3.3. The two first ones are considered due to their standard-like status originating from the FND committee endorsement, and wide recognition. The latter one was included as a consequence of wide peer-support about its superiority over the House-Brackmann scale and FNGS 2.0. Table 2.7 summarizes the introduced scales.

Table 2.7 *The summary of introduced scales: the most typical criticism towards them, and their strengths.*

Scale	Year	Type	Criticism	Strengths
House-Brackmann [3]	1985	Gross 6-point scale	Not sensitive enough [4], high interobserver variability [22, 23], ambiguous secondary movement grading [20]	Fast global evaluation [24], widespread [5]
FNGS 2.0 [20]	2009	Unweighted regional 6-point scale	Not sensitive enough [20]	FND endorsement [20]
Sunnybrook [4]	1996	Weighted regional 100-point scale	Interobserver variability [29]	Sensitive [29], wide-spread [5]

To conclude, the Sunnybrook scale has been supported to be used when accurate and precise documentation is required, basically when change over time in the facial function is observed. [32] In general, the Sunnybrook scale is viewed as a reliable tool that should be used until an ideal system is presented [17, 31], even though it is affected by interobserver variability [29]. Thus, the Sunnybrook scale will be used as a reference scale later in this thesis.

2.4 Detection and Perception of Asymmetry

Literature identifies the facial asymmetry to be a composite of two different thresholds; the perceptive and intervention. Together these thresholds form the three categories of facial asymmetry: normal variability, acceptable asymmetry, and unacceptable asymmetry. The **perceptive threshold** alies between the normal variability and acceptable asymmetry, and the **intervention threshold** is located between the acceptable and unacceptable asymmetry. These thresholds and the three categories are illustrated in Figure 2.4. The composite nature of facial asymmetry can be understood as different thresholds for face at rest, the **static threshold**, and for the face during movement, the **dynamic threshold**. The static threshold is a spatial limit, and is often expressed in millimeters (mm). The dynamic threshold has inter alia a temporal nature, and is generally communicated in milliseconds (ms). [33]

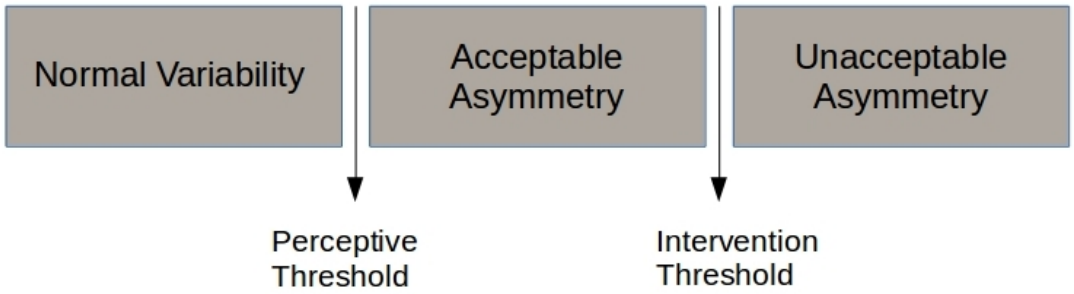

Figure 2.4 *The three categories of asymmetry and their borderlines, or thresholds, illustrated. There are perceptive and intervention thresholds for facial features at rest (static threshold) and in motion (dynamic threshold). Figure is adapted from [33].*

The relationship between progressive asymmetry and the rate of detection is positively correlated and has the shape of a **sigmoid** [33]. Figure 2.5 represents the relation. The purpose of the image is to illustrate the trend, not to be an exact mathematical graph. The x-axis describes the increasing asymmetry of a facial feature, and the y-axis represents the rate of detection. Each facial area has their own unique threshold of perception marked by a sudden statistically significant increase in the perception rate when most of the observers rate the feature asymmetric. Finally, the rate of detection reaches a plauteau phase. [33–35]

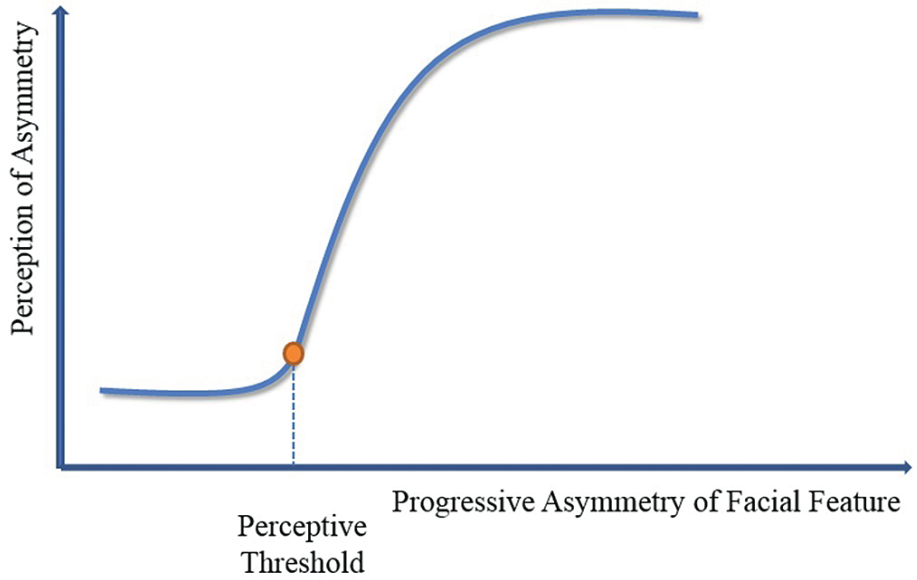

Figure 2.5 *The relationship between facial progressive asymmetry and the asymmetry detection rate is positively correlated, and has the shape of a sigmoid. The perception threshold is the point of statistically increased rate of perception. Figure is not mathematically accurate: the trend of the behavior is the key. Figure is taken from [33].*

There are several factors affecting the perceptive (and intervention) threshold. Three are identified for this thesis. To begin with, a major contributor is the **location**

of the asymmetry: each facial area has its own static and dynamic threshold. [33]. Additionally, the **evaluation time** influences the thresholds; the longer the observation period, the smaller the static threshold [36]. Thirdly, the **observer** may be a clinician [34], informed layperson [34], naive layperson [36], or a facial palsy patient [33]. From the given categories, patients tend to have differing intervention thresholds [33].

The topic of asymmetry thresholds is not as widely studied as subjective grading scales. For the purposes of this thesis, a comprehensive and recent review by Wang et al. [33] is discovered. Additionally, original articles reporting threshold studies are accessed, for example by Chu et al. [36], Kim et al. [37], and Hohman et al. [34]. Table A.1 summarizes the most important research used to compose this section. The purpose of that table is to emphasize the slightly different methodologies, such as different intervals and areas used, to give plausible threshold summary based on them mutually. The referenced studies used progressively increasing level of asymmetry. This type of methodology is required in order to detect the threshold [33].

The following Subsections 2.4.1 and 2.4.2 detail the static and dynamic aspects of facial asymmetry. In other words, the thresholds of different facial units at rest and in motion are reviewed. Subsection 2.4.3 gives insight to the humans' general scan path and visual information sampling frequency of the brain.

2.4.1 Static Threshold

There is a static asymmetry threshold of perception for each facial subunit [33]. These thresholds are summarized in Table 2.8. The table represents the static asymmetry thresholds for different facial areas. The leftmost column gives the facial area, the column in the middle the threshold, and the rightmost column the reference. As shown in the uppermost row, Hohman et al. [34] found the perceptive **static threshold for eyelid closure** asymmetry to be 2 mm. At that threshold, only 10 % of the observers perceived the eye closure to be symmetrical. In contrast, at the previous interval at 1 mm asymmetry, 85 % of the same informed observers experienced the eye closure to be symmetrical. Hohman et al. used 145 observers consisting of both laypeople and clinicians. The model used was a digitally modified female photograph. [34]

The second row in Table 2.8 lays out the perceptive **static threshold for eyebrow** asymmetry; at asymmetry of 3.5 mm 73 % (or 22 people out of 30) of the naive observers perceived the eyebrows to be asymmetrical [36]. When the asymmetry was digitally modified to be an interval less, 3 mm, then only 10 % of the observers stated the eyebrows to be asymmetrical. [36] Chu et al. used a digitally modified

Table 2.8 *The facial subunits' static asymmetry perception thresholds. Table is loosely adapted from [33].*

Facial Area	**Static Asymmetry Threshold**	**Source**
Eye lid	2 mm	[34]
Eyebrow	3.5 mm	[36]
Oral commissure	3 mm	[36]
Nose	4 mm 2.92°	[35, 38] [39]
Chin	6 mm	[35]

photograph of a male as a model in the progressive asymmetry study [36]. Results gained by Hohman et al. [34] support this perception threshold of 3.5 mm. Hohman et al. used an interval of 1 mm and concluded that eyebrow lift with the asymmetry of 3 mm was perceived as symmetrical whereas 4 mm or more as asymmetrical.

In addition to the eyebrow asymmetry threshold, Chu et al. have reported the **static threshold for oral commissure** asymmetry in the same study [36]. The detection threshold is given in Table 2.8 on the third row and was found to be 3 mm by Chu et al. [36]. At that degree of asymmetry 73 % (or 22 people out of 30) of the naive observers identified the oral commissures to be asymmetrical. Furthermore, the asymmetry of 5 mm was found to be unacceptable, or in other words, to be the intervention threshold.

The **static threshold for nasal** asymmetry has been studied at least by Meyer-Marcotty et al. [35], Silva et al. [38], and Kwak et al. [39]. Meyer-Marcotty et al. used a male 3D facial model, and 60 informed clinicians and 30 laypeople as observers. They concluded that the observation threshold of nose asymmetry was 4 mm; at the previous asymmetry interval of 2 mm the nose was still perceived to be symmetrical. The result is supported by Silva et al. who used a 2D digitally modified photograph of a female, and 100 laypeople as observers. Their interval of progressive asymmetry was smaller, 1 mm instead of 2 mm, and the study results the threshold to be the 4 mm. The study conducted by Kwak et al. measured the nasal asymmetry in degrees. Otherwise Kwak et al. used a similar model as Silva et al.; a digitally modified 2D photograph of a female, and such as Meyer-Marcotty et al. beacuse both clinicians (120) and informed laypeople (40) were observing. Kwak

et al. concluded 2.92 degrees to be the perceptive threshold, and the intervention threshold of nasal asymmetry to be 4.28 degrees.

The perceptive **static threshold for chin** asymmetry has been set by Meyer-Marcotty et al. in [35] to be 6 mm. In more detail, Meyer-Marcotty et al. reported that the 6 mm progressive asymmetry was perceived as asymmetrical by the informed observers: clinicians and laypeople. At the previous interval of 4 mm the chin was recognized as symmetrical. This conclusion has been reinforced by Naini et al. [40] that deviation of 5 mm was largely imperceptible. Naini et al. used female and male 2D models, and laypeople, clinicians, and patients as observers. On the other hand, their progressive interval was relatively big when compared to other studies; the interval was set to be 5 mm. However, another study conducted by Silva et al. [38] used smaller incerements of 1 mm and reported that the observers detected no asymmetry in deviations less than 6 mm.

Thus, the location of the asymmetry affects the static threshold [33] in the following way; the eye lid had the smallest threshold of 2 mm [34], and the eyebrow and oral commissure have the second most smallest thresholds of 3.5 mm and 3 mm, respectively [36]. Finally, the nose tip and chin have larger thresholds of 4 mm [35, 38] or 2.92° [39], and 6 mm [35], respectively. However, for example Wang et al. [33] suggest that further studies are needed to establish the static perception and static intervention thresholds.

In addition to the location of the asymmetry, the **evaluation time** affects the static perceptive threshold, too. To be more precise, the evaluation time influences the asymmetry perception in two ways: the detection threshold, and the number of detected asymmetries. This has been studied by Chu et al. in [36]. They investigated the detection of progressive asymmetry with two different observation times: 2 s that corresponds to a casual glance of a passer-by, and 10 s which resembles prolonged observation by an interlocutor.

Firstly, in case of a *single* defect of oral commissure or eyebrow, the static perception threshold decreased for both asymmetries as the observation time got longer. In other words, longer observation period causes smaller asymmetry perception thresholds. [36] Secondly, Chu et al. [36] studied the relationship between observation time and *multiple* simultaneous defects. In more detail, the digitally modified 2D photograph had simultaneously an oral commissure and an eyebrow asymmetry. If the naive laypeople had 2 s to observe the photograph, 90% of the viewers detected a single defect. With the 10 s observation time, 77 % (or 23 people out of 30) of the observers perceived both defects. This result was reported to be statistically significant. Thus, the observation time affects the number of perceived asymmetries. Chu et al. also suggested that with 2 s observation time and both studied defects present, the eyebrow asymmetry to be detected before the oral commissure anomaly. The

eyebrow asymmetry was detected first in 60 % (or 18 people out of 30) of the instances. However, this was not statistically proven. [36]

2.4.2 Dynamic Threshold

The **dynamic perceptive threshold** is a more complex concept than the static one, as the dynamic threshold consists of two main components. Not only does the difference between the two facial sides count (expressed in ms), but also the motion that how the excursion was attained. In other words, the dynamic perceptive threshold depends on the time-delay and the **vector of excursion**. [33, 37, 41]

The vector of excursion has even been suggested to be the critical component in perception of normalness [37]; the difference in distance of the resulting expression may vary up to 52 % in smile, if the vector of excursion is symmetric [41]. However, no thresholds or key values were available for vector of excursion, and thus the concentration here is on the temporal nature of the dynamic perceptive threshold. The time-delay between the lateral facial sides has been studied by Kim et al. in [37] and the dynamic thresholds for the five locations, or actually expressions, they studied are given in Table 2.9.

Table 2.9 *The facial subunits' dynamic asymmetry thresholds. Table is loosely adapted from [33].*

Facial Expression	Dynamic Asymmetry Threshold
Blink	33 ms [37]
Slow eyebrow lift	100 ms [37]
Fast eyebrow lift	67 ms [37]
Lip depression	67 ms [37]
Smile	67 ms [37]

Table 2.9 represents the dynamic asymmetry perception thresholds for five different facial expressions. The leftmost column holds the expression, the middle column the dynamic threshold in milliseconds, and the rightmost column has the source. Kim et al. [37] used progressive temporal asymmetries of 33 ms, 67 ms, 100 ms, 133 ms, 200 ms and finally 267 ms, and the observers of the digitally manipulated video were 58 informed laypeople. As summarized in Table 2.9, the eye blink is the most sensitive movement with the detection threshold of 33 ms. The fast eyebrow lift, lip depression, and smile have the perception threshold of 67 ms. The slow eyebrow lift

has the biggest detective border at 100 ms. These given thresholds provide at least a proper indication of the dynamic asymmetry perception thresholds. [37] However, Wang et al. [33] have stated that further studies are still needed to reproduce and conclude the dynamic perception and dynamic intervention thresholds.

2.4.3 Central Triangle

The static and dynamic thresholds of asymmetry perception are both smallest for the eyelid and then the eyebrows, followed by increase in the thresholds when moving to nose and mouth, and finally the biggest when considering the peripheral facial areas such as chin. [33] This effect of the location to the thresholds is explained by a concept of the **central triangle**. The central triangle is a scan path where observers focus their gaze on clear facial structures - eyes, nose, and mouth - when observing an unfamiliar face. That is, humans observe faces with a general scan path that forms a triangle on the central facial area. It has been shown by eye-tracking studies that a deformity may deviate the observer from this triangle. Also, the eye-tracking studies explain the threshold trend: the observer's gaze reflects directly the observer's attention. Hence the central areas have lower thresholds than the peripheral ones. [33, 42]

In addition to the central triangle, the trend of the dynamic asymmetry thresholds are further explained by the duration of the movement. The unilateral delays are more perceptive if the duration of the movement is shorter when comparing to slower movements such as eyebrow lift or smiling. In other words, the **speed of the movement** affects the threshold in an inverse manner: the faster the movement, the smaller the dynamic asymmetry perception threshold. [33, 37]

The eye is a discriminative facial structure and, thus, vital in the central triangle [42]. Also, blinking is a dynamic facial expression that is fast and short in duration [37]. Therefore, as Wang et al. summarizes; "*[e]yelid position is the most sensitive static facial asymmetry, with blinking the most perceptive dynamic facial expression.*" [33, 36, 37] The dynamic threshold of blink was found to be 33 ms [37] and static threshold 2 mm [34]. Interestingly, Rodriguez and Mendez [43] have reported that visual tasks are processed in the human brain in a discontinuous manner, in quantals of 30-40 ms. In other words, human brain has an update rate, or sampling frequency, for the visual information and that is 30-40 ms. This sets the lower limit in milliseconds for the dynamic asymmetry perception [44].

3. TECHNICAL BACKGROUND

This chapter provides the technical theoretical context for this thesis. Section 3.1 begins the chapter covering why 2D systems are eliminated from further discussion. The following Sections 3.2-3.4 then detail the existing technical solutions to measure facial expression and palsy level. These are the competing solutions to the one examined in this thesis. To clarify, the current technical approaches have been divided by the analysis methodology and there may be overlaps; for example, landmarks may be detected from a surface model. Obviously, there is a myriad of technical suggestions to provide measuring means for facial palsy. Thus, a general overview and common and relevant, or recent and promising, examples are given in this chapter rather than all-encompassing examination.

Finally, the last two sections provide the basis for this thesis' experimental approach. Section 3.5 lists the requirements for optimal grading system based on literature. Section 3.6 introduces the prototype and the available program for measurements.

3.1 Excluding 2D Systems

The development of objective facial measurement systems began with the 2D systems. The 2D systems can be divided into **photograph- and video-based systems**. [45] Within those categories, there are various techniques used to analyze the facial movements [46]. In general, the photographic methods typically study the movement of specific facial locations or landmarks. The distance between the landmark at repose and then at the extreme expression is commonly of interest. [41, 45, 46] The symmetry result is then gained by comparing the affected side's motion result to the healthy side's, for example by computing a percentage [46]. Some methods use a specific facial measure to standardize the photography measurements [45, 46]. The photographic method cannot provide the path of motion [41, 47], a limitation that the video-based methods do not possess; video-based systems can produce both the direction and the amplitude of the landmark motion [41].

However, literature identifies fundamental problems related to the 2D assessment of facial palsy [46, 48–50]. Gross et al. [48] compared 3D and 2D amplitudes of facial

landmark motions. They reported a type of an error, the **projection error**, that is specifically related to the 2D nature of facial palsy measurement. The projection error arises from the three-dimensional facial movement being projected to a plane; in facial motion there is a significant **anteriorposterior component** and the 2D systems project that component into the frontal plane. [48] The anteriorposterior component and the anatomical planes are further illustrated in Figure 3.1. The projection error caused the 3D amplitudes to be notably larger than the 2D as the 2D measurement method lost a major part of the anteriorposterior data [48].

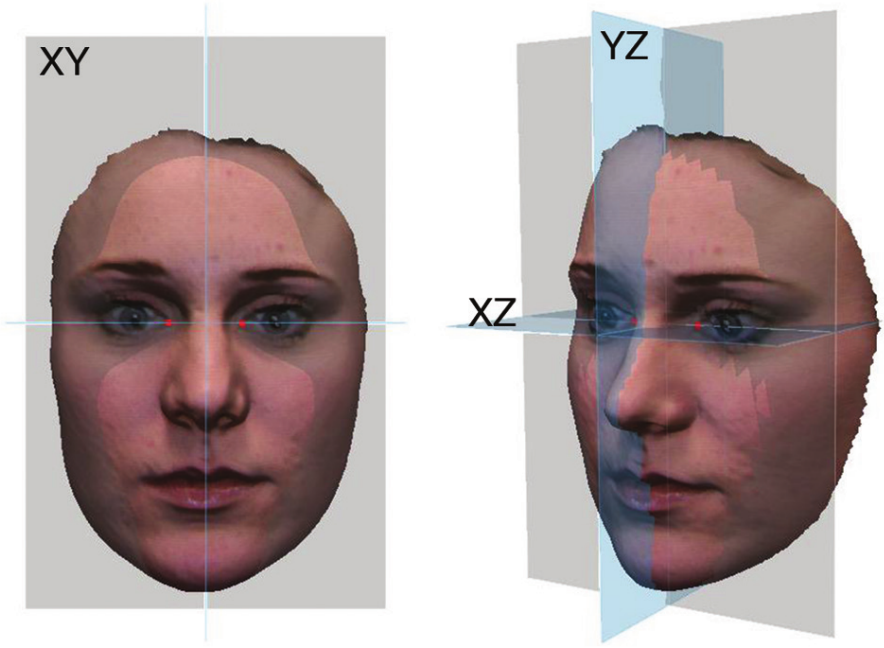

Figure 3.1 *This image shows the anatomical planes on a face: the coronal/frontal XY plane, the tranverse XZ plane, and the sagittal YZ plane [51]. 2D assessment of facial symmetry ignores other planes than the coronal XY plane, and thus the anteriorposterior component is projected to the XY plane and information gets lost [48, 51]. Figure is from [51].*

More specifically, Gross et al. [48] described the projection error to affect all animations (smile, lip-purse, grimace, eye closure, and cheek-puff) and the used fifteen facial landmarks. When the discrepancies were measured in percentages, they were noted to be large. Especially the landmarks of the lower face were influenced: the amplitudinal differences between the 2D and 3D methods were reported to be as much as 43 % during smiling. This result for lower face was explained by two factors. Firstly, the difference between the 2D and 3D results was found to be the biggest with large movements such as smile. Secondly, smile possesses a notable anteriorposterior component as the oral commissures move diagonally backwards during smile. [48]

The result obtained by Gross et al. [48] that 3D systems are superior to 2D has been since supported by peers. Mehta et al. [49] made a proof-of-concept study to

quantify facial movement from a 3D video. They also reported the used 3D system to outperform the 2D system thus supporting the findings of Gross et al. In more detail, Mehta et al. mentioned the possibility to employ different kinds of facial creases and wrinkles, that originate from motion, in the 3D analysis. Finally, Katsumi et al. [50] verified the result and they concluded that the 3D analysis was capable of measuring the mouth's anteriorposterior movement whereas the 2D based analysis could not. An interesting detail is that Katsumi et al. compared the 2D and 3D systems' results to clinical grading scales' results as the clinical scales are commonly used. They summarized that the 3D assessment correlated better, especially in the mouth movement, with the clinical scales than the 2D results did. The clinical grading scales are further introduced in Section 2.3.

To conclude, 3D systems should be preferred over 2D systems in development and usage [48–50]. This prompt has been repeated recently in a comprehensive review by Revenaugh et al. [46]. The fundamental challenge of 2D systems has been reported to be the projection error that causes the anteriorposterior component to be ignored especially in mouth movement [48, 50]. Due to the lost of data, the 2D amplitudes are significantly smaller than the 3D motion amplitudes [48, 50]. Thus, the usage of 2D systems may possibly leave clinically relevant changes undetected [48]. In this thesis, the 2D systems are excluded from further discussion.

3.2 Landmark-based 3D Systems

Facial landmarks can be located purely on soft tissue, on skeletal points, or on functionally important loci of ligaments and muscle attachments. They are specific keypoints of the human face that have a spatial definition and a name that reflects the biological location. Within the average form of the same homologous species, a landmark should have the same position between individuals. In classical biological studies the landmarks are further divided into three categories: type I anatomic landmarks, type II geometrically defined landmarks, and type III extremal landmarks. [52] However, for the purposes of this thesis the different types are rather a curiosity, and no distinction between the types is later made. The amount and configuration of the landmarks used depends on the application and/or the database used [53]. Some typical landmarks are illustrated in Figure 3.2.

According to Katina et al. [54], the standard reference to the detailed definitions of the landmarks has been Farkas et al. [55] since their book's publication in 1994. However, Farkas et al. gave most of the definitions in 2D perspective [54, 55]. After the book's publication, the 3D imaging has developed significantly and thus Katina et al. [54] proposed a new technique and set of landmark definitions in 3D in 2016. They used anatomical curves to identify landmarks and concluded that most of the

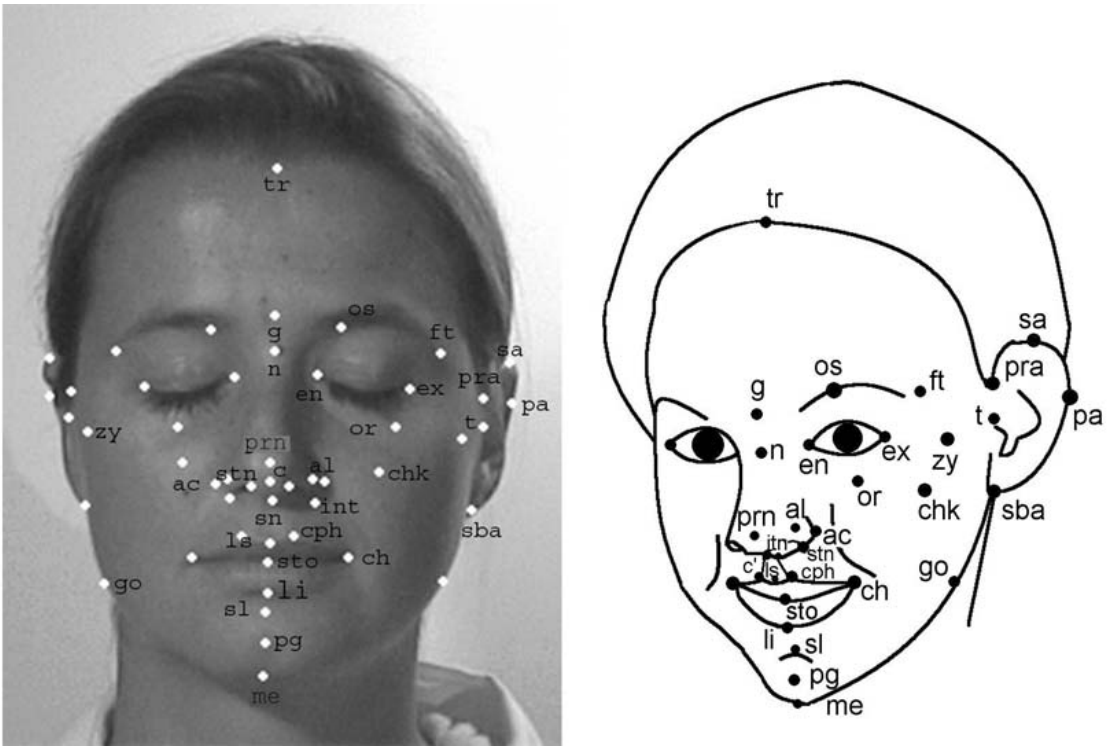

Figure 3.2 *Some landmarks illustrated; midline and paired landmarks identified on soft tissue and marked with their typical abbreviations. The midline landmarks are: tr (trichion), g (glabella), n (nasion), prn (pronasale), c (columella), sn (subnasale), ls (labiale superius), sto (stomion), li (labiale inferius), sl (sublabiale), pg (pogonion), me (menton). The remaining landmarks are paired, and thus found on both sides, left and right, of the face. The list of landmarks and the image is from [52].*

landmarks can be defined either based on the maximum curvature or crossing of curves. The used **curve-based approach** was fundamentally different than the previously used **orientation-based approach** [54].

Obtaining the facial landmarks is merely a midstep to further applications where the facial landmarks are used to represent the face [54]; the surface between the landmarks is ignored in the landmark based analysis [52, 56]. Katina et al. [54] defined shape as "*the information that remains after location, orientation, and possibly also scale, have been removed.*" Facial landmarks have allowed the shape in general being studied across individuals statistically due to the well-defined points [54]. Additionally, **derived applications** have arisen [54]: for example in the field of computer vision, the facial landmarks are important in applications such as facial recognition and verification, and facial expression analysis [57–59]. Also, to give an example from the biological perspective, the origins of schizophrenia have been studied with facial landmarks [54, 60]. Finally, in relation to this current thesis, the facial landmarks have been applied in facial surgery assessment and in facial asymmetry evaluation in general [54].

The facial landmark data can be acquired in three different ways:

a) using physical markers on the face [45]

b) manually selecting the landmarks from a digital source [54]

c) automatically detecting the facial points of interest [54]

Depending on the data obtaining method, there are several sources of error related to the landmark positions. These include the intrinsic accuracy of the measuring equipment, the test subject's expression, environmental conditions such as lightning, and the following landmark identification. [54] The manual selection and physical marker methods are covered in Subsection 3.2.1. Subsequently in Subsection 3.2.2 the automatic detection is explained further.

3.2.1 Manual Landmark Systems

There are two types of systems handled under this section: systems that use physical markers, and methods that use manual landmark identification to digital media.

In brief, the **manual identification of landmarks** on digital facial data refers to first obtaining the 3D (or 2D but that is excluded from this thesis) model and then a human/humans determining the landmark positions [49, 50, 61, 62]. Thus, the landmarks are not physical landmarks in this method; they are just coordinates in a 3D model. The 3D model can be build in various ways: for example, Mehta et al. [49] used a real-time video acquisition system called 3D VAS, Toma et al. [61] and Djordjevic et al. [62] utilized two high-resolution laser scanners, and Katsumi et al. [50] employed a 3D image capturing system that combined two laser-scanners and a color charged-coupled device (CCD) camera in the middle.

The selected amount of facial landmarks were then identified manually from the built 3D models. From the mentioned studies, only Toma et al. who were studying reproducibility of the landmark positioning, mentioned the amount of people positioning the landmarks. [49, 50, 61, 62] Toma et al. [61] used two examiners and localized the landmarks on two occasions to produce the intra- and inter-rater reproducibilities. Djordevic et al. [62] reported that their 3D model could be rotated and thus the landmark position observed from multiple directions before finally determining the landmark coordinates. Katsumi et al. [50] mentioned that their system allowed zooming into an interesting area by the factor of one to twenty before deciding the landmark location, and that the landmark registration procedure was relatively fast by taking less than five minutes. Mehta et al. [49] concluded that the system has potential to become a clinical tool to measure facial movement in palsy cases. Katsumi et al. [50] compared their method to conventionally used Yanagihara

and House-Brackmann scales, summarized that the results aligned, and suggested the investigated system to be clinically used.

However, drawbacks have been reported related to the manual identification of landmarks from 3D models [51, 54, 61]. The manual recognition and registration may be time-consuming and especially some landmarks have high inter-rater variability [51, 61]. Katina et al. [54] pointed out that as landmarks are used to describe the facial shape in general, and derived measures such as symmetry analysis is based on the landmark positions, the variation in their manual identification is extremely problematic as the error gets carried forward in the further analysis.

The second system type to be discussed in this current section is the **marker-based system** which refers to an analysis type where physical markers are attached to the patient's face typically on the facial landmark locations. Then, a group of cameras, or **stereophotogrammetry** technique, is used to capture the motion of the markers in 3D. Different amount of markers have been reported to be used, up to 35 pieces, and the marker size has alternated too between 2 mm and 6 mm in diameter. [45]

The marker-based systems can be further divided into two subcategories; the passive and active marker-based systems. In **passive marker-based systems** light is generated in a source that is close to the measuring camera's lens, and the markers on the patient's face then reflect that light to the camera's lens. In other words, the marker is not the source of light: rather, the marker is passive. The camera's threshold may be set to a level to track the reflecting markers but ignore the soft tissue and fabrics. [45] An example of a passive marker-based system is FACIAL CLIMA, an optical tracking method developed by Hontanilla et al. [63] to assess the functional result after facial reanimation surgery. This method uses infra-red cameras and enables measuring the motion trajectory (displacement, velocity, angle, and areas of the key point movement) and comparing the affected side to the healthy side in 3D terms [63]. Revenaugh et al. [46] reported that several authors have used it since, and the measurements have been validated.

In **active marker-based systems** markers are attached to the patient's face as well. The difference to passive markers is that active markers send signal after being activated and therefore are the source of the signal. The upside of active markers is that they may be set to function at individual frequencies thus easening the data processing step. [45]

However, based on literature on the field, marker-based systems possess challenges [45, 47, 51, 64]. Firstly, there is an error introduced in the physical marker placement between different measurement sessions [45]. A reproducibility threshold of 0.5 mm has been suggested: the standard deviation in every dimension should be kept at or under that when measuring the distance changes from the marker centers [45, 65,

66]. On the other hand, it has been reported that only four landmark locations are manually reproducible when observing intraoperator data. Those locations are right and left cheilion (points at labial commissures), labrale superius (the midpoint of upper vermillion line), and left exocanthion (the outer commissure point of the eye) [45, 66]. However, when interobserver deviation was studied, none of the used 24 landmarks were judged to be reproducible in 3D location [66]. Secondly, the physical markers have been proposed to constrain natural spontaneous facial motion [47]. Thirdly, manually placing the markers has been reported as time-consuming and labour intensive [47, 64]. Fourthly and finally, the correct landmark locus may be difficult to estimate as facial palsy causes deformity on the paralyzed side, and thus the landmark loci shift [67].

To summarize, there is a growing body of evidence to support that facial asymmetry should not be measured with marker-based systems. In 2009 Popat et al. [45] reported how development is moving towards marker-free systems whereas Shreve et al. [64] described in 2011 how they are working to eliminate the need of markers, and Katsumi et al. [50] simply stated in 2015 that landmark-based systems have not become an industry standard due to the requirement of markers.

3.2.2 Automatic Landmark Detection

The automatic detection of facial landmarks can be divided into two categories. The knowledge may be presented as empirical rules that have been observed to provide acceptable results, or by building a machine learning (ML) model. [54] This subsection concentrates on the machine learning approach.

Numerous algorithms have been developed to automatically detect the facial landmarks from images and videos. These methods can be divided into three paramount categories that differ in the information utilization: holistic methods, constrained local model (CLM) methods, and regression-based methods that include inter alia the deep-learning techniques. The training data may be gathered in controlled conditions or in-the-wild when facial expressions, head poses, facial occlusion, or envinronmental conditions such as illumination are not controlled. [53] It has been reported in 2018 by Wu et al. [53] that the current facial landmark research concentrates on landmark detection in-the-wild, although a robust method is still to be developed. Recently, regression-based deep learning methods, Convolutional Neural Network (CNN) models that follow direct regression have been showing the highest performance in landmark detection and tracking. [53]

However, the mentioned categories do not include any 3D methods, the importance of having 3D data was discussed in Section 3.1. There are factors that hinder 3D

landmark detection development. Firstly, there is a lack of 3D face scan databases. This is partially related to the fact that the 3D landmarking is new compared to 2D landmark detection, but 3D face scans are also harder to obtain than 2D images or video. Secondly, 3D landmark labelling is typically more difficult than 2D labelling. However, several different 3D landmark detection methods have been reported and those methods can be classified based on the dimensions of the used data. There are methods that utilize 2D images and videos, and 3D methods that operate on 3D face scans; both method types yield three-dimensional landmark coordinates. [53] In order to obtain 3D landmark coordinates from two-dimensional videos and images, the current methods either utilize the limited 3D training data available [53, 68], or use a pre-trained facial shape model [53, 69, 70], and unite them with the ML approach. Table 3.1 lists the methods of obtaining 3D coordinates from 2D input data and those methods are explained further after the table.

Table 3.1 *A summary of the methods of detecting 3D coordinates from 2D facial data such as images or video.*

Method	**Source**
3D training data	[68]
Pre-trained shape model	[69, 70]
3D training data and pre-trained shape model	[71]

Tulyakov and Sebe [68] estimated 3D facial shape from two-dimensional images by using **limited 3D training data** to predict the necessary depth information. The novelty of the method was the single-step nature: the 3D information was used in the learning pipeline to produce 3D feature indexing. Thus, as the depth was taken into consideration at every level of the cascade, there was no need for further phases to gain the depth. Another contribution worth mentioning is that Tulyakov and Sebe estimated all landmarks under consideration even in the cases of self-occlusion instead of ignoring the landmark or taking a nearest visible point as a replacement. [68]

Another approach to yield 3D facial landmark coordinates from two-dimensional images or videos is to use the limited 3D data to build a **3D shape model**, detect 2D landmarks, and then fit those estimated landmarks on the 3D shape model to gain the 3D facial landmarks [53]. An example of this method is provided by Gou et al. [69] and Jeni at al. [70]. Gou et al. suggested in 2016 a novel 3D face alignment method that used a 3D deformable shape model. They had two steps; at first to detect the 2D landmarks with regression, and then to estimate the correct 3D shape

based on the initial deformable model and the obtained landmarks [53, 69]. Jeni et al. on the other hand used a dense model and thus their procedural step after gaining the 2D landmarks was to iterate the 2D coordinates to fit the pre-trained rigid model [53, 70].

However, there are also methods that use both of the aforementioned methods in the same approach [53]. In other words, the **pre-trained shape model and the usage of 3D data in the training** step have been reported to be combined by Jourabloo et al. [53, 71]. Jourabloo et al. [71] developed a novel method to estimate 3D landmarks and their visibilities in a pose-invariant manner. The training of the cascaded regressor-based model included building a 3D point distribution model from an available set of 3D scans (method of using a shape model), and using a 2D image with manually labelled 2D landmarks combined with the 3D ground truth [71].

Finally, also methods that localize 3D coordinates for facial landmarks from **3D scans** have been reported [53, 72, 73]. Papazov et al. [72] developed a method that first detects location candidates for every landmark under investigation from facial 3D scans. To be more precise, Papazov et al. used a triangular surface patch (TSP) descriptors, that find the 3D shape of a triangular facial area, to get the landmark candidates. Then, the final estimates for both landmarks and pose were gained by fitting the candidates on a pre-trained shape model. [53, 72] Liang et al. [73] presented an approach that detects dominant landmarks on 3D facial scan by using the landmarks' particular geometric properties. Their next step was also to match a deformable 3D model to gain the remaining supporting landmarks. [53, 73] Papazov et al. [72] used synthetic 3D images as training data due to availability. They concluded that the method is viewpoint invariant, works in real-time and produces improved pose and location estimates compared to earlier published work.

Wu et al. [53] discusses the current problematics of automatic ML-based landmark detection and tracking. They report that as facial detection in general is a prior step to landmark detection, the landmarking results depend heavily on the face detection accuracy and performance. This is a limitation. Additionally, Wu et al. remind that some landmark detection and tracking algorithms are still computationally expensive. Finally, they note that the focus has been on simply detecting and tracking the landmarks; the dynamic information has not been employed in its full potential. [53] Regarding to facial palsy and facial symmetry analysis, automatic facial landmark detection and tracking are merely a midstep; landmarking is a task to be completed prior to further facial analysis [71].

3.2.3 Joint Landmark Criticism

The landmark-based systems have been stated to limit the analysis to the marked points [47, 51]. Some regions may be underpresented with landmarks and some areas such as nose are too complex to be analyzed with few landmarks, thus resulting in missing clinically important information [51].

In 2015, Alqattan et al. [51] concluded that a landmark-based method may be used to diagnose orthodontic patients' asymmetry as long as the limitations of the method are acknowledged. Nonetheless, they also suggested that more modern technologies could provide much more insight by extending the analysis beyond specific landmarks, thus including the digitally recognized landmark systems to the criticism [51]. This proposition was repeated in 2016 by Katina et al. [54] when they concluded that their method of characterizing the facial landmark from anatomical curves and surfaces in 3D could be used in "*a much richer way than individual point locations*".

3.3 Surface-based 3D Methods

In contrast to point- or landmark-based approaches to quantify facial measures such as symmetry, surface-based analysis methods may be used as well [51]. Several methods can be placed under this title [54] and in this section, insight is given to laser scanning based techniques (Subsection 3.3.1), stereophotogrammetic approach (Subsection 3.3.2), and infra-red (IR) camera systems (Subsection 3.3.3).

3.3.1 Laser Scanning

In general, facial laser scanning is a method to gain 3D surface representation of the face [54]. Similarly to stereophotogrammetry, another surface-based approach discussed in Subsection 3.3.2, the 3D surface construction is just a step in the entire analysis process [62]. The development of 3D acquiring systems such as laser scanning has enabled the usage of richer analysis metrics; facial areas, volumes and 3D to 3D distances [74] in addition to more conventional ones of angles and linear measurements [75]. Laser scanning has been used inter alia to study facial symmetry in orthodontic patients [51] and in healthy adolescents [62] as well as to examine surgery results, anthropology and genetics [74].

Laser scanning process works by directing a beam of light to a patient and constructing the 3D image by the means of triangulation [54]. Two types of laser scanning set-ups have been reported [76]: a simultaneously scanning stereo pair set-up [61, 77] or

a consecutively scanning set-up [76]. The more common set-up is the stereo pair set-up where two cameras are positioned around the patient [61, 76, 77]. Figure 3.3 illustrates an example of stereo pair set-up where the cameras simultanously scan to produce a right scan and a left scan. Those scans are then merged computationally to form a whole face 3D scan. [77] The consecutive set-up uses only one laser scan and is visualized in Figure 3.4. The right and left lateral profiles are captured in turns, not simultaneously, and also a frontal profile is scanned. [76]

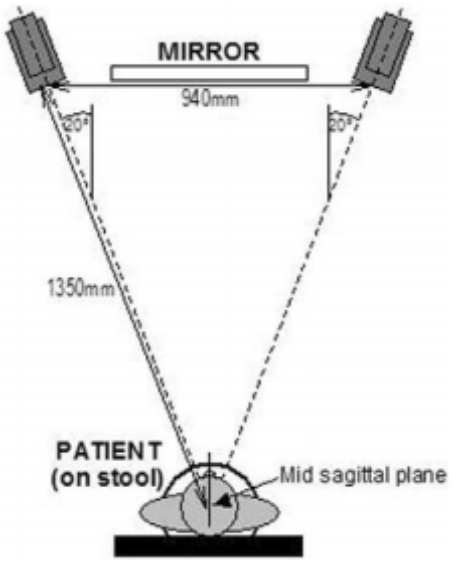

Figure 3.3 *The laser scanner stereo set-up illustrated. The patient was positioned on a chair and the mirror helped finding the correct spot. The two laser scanners then captured the facial 3D surface simultaneously as a stereo pair resulting in two raw data sets; one from each side, right and left. The merging was completed computationally. Figure is taken from [77].*

The main difference between the aforementioned studies is the amount of used laser scanners and the protocol change as a consequence. The consecutive scanning study concluded that linear distances, angles and even surfaces were reliable between the examined consecutive laser scanning method and stereophotography that served as a control [76]. However, volumes and surface registration results should be compared between the surface acquirement methods with caution [76]. The authors also suggested that stereo pair set-up leaves some space for involuntary movement [76].

The high price has been mentioned as a major limitation of laser scanner products and as a hindering factor for their usage [74]. Portable solutions have been suggested to be the economical solution to this problem [74] and also a more convenient option than the traditional static equipment [54]. Portability as a feature is a recent advancement in 3D solutions, and has been proposed to increment the facial databases' contents and application count [75]. However, Gibelli et al. [74] studied the reliability of a low cost portable laser scanner Sense (3DSystems, Rick Hill, SC, USA) and concluded

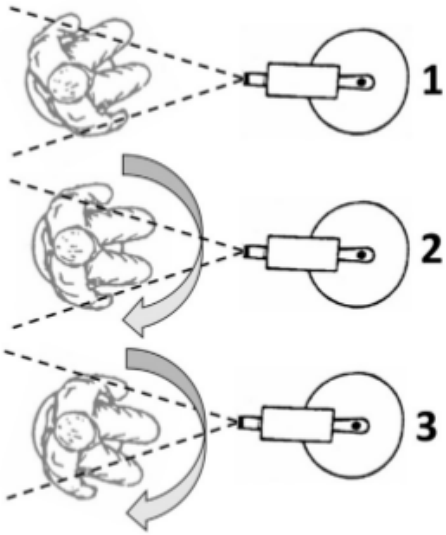

Figure 3.4 *A consecutive laser scanning process illustrated. There were three steps; at first (the uppermost image) the right lateral profile was captured following by frontal profile (the midlle image) recording and finally the left lateral profile (the lowermost image) was scanned. The laser scanner position was untouched during these three consecutive steps and an examiner rotated the chair instead. Figure is taken from [76].*

that the particular low cost scanner was not reliable in living persons but possibly in inanimate object scanning. Thus, the current solution as a portable scanner was not found suitable for facial studies due to the effect of involuntary facial movements on volumes [74].

Beyond the price and portability, laser scanning has also other strengths and weaknessess. The advantage of high resolution has been mentioned at least in [54, 77]. The relatively long capture time when the patient needs to stay still has been listed as a major disadvantage [19, 76, 78]. Involuntary movements during the lengthened scanning may cause motion artefacts [19] or dislocation of landmarks [76, 78]. Mishima et al. [19] summarized that scanning process typically require 8 - 30 s as a basis for their criticism. However, a system used by Gibelli et al. required 2.5 s [76] which is still longer than the human ability to stay still - 1 s has been suggested as a threshold in [79] - or the time typical stereophotogrammetry requires [76, 78]. Finally, the utmost downside of laser scanning in facial studies is that no information between resting and maximum expression is gained; in other words, the approach is static [19]. Nevertheless, to conclude, laser scanning techniques are currently of "*growing importance*" [76].

3.3.2 Stereophotogrammetry

In general stereophotogrammetic method refers to a system that captures an object with at least two cameras in order to produce a 3D reconstruction with a computer. Stereophotogrammetry has three different approaches; active, passive and hybrid. [80] Obtaining the surface representation by stereophotogrammetry is merely an intermediate result, and in the case of facial palsy analysis the gained surface model is then further utilized [66, 81, 82].

Passive stereophotogrammetry refers to measuring approach where the target is recorded from at least two different viewpoints in order to produce a 3D reconstruction with a computer [54]. The passive method may refer to taking photographs from multiple angles [66] or video when also the dynamic facial action is captivated [81]. In the latter case, the system is often referred to as a **4D system** time being the fourth dimension [45, 81]. Figure 3.5 shows the passive stereophotogrammetry principle.

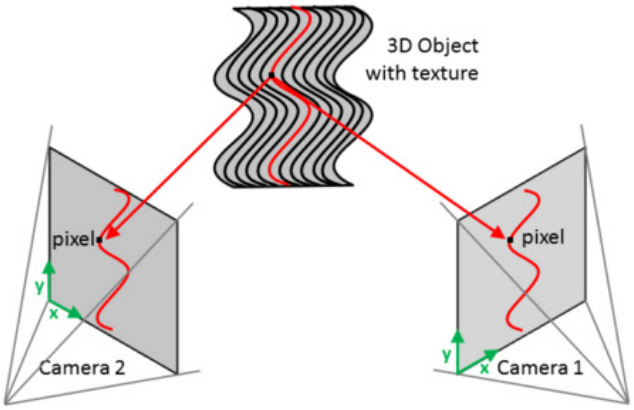

Figure 3.5 *Passive stereophotogrammetry principle illustrated. The same point of the object being recorded is represented in two images taken from different viewpoints. This enables 3D image formation. Figure is taken from [80].*

Passive stereophotogrammetry has been used in facial palsy measurements for example by Alagha et al. [81] and Codari et al. [82]. Both of these studies are recent and promising, and utilize dynamic stereophotogrammetry to study facial palsy. They are overviewed next.

Alagha et al. [81] used a 4D system in their recent unilateral facial palsy assessment. To be more precise, Alagha et al. examined the reproducibility of facial expression

dynamics on 20 facial palsy patients. Five movements (lip purse, cheek puff, maximum eyebrow lift, maximum smile, forceful eye closure) were considered; each movement was conducted in 3-4 seconds taking 60 3D images per second thus resulting in 180 3D images per movement. This procedure was repeated after 30 minutes to gain comparison data. Finally, five key frames (start of motion, first quartile, maximum expression, third quartile, end of motion) from each image sequence was selected and used to compare the two capture sessions movementwise. [81]

The results obtained by Alagha et al. [81] was that three movements of lip purse, cheek puff and eyebrow lift were found reproducible whereas the maximum smile and forceful eye closure not. The authors propose that a possible explanation for the results is the amount of muscles participating in the expressions. The cheek puff, lip purse, and maximum eyebrow lift expressions are created by a limited amount of muscles whereas maximum smile and forceful eye closure are more complex and thus involve varying amount of muscles. Also, the greater amount of used muscles requires more coordination. [81]

In the research paper by Alagha et al. [81], the dynamic 3D facial data was collected by using a commercial system named **Di4D** (Dimensional Imaging Ltd, Glasgow, UK). Once the data was gathered, the analysis began by conforming a generic mathematical mesh to the first frame of each image sequence. The conforming was done by matching a set of 23 landmarks between the mesh and the captured 3D frame. The authors assured that those landmarks have shown to be reliable and were used only for the conforming purpose. Once the mesh was adjusted to its personalized shape, the 42 000 vertices of the mesh were tracked through the expression sequence and thus a dynamic 3D animation, or 4D representation, gained. [81] Figure 3.6 shows the generic mesh (leftmost), 3D frame (in the middle) and conformed mesh (rightmost).

Alagha et al. [81] concluded that the 4D facial palsy assessment utilizing stereophotogrammetry, generic mesh model and advanced geometric morphometrics is a useful method to examine facial dynamics in falsy patients. They also suggested that the approach could be further utilized clinically.

To briefly give another recent example of passive stereophotogrammetry usage in facial palsy studies, Codari et al. [82] utilized stereophotogrammetry to investigate facial symmetry in areawise manner. They used a commercial system as well; **VECTRA M3** stereophotogrammetric system (Canfield Scientific Inc., Fairfield, NJ) captured 3D models of 40 healthy controls and 30 palsy patients and the scanning process was fast. Prior to scanning, Codari et al. marked 50 soft tissue landmarks that were digitized once the 3D model was ready. The authors underlined that the landmarking system followed a validated protocol, and pointed out that the marking was the only manual step of the process. The reason for using physical landmarks was to obtain

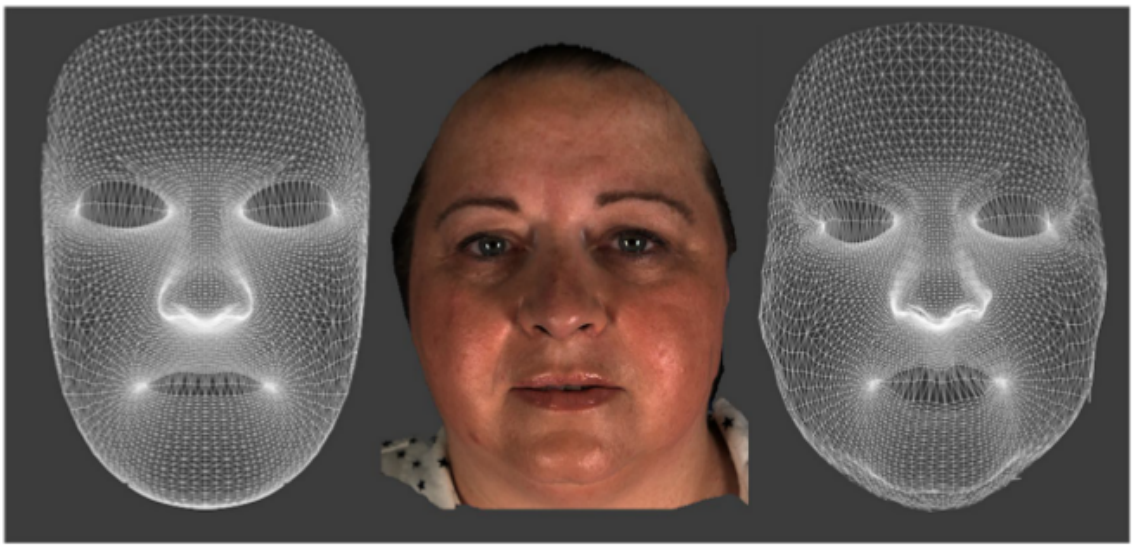

Figure 3.6 *The generic deformable mesh model on the left, collected 3D frame in the middle, and the conformed mesh on the right. The conforming was performed for the first frames of each expression sequence. The figure is taken from [81].*

a single-valued and repeatable definition for the three areas. There was an area for each trigeminal branch of a facial side and the surface area was automatically computed by a commercial software from the same company. The actual facial palsy analysis was then completed by finding the plane of maximum symmetry, mirroring each area, and calculating the root mean square deviation (RMSD) between the corresponding points of the original and mirrored surface areas. An example result image is represented in Figure 3.7. Codari et al. concluded that the method was capable of detecting differences between the control group and the patients, and that approach was suitable for estimating conditions affecting only part of the face. [82]

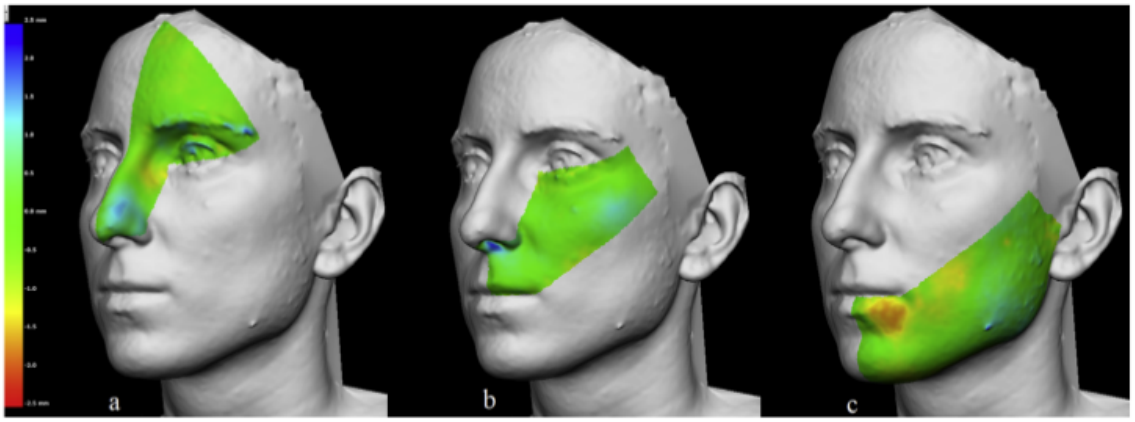

Figure 3.7 *A result of surface symmetry analysis illustrated as a heatmap. The face was divided into three areas that were defined by soft tissue landmarks. The colors represent the difference of corresponding points between the original and mirrored image. The scale ranged from 2.5 mm (blue) difference via 0.0 mm (green) to -2.5 mm (red) difference. The figure is taken from [82].*

The second approach of stereophotograннetry is **active stereophotogrammetry** that is often based on a method called **structured light**. The fundamental idea of structured light approach is to project a light pattern on the target object. As the target, here the human face, has uneven surface, the pattern deformation provides the depth information. In more detail, cameras record the distortion and the object from different viewpoints, and together with system information such as camera distances and positions, the 3D surface can be formed by triangulation. [80] Figure 3.8 shows an example of structured light pattern that is then projected.

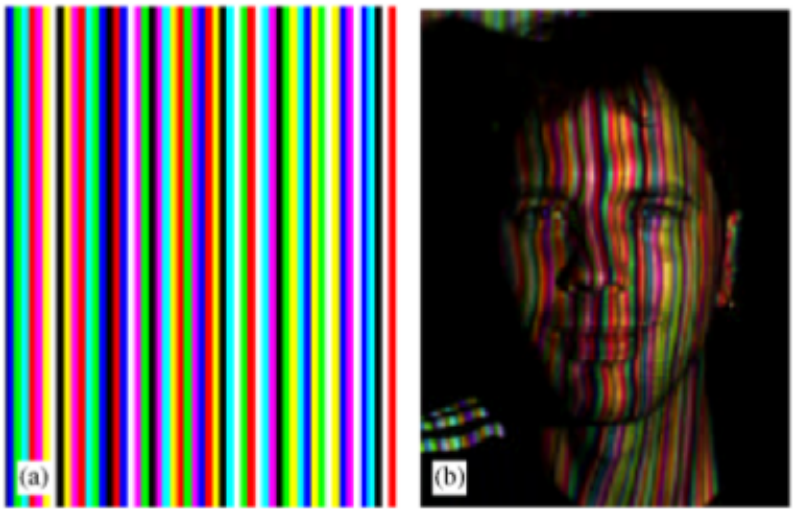

Figure 3.8 *Figure a) shows the used structured light pattern and image b) visualizes the pattern projected on a human face target. The distortion of the pattern is visible. Figure is taken from [83].*

Sandbach et al. [84] reviewed structured light patterns from multiple sources. The bottom line is that the projected structured light pattern can have a **single pattern**, **fringe pattern** or **multiple patterns**, and be a **color pattern** or **binary-coded** [84]. Figure 3.8 shows a single color-patterned light and its application on a face. The chosen pattern type affects the shape acquisition speed and spatial resolution. For example, single color pattern could be used to obtain facial shape in real-time but with the cost on accuracy. Switching quickly between white light and a pattern has been reported to produce a color image and its depth information approximately in sync with each other. [84] Active stereophotogrammetry can also be based on infrared (IR) light [45]. The working principle of such systems is covered in detail in Subsection 3.3.3.

Hybrid stereophotogrammetry systems combine active and passive systems [80]. There are commercial systems such as 3dMD (3dMD, Atlanta, GA, USA) that exploit hybrid stereophotogrammetry to provide 4D systems to examine inter alia the human face [80]. Their face specific system 3dMDface has been utilized for example by Popat et al. in [45]. The system utilizes four gray cameras, two color cameras which are all placed on two pods, and an IR light system. The color cameras capture facial

texture images while the IR system detects the depth information. These passive and active systems record simultaneously. [45] Tzou et al. [80] gathered together several studies that have validated the medical usage of 3dMD. 3dMD system is considered clinically validated and is used as a golden standard reference for 4D measurements in [85]. The accuracy of the system needs to be examined before clinical usage: that was done in [85].

To conclude, non-invasive imaging methods such as stereophotography has sped up the symmetry quantification research [82]. Two recent facial palsy studies utilizing two different commercial stereophotography equipment were introduced here. The advances of stereophotogrammetry systems may be summarized as: non-invasiveness [82], reachable high accuracy if quality cameras used and short capture time [54].

3.3.3 Infrared Cameras

The third method to obtain facial surface to be analyzed are the infrared cameras. IR light is used as a component of active and hybrid stereophotogrammetry systems as explained in Subsection 3.3.2. However, in this current subsection the main emphasis is on IR systems that do not utilize the stereo measuring possibility but use the IR characteristics for depth vision.

Um et al. [86] have described different working principles behind IR-based **mono vision** depth sensors, and two of them are explained here. The first working principle to be described is the **speckle pattern depth measurement**. The working principle of such systems is very similar to the structured light method explained with active stereophotogrammetry but cross-correlation is used to obtain the depth information from the displacement of the recorded IR dots. Also, IR light is used instead of visible light in the projected pattern. Thus, the key components are an IR projector that projects IR speckles and an IR camera to record the projected IR dots. [86, 87] A possible problem related to this speckle pattern approach is a challenging surface geometric that may cause issues in establishing the correspondence between the original emitted speckles and the recorded ones, or between the undistorted and distorted speckles. This would show as holes in the generated point cloud or 3D model. [87]

The second operating principle is the **Time of Flight** (TOF) method. While the TOF principle utilizes the same key components of IR projector and recorder, its measuring principle deviates from the previously mentioned speckle pattern method. TOF system emits an IR wave and detects the reflected IR signal. The phase difference of the leaving and coming wave allows the depth image computation from the speed of light. [86, 87] Figure 3.9 illustrates the TOF principle.

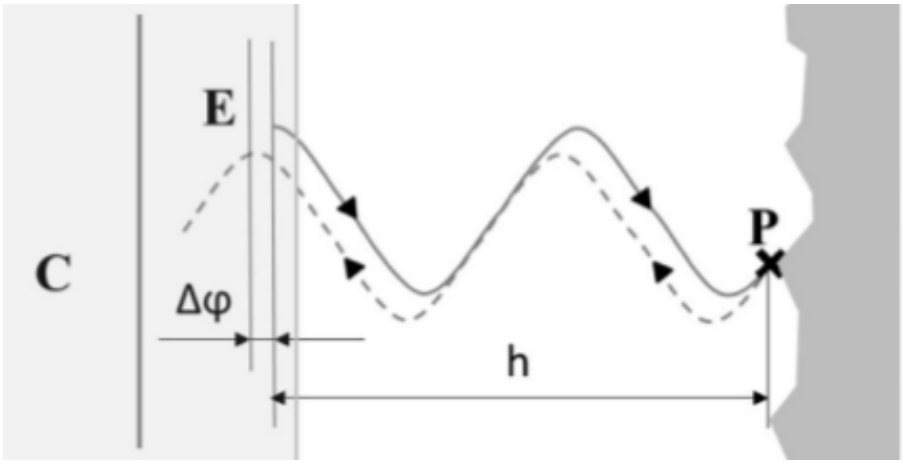

Figure 3.9 *The working principle of a TOF sensor illustrated. An IR pulse is emitted by IR projector (E), reflected from a surface point (P), and the reflected IR wave's phase shift in comparison to the emitted signal is measured by sensor (C). Thus, the distance h between the sensor and object point can be solved and a depth image formed. Figure is taken from [87].*

Figure 3.9 shows the working principle of TOF sensors; the quantities needed for forming the depth image. C stands for IR sensor, E for IR emitter, P for point on the object surface which distance is being measured. The marked quantities are: h which is the unknown distance between the sensor and the object being solved, and ϕ that stands for the phase shift being recorded. From these quantities, the distance h can be solved as the speed of light and IR frequency are known, and the phase shift measured. [86, 87] The weak spot of this approach is certain surface characteristics; sharp edges and highly reflective or transparent surfaces may cause noise, outliers or data shift to the depth measurement. This would appear as blurring or locally distorted 3D image. [87]

An example of an IR based depth sensor is the widely spread off-the-shelf product **Kinect** (Microsoft, USA). There are two versions released, Kinect I and Kinect II, and they work on the structured light/speckle pattern and TOF principles, respectively. [87] In addition to the IR projector and recorder, Kinect cameras use color stream [87] and is thus a hybrid method if following the division made in Subsection 3.3.2. Kinect was originally designed for a video game console but its applicability to health care solutions has been assessed successfully for example by Pöhlmann et al. [87] and by Sathyanarayana et al. [88].

In more detail, the aforementioned studies listed healthcare applications that Kinect has been utilized in. Those included inter alia large scale purposes such as pose and body movement estimation but also usages requiring more detail such as sleep monitoring including breathing rate and depth of sleep detection, heart rate measurements, and also surgical planning and support, to give examples. [87, 88] In addition, Pöhlmann et al. [87] studied the performance of Kinect I and II in terms of accuracy,

precision and resolution of the generated images. The accuracy of both sensors was better than 2 mm in majority of the measurements, and the error describing precision less than 2 mm when the average resolution was at least 390 points in cm^2. Pöhlmann et al. [87] concluded that even though Kinect sensors are not designed for medical purposes they "*show performance adequate for a range of healthcare imaging applications*". A distinction that Kinect I is more suitable for short range imaging and has higher resolution whereas Kinect II is more capable of measuring highly curved surfaces, say face, and has higher precision was made [87]. Kinect II has been used to grade facial palsy for example by Gaber et al. [89]. They utilized Kinect II to assess facial symmetry at rest and in voluntary motion, and used automatically detected 3D coordinates of certain landmarks.

Another portable and inexpensive off-the-shelf IR camera is RealSense F200 (Intel, USA), further on referred to as **RealSense**, designed for home usage, too. Its core components are the same as described for IR depth sensors in general and specifically for Kinect above; an IR sensor and a projector. Again, also with RealSense, a color sensor is included to provide 2D color image alongside with the depth data. The RealSense camera has been developed for close range purposes, such as capturing facial movement, and its recording distance is between 20 and 120 cm. RealSense's working principle is structured light and the composed depth information consists of *x, y, z* point cloud. [85]

Harkel et al. [85] investigated the depth accuracy in a clinical setting for the RealSense sensor. The test subject cohort consisted of 34 facial palsy patients. Five movements from the Sunnybrook scale and a rest pose were used to provide consistency in scientific results. 3dMD system (see Subsection 3.3.2) was utilized as a reference system. The patients' healthy side was used to determine the sensor's accuracy in a normal situation, and the palsy side to investigate if palsy affected the accuracy. Harkel et al. concluded that facial palsy does not affect the sensor's accuracy: the palsy side depth accuracy was found to be 1.48 ± 0.28 mm and healthy side depth accuracy 1.46 ± 0.26 mm. However, the measuring distance did affect the accuracy and the measuring distance of 35 cm was found to provide the highest depth accuracy of 1.07 mm. The authors also stated that the accuracy of RealSense is within the limits of measured Kinect accuracy and may even be higher, as their own study used only a single frame for the analysis. The single frame choice was made to be able to compare the results to the static 3dMD in use. Further on, RealSense accuracy was found to be an order of magnitude poorer than the 3dMD's. In practice this meant a smoother 3D image and areas of high curvature, such as nose region and mouth, had decreased accuracy. However, the decreased accuracy especially around the nose area was suggested to be partially explained by a blocked view; 3dMD system has a stereo view from two pods whereas RealSense captured the images from a single frontal

viewpoint. Harkel et al. concluded that RealSense provides reliable and accurate facial depth information measurement option. [85]

To conclude this section, IR-based depth sensors provide a portable, inexpensive and widely available option for clinical purposes [87] and also for telehealth usage thus not limiting the 4D measurements to health centers [85]. Two examples of such sensors, Kinect and RealSense, are presented above. These portable IR monovision sensors are not as accurate as the traditional 4D imaging systems but this has been stated to be "*an expected result, considering the difference in cost, size, and complexity of the two systems*" [85]. An option to increase the accuracy of IR based depth sensors would be combining two sensors to work in sync in stereo vision sense [85]. For example Mazurek et al. [90] combined two Kinect sensors. However, this multisensor approach introduces the need for calibration and increases the computational costs [88].

3.4 Other Methods

This section details methods that did not fit under the landmark or surface based approaches. To begin with, the convolutional neural networks (CNN) approach to grade facial palsy are described in Subsection 3.4.1. Then Facial Action Coding System (FACS) and especially its automated versions are introduced in Subsection 3.4.2. Generally this section includes recent and automated approaches such as proposed by Liu et al. [91] and Romero et al. [92]. Due to the novelty of the reviewed systems, also 2D approaches are included to this section despite of the discussion in Section 3.1 as the 3D methods have simply not been developed yet.

3.4.1 Convolutional Neural Networks to Classify Facial Palsy

Convolutional neural networks (CNN) have become a central solution method - as mentioned earlier; in landmark detection and tracking [53] - but also in general in image and video recognition [93]. This success is inter alia due to increased computation performance especially based on the usage of graphics processing units (GPU), large open access image databases such as ImageNet, and an annual competition called ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [93, 94].

ML models [95–97] and CNNs [67, 98] have been used to rate facial palsy. However, Sajid et al. [67] referred to [95–98] and reported that to the best of their knowledge, their study is the first one to grade facial paralysis on a large dataset. Sajid et al. used 2000 facial images and then further augmented them to result in 7000 facial

pictures. The used augmentation method produced different levels of paralysis from a single image. The substantial number of images containing varying levels of palsy, addresses the common issue in facial palsy grading; repeatability. Also, in the case of CNNs, adequate training data rejects overfitting. [67]

Additionally, Sajid et al. [67] emphasized their automatic feature selection and its accuracy improving effect to the underlying classification task. In more detail, they juxtaposed the automatic feature selection completed by CNNs and the human-based, or hand-crafted, feature picking (representative studies of that category referred by [67] are for example [95–97]) in their discussion. Their conclusion was that hand-crafted methods may easily choose the wrong features for the task, here facial palsy evaluation, and thus result in poorer accuracy. The rest of this section concentrates on the method developed and represented by Sajid et al. in [67]. This choice is based on the discussion above; their recent method published in 2018 uses a large amount of adequate images and a CNN-based feature selection.

Figure 3.10 summarizes the five facial palsy classification steps of Sajid et al. [67].

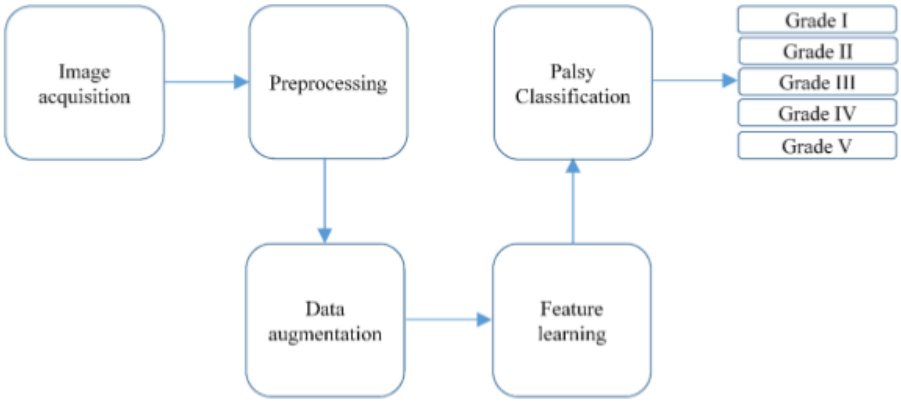

Figure 3.10 *An illustration of the facial palsy classification procedure's major steps. Image is taken from [67]. It is worthy of comment that Sajid et al. [67] used range nulla - V whereas House-Brackmann intended their system to range from I to VI. This systematic shift should be kept in mind when reading further.*

The beginning step marked in Figure 3.10 is **image acquisition**. Sajid et al. [67] acquired 2000 annotated and labelled images from several sources. The labelling utilized the wide-spread House-Brackmann scale and the formed dataset had images of each House-Brackmann palsy level excluding normal. The second step shown in Figure 3.10 is **preprocessing**. In more detail, Sajid et al. rotated all the images to be in an upright-position, converted them into grayscale, completed histogram equalization, cropped to certain dimensions and applied a filter. **Data augmentation** was the third step. The authors' decision to keep the whole pipeline automatic, and thus augment the dataset blindly, limited the choice of augmentation method. A novel method of generative adversial network (GAN), proposed in [99], was selected for

image synthetization. [67] The GAN is based on an adversial procedure; a generator confuses the other model, the discriminator, and tries to maximize the discriminator's errors. The discriminator on the other hand tries to differentiate whether the data came from the original sample or from the generator. [67, 99] The GAN procedure generated every House-Brackmann scale level of facial palsy based on the original picture [67]. The fourth step shown in Figure 3.10 is further illustrated in Figure 3.11.

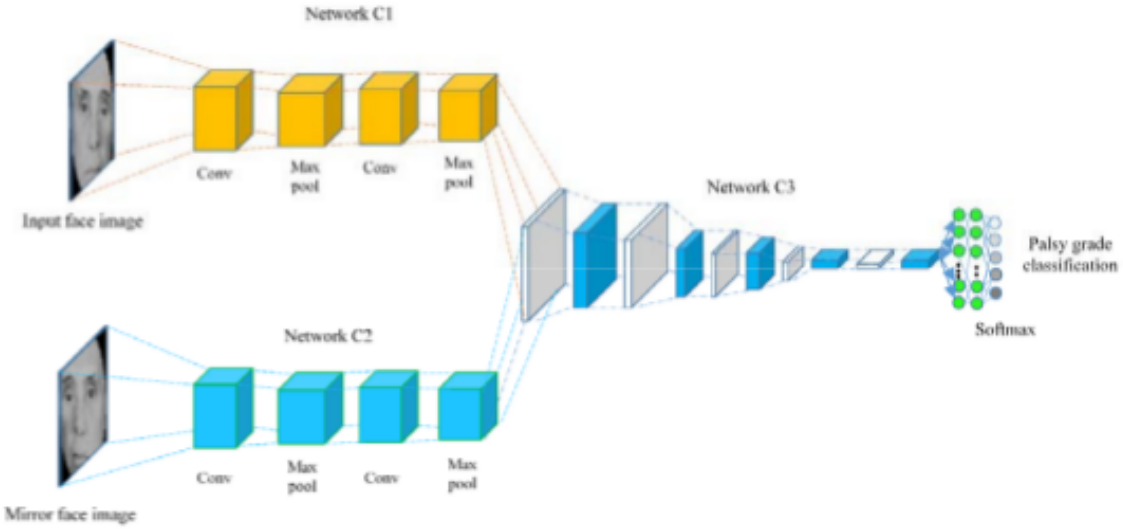

Figure 3.11 *The CNN structures responsible for feature learning (C1 and C2) and palsy grading (C3) illustrated. Image is taken from [67].*

Figure 3.11 lays out the details of the fourth step, the **feature learning**. The keypoint is that there were three neural networks; two first neural networks, labelled as C1 and C2 in the figure, consisted of two convolutional and two max pooling layers, and were responsible for feature encoding. [67] According to the authors, Sajid et al. [67], the C1 and C2 structures have been described by Brachmann and Redies in [100]. Brachmann and Redies proposed a CNN method to measure image symmetry, and Sajid et al. modified the method to suit automatic facial paralysis feature encoding by utilizing two CNNs, C1 and C2, instead of a single CNN. Then, a facial picture and its mirror image were used as inputs to the two CNNs by Sajid et al. in [67].

After the second max pooling, the C1 and C2 outputs were connected to the third CNN denoted by C3 in Figure 3.11. The purpose of this third neural network was to serve as a deep architecture framework for the paralysis classification. [67] The third CNN was a pre-trained network called VGG-16 originally trained with ImageNet pictures [67, 101]. Sajid et al. [67] chose the VGG-16 as their classification framework due to reported excellent performance by [93]. The used VGG-16 had 16 convolutional and 5 pooling layers [67].

The fifth step of Figure 3.10 is **classification**. The softmax layer of the VGG-16 marked in Figure 3.11 was modified by Sajid et al. [67] to contain five neurons; one for each used House-Brackmann palsy grade. Thus, each of the five neurons produced a posterior probability to describe the likelihood of the image belonging to the corresponding House-Brackmann grade. During training, this softmax layer was responsible for adjusting the neuron weights by back propagation. [67]

To be able to produce trustable results, Sajid et al. [67] divided their collected dataset into three subject-exclusive groups at the very beginning; training, validation, and testing subsets. From the 2000 original images, 1000 were extracted for training. After augmentation, the training dataset contained 6000 images with varying palsy levels. The remaining original 1000 images were then split equally for validation (500) and testing (500). The validation or testing data were not augmented. [67]

The results were then expressed in terms of average accuracy percentage, F1 score that describes the accuracy too, precision and sensitivity. The performance of the CNN system was reported to be consistent and superior to existing methods. Thus, Sajid et al. [67] concluded with a suggestion to apply their method in wide-scale for example in health care. To summarize, the developed CNN-based system took an advantage of novel methods developed by others by modifying them; GAN to automize augmentation, pre-trained CNNs to extract features automatically, and VGG-16 to grade them. Sajid et al. acquired large dataset, avoided overfitting, addressed repeatability, and developed scalable narrow artificial intelligence (AI) to House-Brackmann grade facial palsy from 2D images. [67]

3.4.2 Facial Action Coding System

Facial Action Coding System (FACS) was developed in 1970s by Ekman and Friesen and revised by Ekman et al. in 2002 [102, 103], and has been described as "*a taxonomy of human facial expressions*" [103]. FACS divides face into its muscle components, or action units (AU), and describes every facial expression as a combination of these numbered AUs [102]. For example, happiness is expressed with AUs of 6 (cheek raiser), 12 (lip corner puller) and 25 (lips part), and fear may include all AUs of 1 (inner brow raiser), 2 (outer brow raiser), 4 (brow lowerer), 5 (upper lid raiser), 20 (lip stretcher), 25, 26 (jaw drop), and 27 (mouth stretch) [102, 103]. As mentioned, these number coded AUs and their FACS names all have listed muscular basis; for example AU 1 involves muscles called frontalis and pars medialis, and AU 12 a muscle named zygomatic major [102]. An advantage of FACS is that as its components are facial muscle sets, any facial expression can be communicated with FACS just as different combinations of these elemental units called AUs [102].

FACS also enables measurement of intensity, dynamics and symmetry [45, 103] which makes FACS interesting in the sense of facial palsy measurement. The intensity is measured with a five point ordinal scale from A to E; A corresponds to trace of movement and E to maximum action [102]. Dynamics refer to the temporal aspects of facial motion; is the expression beginning or disappearing, how are the duration of different steps of the expression, how is the symmetry of individual AUs and how is the interaction between different AUs [103]. Marquez et al. [103] discussed studies about differences between voluntary and involuntary expressions (in other words, faked and spontanenous expressions) and summarized that dynamics reveal the difference. Thus, FACS is comprehensive [102, 103] and has been used in a variety of psychology and neuroscience studies [103]. However, FACS is very labourious to learn: 100 hours of practice has been reported in order to gain satisfactory skill level to scoring [102], and even after mastering FACS it may demand hours to manually score a one minute video [45].

Thus, automatizing the FACS scoring process would speed up research and make FACS available without the labour-consuming training [103]. It has been suggested that an automated scoring method could also increase the FACS reliability, precision, reproducibility and dynamic measures [103, 104]. Martinez et al. [103] reviewed current solutions of FACS machine analysis. Their survey is summarized in Figure 3.12 which shows the main steps of a generic automatized analysis (pre-processing, feature-extraction, and analysis), and their substeps and/or approach type. The steps are not further elaborated here as the review considered 2D methodology.

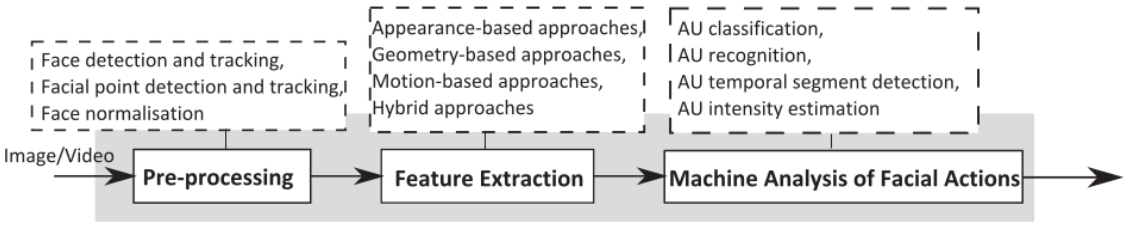

Figure 3.12 *The main steps and substeps of 2D automised FACS solutions reviewed in [103]. It is worthy to mention that most feature extraction methods listed in this figure are so called hand-crafted feature extraction methods, when comparing to CNN-based feature extraction (see Subsection 3.4.1). Figure is taken from [103].*

Martinez et al. [103] concluded that there have been major advancements to obtain automatic (and real-time) FACS based on 2D data but the detection of AU segments and their intensities still remain an open problem. Martinez et al. [103] included central issues to include inter alia occlusion handling, non-frontal head poses, co-occurring speech, varying illumination conditions, and detection and grading of low intensity AUs. Additionally, Martinez et al. [103] higlighted the need for 2D/3D open source databases; realistic and natural in-the-wild database is required to avoid algorithmic local maxima.

Sandbach et al. [84] reviewed 3D methods and concluded then in the same manner that the development of databases is in the central position of transferring the 3D facial analysis from its then infant stage. More specifically, Sandbach et al. mentioned need for spontaneous, natural and dynamic data that would contain complex states. They also suggested that progress in 3D data acquisition should increase the degree and the number of these 3D databases. [84] This urgent requirement has been addressed recently by Liu et al. [91]. Liu et al. [91] proposed an AU-level synthesis framework that was based on a geometric model and an adversial environment. They acknowledged the need for large-scale database in order to employ deep-learning approaches to rate for example AU intensities and thus suggest an augmentation method.

Another recent study was conducted by Romero et al. [92] and diverged from the current state-of-the-art approach of automating facial expression analysis and FACS to be more precise. Romero et al. proposed a novel CNN-based solution to detect AUs from multi-view videos by directly analysing the whole human face. Thus, the standard benchmark of FACS automatisation, landmark localisation, is skipped. The system by Romero et al. is called AUNets and consists of several modules; optical flow computation, ensemble of AU detectors, and a view point classifier. The optical flow field of the input video investigates the shared and individual appearances of AUs, the viewpoint classifier selects suitable AU detectors from the ensemble of them based on the input video's angle, and finally the CNN based detectors conduct presence predictions of the AUs. This method was concluded to be flexible due the modular design, efficient and advance the solution of the problem. However, limitations of the approach were listed to be a big number of parameters and the long time required to train them. [92] A result example is visualized in Figure 3.13.

Figure 3.13 *An example result of the AUNets that shows the detected AU-name and its code, and also the probability of its presence. The AUNets was trained to recognize 12 different AUs. Figure is taken from [92].*

The fact that FACS has a physiological basis [92, 102] easens the computer vision solution development; it is possible to concentrate on the core of the task [92]. Solving automated facial expression analysis task would be a step towards a highlevel human-computer interaction [92], and have a major effect on games, security and health industry [103].

3.5 Demands for Facial Grading Scale

The purpose of facial grading systems is to provide means to document facial nerve functioning, the tracking of recovery due to time and/or intervention, and to enable communication between clinical professionals and researchers [17, 18]. Numerous publications related to facial palsy and facial symmetry have listed demands for FGS; for example Fattah et al. [17] reviewed literature systematically and inter alia collected demands, and Kang et al. [28] reported of the "*generally accepted goals of an ideal facial nerve grading system*". The following list gathers the demands together.

Thus, facial nerve grading system should:

1. be able to provide differential information about the facial nerve's main branches' movements. In other words, provide the means for dynamic measurement. [1, 17, 18, 28]
2. be able to provide information of facial symmetry at rest. In other words, provide the means for static assessment. [1, 17, 18, 28]
3. be able to provide information of secondary symptoms such as synkinesis [1, 17, 18, 28]
4. have high reproducibility which translates to low intra- and interobserver variability [17, 28]
5. be objective [17]
6. be sensitive enough to provide information of the recovery over time and for possible intervention's evaluation [4, 17, 18]
7. be minimally invasive [17]
8. be clinically convenient to use [4, 17, 28]
9. be specific [17]
10. be quantitative [4, 17]
11. be cost effective [17, 28]

12. be fast [4, 17, 28]
13. not require plenty of equipment [4, 28]

The experimental part of this thesis will be compared against these demands.

3.6 Face Hugger: Prototype and Software

An alternative method to detect and measure facial movements and expressions, and their intensities, has been developed by Rantanen et al. in dissertation [6] and in its publications [7–9] to be more precise. At the center of the method is a prototype [6] that is visualized in Figure 3.14 and was casually named as Face Hugger.

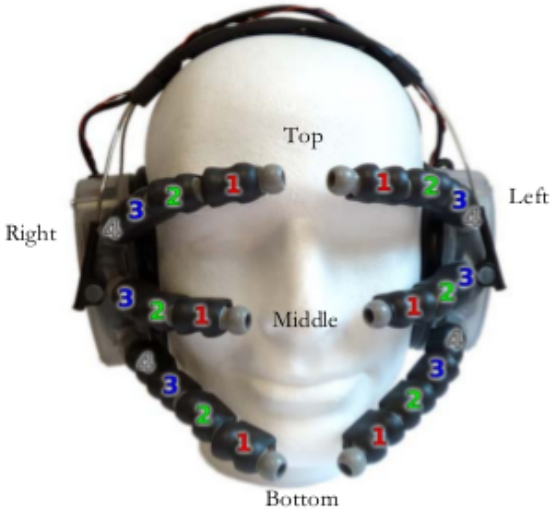

Figure 3.14 *The prototype for capacitive distance measurement illustrated. There are 22 channels in total; 11 on both sides. The colored numbers mark the sensor spots. Figure is taken from [7].*

Face Hugger from Figure 3.14 is a prototype device for capacitive facial measurement. It was build on top of an off the shelf headset, where the earmuffs contain the electronics and support three branches from both sides. The branches are visible in Figure 3.14: top, middle and bottom branches on right and left. The measurement electrodes were placed in these extensions; small pieces connected to each other with ball-and-socket joints hold the sensors. The pieces have three degrees of freedom to provide adjustability. There are 22 pieces, and thus sensors and therefore channels, in total. The colored numbers in Figure 3.14 indicate a sensor position. [6]

Face Hugger uses the measurement principle of **capacitive distance measurement**. There are two plates in the measuring capacitors: one is the sensor with an area of of 2.4 cm^2 and the other the face. Once the person wearing the prototype does facial

expressions the distance between the plates vary thus alternating the measurable capacitance according to Eq. (3.1):

$$C = \epsilon \frac{A}{d}, \tag{3.1}$$

where d stands for distance, A for area, ϵ is a permittivity constant of the medium between the plates, and C is the capacitance. Capacitive touch sensors work utilizing the same concept. The multichannel capacitive raw data is then transferred over Bluetooth connection to a computer. The prototype's maximum sampling frequency of 29 Hz when all the channels are set to measure originated from the electronic choices made in the dissertation [6] process. The capacitive raw data is converted to distance signal by solving Eq. (3.1) for a quantity $d_\propto$ that is proportional to distance:

$$d_\propto = \frac{1}{C} \tag{3.2}$$

where $d_\propto$ is the computed quantity proprotional to the absolute distance, and C the measured capacitance. When comparing Eq. (3.2) to Eq. (3.1), it is worthy of mention that they correspond to each other; the area A and permittivity ϵ are reduced in Eq. (3.2) as they are assumed to stay constant and as absolute distance is not required. After the capacitance samples have been converted into distance proportional values, preprocessing steps such as baseline removal should be conducted before suitable analysis steps. [6]

Face Hugger prototype has been used to **detect and classify facial expressions** [7, 8]. Ten volunteers performed voluntary expressions of frowning, eyebrow lift, eye closure, opening mouth, and raising and lowering mouth corners while wearing the prototype [7, 8]. The analysis utilized principal component analysis (PCA) to locate the facial activity from the multichannel data [7, 8] and hierarchical clustering to identify the movement [8]. Other expressions than eye closure were detected and located correctly with percentages over 90; eyebrow movements were located from the multichannel data successfully in average of 99 % of the cases with standard deviation of 3 %, and lowering of mouth corners reached the success average of 96 % with standard deviation of 13 %. Mouth opening and mouth corner raising had success in 100 % of the locating cases. [7] The eye closure movement was problematic also in classification [8].

Face Hugger has also been utilized in **facial expression intensity measurements** [9]. The test set-up included ten volunteers, and movements of frowning and eyebrow lift which involved frontalis and corrugator supercilii muscles respectively, and lowering and rising the mouth corners that necessitated triangularis and zygomaticus major muscles again, respectively. The movements were repeated ten times with

three different intensity levels (low, medium, high). The control measurement was done with conventional and invasive electromyography (EMG). It was concluded that the prototype is able to record muscle activation intensities well compared to the EMG. The muscles investigated had small differences in error when juxtaposed with EMG excluding triangularis that is responsible for pulling the mouth corners downwards. The differentiation of the three used intensity levels was reported to be straightforward and successful. [9]

Face Hugger will be used in this thesis' experimental part. To conduct the measurements with Face Hugger, an existing software is utilized. The **current software** is an in-house software developed within the research group to be used with Face Hugger. The software runs on Windows operating system and can be used via graphic user interface (GUI). The software receives data over Bluetooth connection. It is a general-purpose program that allows 1) measurement configuration and 2) conducting measurements. The configuration step includes for example choosing the recording channels, setting the sampling frequency, and defining instructions to be shown for the test person and the duration of each instruction. The measurement conduction from the software point of view means showing the configured instructions, and annotating and reading the measured data into a text-file.

The current software was designed to be a general program for Face Hugger and therefore no study-specific special needs were taken into account. There is no data analysis build into the software as the analysis is often survey specific. The GUI allows numerous adjustments in order to be a multipurpose and thus for simple studies it is rather complex. Also, the way the result file is written is not user-friendly; it is difficult to observe by naked eye. The results are also written on a file only after the entire measurement is conducted and the user chooses to save the measurement. Thus, a lost Bluetooth connection forces retake on the entire measurement.

Unfortunately there is no documentation, whitepaper, or reference in article to be cited here. This section's information was gained by using the software and getting oral instructions and demo from the research group members.

4. MEASUREMENTS AND DATA ANALYSIS

This chapter covers the research methods; how are the results gained. To begin with, Section 4.1 ties this experimental part to the theoretical context laying out the motivation to this thesis. Section 4.2 gives a general overview to the used approach. Section 4.3 introduces the used framework, build-measure-learn (BML) feedback loop, and Section 4.4 describes the parts of the methodology that are shared between the feedback loop iterations. Finally, Sections 4.5-4.7 give insight to the iterative BML loops.

4.1 Fundamental Problems with Facial Grading Systems

The speed of scientific progress to treat facial palsy is not as high as possible. This is due to the lack of a proper universal method to measure the facial status. Both before and after the development of the House-Brackmann scale in 1985, there has been a **myriad of methods**. In the 1983 research House compared eight different systems [2]. The weak spots of the House-Brackmann scale led to proliferation of numerous new scales [28] and later on their characteristics and commonness have been studied multiple times, for example in [5, 17, 28].

The bottom line is that **peer-to-peer comparison** is at the heart of scientific achievement. Now it is partially prohibited as the studies express results on multiple grading systems with different ranges. In other words, studies are difficult to compare between each other [5, 105] as the results may be difficult to interpolate from scale to scale [20].

The comparison and conversion issues originating from the wide amount of different scales is not the only factor restraining the medical advancements to treat facial palsy. The need for precision in the outcome results is enhanced when a trial is conducted in multiple locations [29] and required when comparing different trials, too. Within a multicenter trial this is possible to some extent: Neely et al. [29] developed and adopted a specific criteria to be used with the Sunnybrook grading system to overcome the different level of experience and personal divergence of to pay attention to details of the coinvestigators. This criteria was then assimilated by

Neely et al. to control the **intraobserver** and **interobserver variability** of the multicenter study.

However, this kind of control is not completed when comparing different studies and trials. Even though the interobserver variability of several distinct grading scales has been widely studied, little attention is paid to the **comparability of studies** when drawing conclusions based on different trials. What happens to the reliability of the conclusion, if the individual trials' outcome results have been measured by a) same person within a hospital b) different people from the same hospital c) different people from different hospitals d) different people from different countries or e) different people from different continents. The human factor of the measurements seems to be ignored to some extent.

Thus, the fundamental problem with the facial grading systems is that they are subjective or contain a subjective component. **Subjectivity** is where the interobserver (and intraobserver) variability arises from. In more detail, the primary obstacles related to subjectivity are possible bias [19], the effect of varied level of experience and education [29], and decreased sensitivity [19].

This issue has been acknowledged for a long time and already in 1983 House began his article [2] with words "*[t]he major problem in assessing the results of facial nerve surgery or medical treatment lies in the subjective methods of assessment and reporting.*" Trying to improve the scale led to wide amount of scales, but none has fixed the underlying cause. Therefore several facial grading systems were left out of this thesis, even relatively competent ones such as the Yanagihara scale. There was no point for more comprehensive processing of systems embodying the same fundamental flaw and thus to be replaced.

Later on, in the articles of other systems than the House-Brackmann grading scale, the subjectivity problem has been recognized as well. Often an objective method, or "ideal method", has been suggested to be the solution. In other words, the **need for a standardized objective method** has been acknowledged as the solution widely. The grading scales covered in this thesis included FNGS 2.0 which limitation of subjectivity was stated alongside its introduction: "*[f]urther improvement of quantification of facial nerve function will require an objective rating scale.*" [20] The Sunnybrook scale was analyzed by Fattah et al. [17] and Kanerva et al. [31] few years after its development by Ross et al. [4]. The analyses suggested that the Sunnybrook scale should be used until an ideal grading system is available [17, 31]. To conclude, in addition to individual yet peer-reviewed articles, also the 12th International Facial Nerve Symposium held in 2013 emphasized the need for a standardized system to measure and discuss especially the surgical outcomes [18].

Indeed, there have been efforts to develop an objective grading system to solve the issue of subjectivity, and a range of those attempts have been represented in Sections 3.1-3.4. The development towards an objective facial grading system started with 2D systems, and then shifted to 3D systems [45] as 2D approach was proven to be insufficient to grade facial 3D movement [48]. Within the 3D system category the field has been further split into landmark- and surface-based approaches [51]. There have been suggestions in the literature that surface-based marker-free systems should be favored over landmark systems inter alia to utilize the whole information set available instead of limiting to individual points [51, 54]. Also methods outside landmark- and surface-based approaches have arisen, as discussed in Section 3.4. Those systems include very recent research on CNNs and for example the FACS that has physiological basis was automated in [92]. Thus, narrow AI is making its progress also in the close proximity of facial palsy studies.

Based on the literature on the field it is safe to claim that in addition to subjective scales, there is a myriad of aiming-to-be-objective methods as well. However, to the best of knowledge, no method fulfills the demands set for an ideal objective facial grading system, that were gathered together in Section 3.5. Thus the need remains an open question even after over 35 years of House and Brackmann's article's opening sentence.

4.2 Our Approach: Overview and Limitations

Based on Chapter 2 and 3, and discussion in Section 4.1 there is still a need for a facial palsy grading system that fulfills the demands laid out in Section 3.5. The research question of this thesis asks whether the Face Hugger prototype introduced in Section 3.6 could provide an answer for this need. The **novelty** of this current research is the application of a capacitive method to measure facial symmetry. To the best of knowledge, capacitive approach has not been applied to study facial symmetry previously.

To answer the research question, this thesis focuses on demands 1 and 10, the capability to measure the main branches' dynamic movement and to provide quantitative results, respectively. Demands such as minimal invasiveness (demand 7), clinical convenience (demand 8), cost effectiveness (demand 11), fast to use (demand 12), and the requirement not to utilize plenty of equipment (demand 13) will be included to discussion. The **limitation of the Face Hugger method** is the inability to fullfil the demand 2 of static assessment. This limitation arises directly from the measuring principle explained in Section 3.6: the *change* in the capacitance is measured and that change is evoked by *varying* distance between the sensor plate and the facial skin. When evaluating the face at rest, there is no movement and thus no alternation

in the distance. The **limitations of this research** are: not conducting an analysis to study secondary symptoms such as synkinesis (demand 3), reproducibility in terms of intra- and interobserver variability (demand 4), or objectivity, sensitivity, and specificity of the method (demands 5, 6 and 9 respectively). The hypothesized suitability of the Face Hugger method and the evaluation plan for each demand are summarized in Table 4.1.

Table 4.1 *A summary of evaluation methods of each facial grading system demand. The demands marked to be the limitations of the research or method are outside the scope of this thesis.*

No.	Demand	Evaluation
1	Dynamic measurement	Analysis
2	Static assessment	Limitation of method
3	Secondary symptoms such as synkinesis evaluation	Limitation of research
4	Good reproducibility; low intra- and interrater variability	Limitation of research
5	Objectivity	Limitation of research
6	Sensitivity	Limitation of research
7	Minimally invasive	Discussion
8	Clinical convenience	Discussion
9	Specificity	Limitation of research
10	Quantitative	Analysis
11	Cost effective	Discussion
12	Fast to use	Discussion
13	Not requiring plenty of equipment	Discussion

The demands relisted in Table 4.1 are further explained in theory part and especially in Section 3.5. Table 4.1 illustrates that from the thirteen demands, two demands will be mathematically analyzed, five discussed, one stated to be a limitation of the method, and six left outside the scope of this thesis.

4.3 Using Build-Measure-Learn Feedback Loop

A **build-measure-learn feedback loop** introduced by Ries in [106] provides the framework for representing this thesis' methodology part. The build-measure-learn principle is visualized in Figure 4.1.

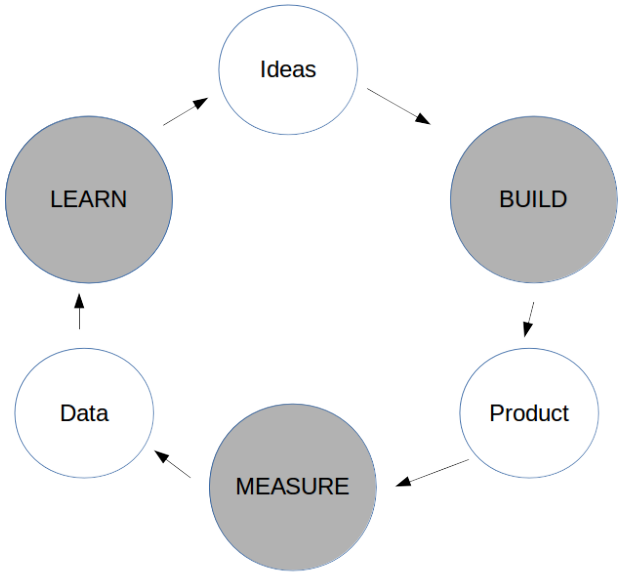

Figure 4.1 *The build-measure-learn feedback loop illustrated. The image is adapted from [106].*

The build-measure-learn principle utilizes **validated learning**: instead of making numerous assumptions and aiming to create a final product by once, or here to answer the research question with a single process flow, adjustments can be made. The build-measure-learn principle gives the guideline of how to adjust the process, or iterate. Making a lap in the build-measure-learn loop represents one iteration cycle. As Figure 4.1 shows, the three-step loop begins by **building** a **minumum viable product (MVP)**, **measuring** how does the MVP answer to the fundamental hypothesis, and then **learning** from it. The learning outcome gives the next steps; should a sharp turn be made - **a pivot**. The fundamental idea of the method is to minimize the total time around the loop, to learn and improve fast. [106]

The build-measure-learn feedback loop is suitable for creating new under conditions of uncertainty [106]. Ries intended the method for entrepreneurs and companies of different sizes. It is used here due to two reasons. Firstly, the research question is vast and intermediate steps enable modifying the approach when necessary. Secondly,

the novelty of the topic limits the sources to look examples and take guidance from. Thus, an iterative research approach is useful in tackling the uncertainty arising from the two reasons given.

Prior to conducting the research, there is no knowledge of the exact amount of build-measure-learn cycles to be performed. Two cycles are intended to **answer the research question**: a preliminary experiment studying healthy test subjects and a patient measurement step to research facial palsy patients. The **preliminary experiment** is conducted to evaluate if the approach can differentiate between symmetrical and asymmetrical movement. These measurements are done with healthy volunteers. The preliminary mock-up experiment also provides an opportunity to assess the set-up, protocol, and data preprocessing and analysis. The purpose of this first step is to provide a **proof-of-concept (PoC)** to avoid unnecessary patient measurements. The second step is to perform **patient measurements** on facial palsy patients in the Hospital District of Helsinki and Uusimaa (HUS) Plastic Surgery ward. The patient measurement step compares the measured data against Sunnybrook scale that was detailed in Subsection 2.3.3 to investigate the quantitativity (demand 10) of the approach. The second build-measure-learn cycle also **addresses the objective** of this thesis; how to develop a software to enable further studies.

The following sections explain the details of the build-measure-learn loops of this thesis. Section 4.4 describes the steps that are shared between the different iterative rounds. Section 4.5 details the choices specific for the preliminary experiment, and Section 4.6 narrates and reasons the specialties of the patient measurement and analysis. Finally, Section 4.7 reports the actions done after pivoting at the end of the second build-measure-learn loop.

4.4 Starting Point

This current section explains the steps and details that are shared between the different iteration steps.

4.4.1 Set-up and Protocol

The measurement **set-up** includes:

1. Face Hugger prototype
2. Laptop with current measurement software

3. Chair and table
4. Video camera and tripod
5. External screen (patient measurements only)
6. Mirror (preliminary measurements only)

Each test person wears the prototype that measures the participant's facial expressions. The capacitive data is send from the prototype to the laptop over a Bluetooth connection. Thus, the laptop requires a Bluetooth radio itself. The current software described in Section 3.6 is used to control and record the measurements and show instructions to the test persons, and thus the laptop needs Windows operating system. The test participant is given a choice between having the instructions shown in Finnish or English. An office chair with a head rest is provided to enable a relaxed and comfortable position for the test person for the duration of the measurement. The test participant sits on the chair that is placed in front of a table. The laptop (preliminary experiments)/an external screen (patient measurements) is set on that table and directed towards the test person to show the written instructions. Depending on the test person, the distance between the laptop and participant may vary slightly; the person is asked to make her/himself comfortable and to be close enough to the table to see the instructions properly. A video camera is placed on a tripod and oriented to capture frontal image of the test person performing the movements. The video is taken in case outliers or other inconsistencies appear in the data; one could verify from the recording if the test participant for example performed an incorrect movement.

The **protocol** of the measurement session begins by informing the test person of the scientific purpose of the experiment and explaining the right to terminate participation at any time without reasoning. Each participant signs an **informed consent** to participate to the study, and allow capacitive data usage for analysis and repsorting. This permit is required in order to participate the study. Additional consent forms are a) allowing video recording for data analysis b) and permitting video data usage in scientific publications and presentations. The participants are notified that the research does not give rise to any danger or contain a risk for the participant, and that no benefit such as payment comes from the participation. An ethics board approval of the measurement set-up, protocol, purpose, and data storage preceded the patient measurements. The plastic surgeons involved inform the voluntary patients on the core contents of the approval.

A **practicing round of the movements** foregoes the measuring as well. In the healthy test participants preliminary measurement, the test subject is given a mirror and a moment to try out the movements. The reason for this is the inclusion of unnatural asymmetrical movements; performing for example one-sided smile requires

thought from the healthy participants and the mirror gives the participants instant feedback how did they succeed. The participants are instructed to choose either right or left facial side to be the "paralyzed" side, and not to switch the side during the entire experiment. The patients are not given any mirror or asked to practice; instead they are encouraged to do the movements as well and big as possible, yet naturally.

The next step is to **adjust Face Hugger** on the participant's head. The uppermost extensions are set to cover forehead area and especially the eyebrow movement. By this positioning, the activity of frontalis and corrugator supercilii muscles can be recorded. The middle extensions are placed parallel to the cheekbones and the bottom extensions alongside the jawbones. The lower extensions target zygomaticus major and triangularis muscles. The extensions should be placed approximately 1 cm away from the skin.

The final preparatory step before measurements is the **practice round**. That round simulates the actual measurement phase; the prototype is worn and the measurement set-up is used. Each movement is included once. After the practice round the participant is given an opportunity to ask questions.

After the aforementioned preparations the **actual measurements** can be conducted. The measurement consists of two movements chosen from the six movements given in Section 2.2. Out of those six typical movements to evaluate the facial function level, **eyebrow lift** and **smiling** are chosen. Those movements are chosen as they involve different areas of the face and the prototype has been used succesfully with these movements as discussed in Section 3.6. The movements are limited to two, since the preliminary study involves also asymmetric counterparts and the patient measurements include Sunnybrook assessment. Thus, as the amount of repetitions is wanted to be relatively high to improve reliability and to assess the technical reliability, the amount of movements used is kept low in order to maintain reasonable measuring time.

4.4.2 Preprocessing

The purpose of the preprocessing is to process the raw data into ready-to-be-analyzed form. The preprocessing step has several substeps each with their own motivation. The main parts of the preprocessing phase are common for the two types of data sets, the preliminary and patient data, and are detailed here. The major steps of the **flow of preprocessing** originate from the signal processing principles used in [6–9].

The steps are visualized in Figure 4.2.

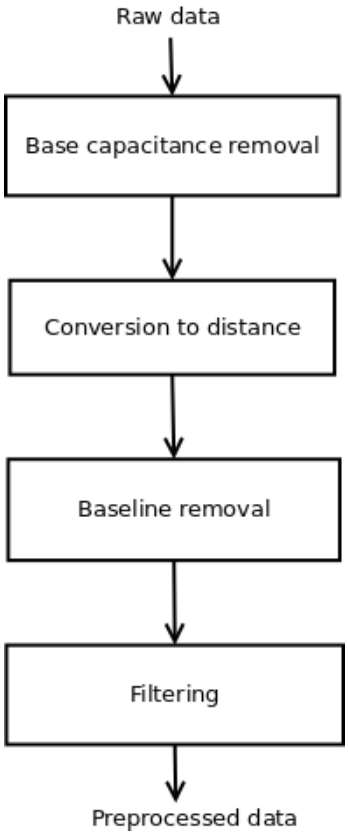

Figure 4.2 *The flow of preprocessing the capacitive data illustrated. The image is adapted from [8].*

As Figure 4.2 shows, the preprocessing protocol has four substeps. The first one is the **base capacitance removal** from the raw capacitance signal. The base capacitance is present in the measured signal as an offset, or noise, and is due to the properties of the used electronic components [7]. In theory, the capacitive signal should be zero without a user wearing the prototype and conducting facial movements. As that is not the case, the base capacitance needs to be removed.

The base capacitance can be removed by measuring it, and then subtracting it from the data. The base capacitance is obtained in the measurement room prior to the measurements. The base capacitance data consists of three separate data sets that are recorded sequentially without anyone wearing the prototype. The base capacitance data recording procedure is simply to use the measurement location and measure the **offset** of the equipment by spreading Face Hugger extensions as open as possible and holding still the sensors far away from each other (approximately 180 degrees can be reached to be between the contralateral extensions by twisting with hands from the ear muffs of the prototype) and other equipment or persons; away from any conducting object [7]. The three capacitive data sets are then used to compute an arithmetic average for each channel over the capacitive base data

measuring time of 30 seconds. Finally, the actual base capacitance removal can be conducted; the just computed base signal is subtracted from the measured raw data capacitance data according to Eq. (4.1):

$$C = C_{\text{raw}} - C_{\text{off}} \tag{4.1}$$

Applying the Eq. (4.1) [7] to raw data literally removes the offset (C_{off}) from the raw data signal (C_{raw}) producing base-capacitance-free signal (C). The same base capacitance data is used for each measured data set.

After the base capacitance removal, the second step is to convert the raw capacitance data into **distance signal** as represented in Figure 4.2. In other words, the raw data contains information in the form of Eq. (3.1). By applying Eq. (3.2) to every data point the distance signal can be gained. The distance signal is thus the inverse of the capacitance signal as the area and permittivity are reduced due to our interest on the proportional distance instead of the absolute distance [7]. Prior to the conversion, the raw signal is checked for zero values to avoid division by zero. If zero is found, it is replaced by the average of its surrounding samples or the neighbor sample in case it is at an end point of the signal.

The third step of the preprocessing is the **baseline** computation and removal, this step is marked in Figure 4.2 as well. The purpose of this step is to normalize the distance signal. There is a baseline present in the measured distance data that originates from the test participant simply wearing the prototype with relaxed non-moving face. As the interest is in the relative change in distance, the baseline should be removed [7]. Thus, the baseline is computed: with preliminary data by using a constant false alarm rate (CFAR) principle, and with patient data by a long median filter. The details and reasoning behind choosing the methods mentioned are explained in Subsections 4.5.2 and 4.6.3, respectively. Finally, subtraction of the baseline from the distance signal in a sample-wise manner produces a normalized distance signal without the baseline.

The final step of the preprocessing is to apply an **average filter** as marked in Figure 4.2. The average filtering is done to each channel separately. The purpose of this step is to remove noise; to smoothen the signal. The length of the filter is chosen to be approximately 20 % of the samples recorded in a second. This is the result of experimental iteration and observing the filtered signals against the unfiltered ones. The goal is to remove the noise but keep the maximum values of the signals as unaffected as possible.

Thus, to **summarize the preprocessing**, firstly, the base capacitance is removed from the raw capacitive data in order to remove the noise of the equipment. Secondly, that base-capacitance-free signal is converted to a distance signal. Thirdly, the

distance data is normalized by removing the baseline and finally smoothened with an average filter. This preprocessing protocol produces data ready to be analyzed; signals that are proportional to changes in distance.

4.4.3 Quantities and Parameters

The basis for the later data analysis phase, for both preliminary and patient data, is to choose which quantities to investigate. This decision is based on theory given in Section 2.4 where the perception of facial symmetry was reported to be based on static and dynamic thresholds. The static threshold translates into contralateral spatial displacement difference, in other words the magnitudinal difference of the facial expressions. The static threshold was discussed in detail in Subsection 2.4.1. On the other hand, the dynamic threshold is stated to depend on the temporal difference between the two facial sides but also on the vector of excursion; the way the facial expression is achieved. The dynamic threshold was detailed in Subsection 2.4.2. Thus, the **quantities to study** are proposed to be:

1. Spatial difference between contralateral sides
2. Temporal difference between contralateral sides
3. Vector of excursion

where the first one is an indicator of the static nature of perception, and the two latter ones of the dynamic characteristic.

The quantities are studied by comparing certain parameters between the contralateral facial sides. To be more precise, each **contralateral channel pair** is compared individually. A channel pair consists of the corresponding channels from the right and left sides. The structure of the prototype has been illustrated in Figure 3.14, and from that picture the channel pairs are the channels with the same number. The parameter difference is computed separately for each 1) test person 2) movement 3) channel and 4) repetition, to give the hierarchy from top to bottom.

The **analyzed parameters** are chosen to describe the listed quantities and measure their changes. To begin with the dynamic symmetry analysis, the **temporal difference** is assessed by studying the contralateral differences in time domain. This parameter is analyzed with different approach in preliminary and patient study, and thus the further discussion is left for the corresponding Subsections 4.5.2 and 4.5.4, and 4.6.3. Further, it should be noted that with this temporal quantity, a significant limiting factor comes from the prototype. As stated in Section 3.6, if all 22 channels are recording, the maximum sampling frequency is 29 Hz. If the number of recording

channels was decreased, the sampling frequency would improve. In other words, the temporal resolution is inversely proportional to the sampling frequency.

The dynamic symmetry analysis then continues to study the **vector of excursion**. **Cross-correlation** is used as an indicator of the facial movement's direction. In other words, the correspondence of the contralateral waveforms is investigated by computing the cross-correlation coefficients. For example in [107], cross-correlation is defined as

$$f \star g = f(-t) * g(t) \tag{4.2}$$

where $*$ stands for convolution. As Eq. (4.2) shows, the cross-correlation of two signals can be computed by turning one signal and then convolving the signals. Thus application of Eq. (4.2) to each channel pair's every repetition produces a series of cross-correlation coefficients. For each repetition, the maximum cross-correlation is used. In general the cross-correlation coefficients are between -1 and 1; the negative end describing a situation of matching inverse shape and the positive end corresponding to matching signal shape. Therefore an absolute value is taken prior to finding the maximum cross-correlation coefficient.

To evaluate the magnitude of the vector of excursion, the **speed of the movement** could be analyzed. The minimum and maximum of distance signal derivative would give the smallest and fastest contraction speeds. Again, these values should be compared between the corresponding channels. However, this assessment is not performed in this thesis and is included here for curiosity.

The static symmetry analysis is conducted by assessing the **spatial difference** between the contralateral sides. In more detail, the magnitudinal difference between the channel pairs is evaluated by comparing the **maximum amplitudes**. However, the methods used for preliminary and patient data differ and are thus discussed in their individual Subsections 4.5.2 and 4.6.3, respectively.

To **summarize the analyzed quantities** and their indicating parameters, the analysis is split into dynamic and static quantities as proposed in literature (Section 2.4. The dynamic characteristics of facial symmetry are investigated by temporal difference and vector of excursion. The static nature of facial symmetry is studied by the movements' magnitudes. The analysis is done by comparing the corresponding channels and computing the differences in four levels; on test person, movement, channel, and repetitition level. Finally, statistical analysis is conducted on intra- and/or inter-person level to provide material for the next steps, and for discussion and conclusion.

4.5 Approach I: Preliminary Experiments

This current section describes the first cycle through the BML feedback loop; the preliminary study, or PoC, on healthy test subjects. The aim of this cycle is to determine, if it is possible to differentiate between symmetrical and asymmetrical movement. Based on this cycle, it is resolved, whether or not the patient studies should be implemented.

Subsection 4.5.1 explains the *build*-part. The basis for the *build*-part comes from Section 4.4. Subsections 4.5.2 and 4.5.3 correspond to the *measure*-part of the loop. Finally, Subsection 4.5.4 represents the *learn*-part; it holds the discussion on preliminary results and whether to persevere or pivot.

4.5.1 Data Collection

Two movements are used in this thesis' experimental part: smile and eyebrow lift. The healthy subjects conduct both symmetrical and asymmetrical repetitions, as the purpose of this PoC step is to deduce if the symmetrical movements can be distinguished from the asymmetrical ones. In addition, two different movement intensities are inspected to provide information on the distinguishment threshold. These choices result in having eight different movements. The **movements and their intensities** are:

- Eyebrow lift: symmetrical, max intensity
- Eyebrow lift: asymmetrical, max intensity
- Eyebrow lift: symmetrical, medium intensity
- Eyebrow lift: asymmetrical, medium intensity
- Smiling: symmetrical, max intensity
- Smiling: asymmetrical, max intensity
- Smiling: symmetrical, medium intensity
- Smiling: asymmetrical, medium intensity

The instruction which movement to perform is shown to the test participant on a laptop screen. A matlab script is made to **randomize the order** of these movements and for each participant a new set of random order instructions is generated. This ensures that the results are independent on the movement order. The instruction set

is imported to the current software as a text file. The software shows **each instruction for five seconds** and then relax instruction for eight seconds. Thus, before and after each movement there is **eight seconds of relaxing** when the participant is instructed to keep a natural relaxed expression and to avoid movements. The relaxing periods of the data are used for data analysis purposes such as baseline removal. Each movement is repeated in the instructions ten times, and additionally the measurement set begins by five repetitions of max intensity symmetrical movements to provide **references**. Thus, the total duration of the healthy subject measurement is **19.5 minutes**.

The test participants are **volunteers** from the university. They are **nonnaive** as the purpose of the study is informed to them. In total **20 people** participated; nine women and eleven men. The test participants have **no history of facial palsy** or major facial trauma, and are aged between 23 and 50 years the average being 33.25 years and median 33.5 years. Each participant consented on capacitive data usage (Permission 1) and video recording for data analysis purposes (Permission 2a). Some gave permission to use video in scientific publication and reporting means (Permission 2b). The measurements took place in a quiet test laboratory room in Tampere University to avoid disturbance. Tables B.1 and B.2 in Appendix B summarize the test participant details and include notes about fitting of the prototype (whether the sensors touched the face) and the participant's ability to conduct the asymmetric movements.

From Tables B.1 and B.2 it is clear that the asymmetric eyebrow lift is difficult for the healthy volunteers. Only three participants could perform it well, six not at all, and the rest to some extent that produces a minor difference between the eyebrows during the lift. Also, the fitting of the prototype shows its problems; Face Hugger fitted only three participants well without any sensors touching the face. The size of the prototype just did not allow good-enough adjustment for every participant. If the sensors touch the face, the measured data is compromised. The fourth sensors of the lowest extension and/or the third sensors of the middle extensions touched 16 participants. These mentioned sensors are the most lateral sensors of the lowest and middle extensions and can be seen for example in Figure 3.14. One participant reported possible touching of sensors during the measurement. Also, for one participant, the prototype moved during the measurement session.

4.5.2 Data Analysis

The quantities and parameters to be analyzed within the scope of this thesis were chosen in Subsection 4.4.3. This section aims to explain that how each parameter is computed in this current preliminary study.

Matlab is used for the data analysis for two reasons: it enables fast **scripting** for PoC and as it was used in dissertation [6], some parts could be reused. Reading correct columns from the files into variables and the preprocessing steps (see Subsection 4.4.2) are taken from the dissertation work, and refactored and complemented when necessary. For example, the data structure to store the measurement data and its processed forms needs to be build, but the logic to convert the capacitance data into distance signal not. After every scripting step, the input and output of the phase in question are observed in graphical form. Also, the dimensions of the variables and data structures are investigated. These actions are taken to assess the correctness of the script.

Other preprocessing steps than the **baseline computation and removal** are detailed in Subsection 4.4.2. For the baseline computation a ready Matlab function developed in dissertation [6] studies is available. In that function, an adaptive threshold is computed by applying the **constant false alarm rate (CFAR) principle**. In a nutshell, the adaptive CFAR filter works as follows; the filter slides over the signal and consideres a sample point and certain amount of its neighbors called reference samples at a time. The reference samples of the point under observation are summed together and multiplied by a predefined constant or sensitivity parameter, this gives the adapted threshold. (In the function made in the dissertation project, Matlab's ready cfar-function is utilized at this threshold computation step.) If the sample point exceedes the just computed threshold, the sample is replaced by the threshold. Once this sliding filter is applied to the whole signal, one channel at a time, the result is the baseline; all the peaks caused by facial movements are removed. Finally, the subtraction of the baseline from the distance signal in a sample-wise manner produces a normalized distance signal without the baseline. In other words, adaptive CFAR removes all the movement induced part from the signal which provides the baseline that exists purely from the sensors coupling with a relaxed face. Once this baseline gets subtracted from the original signal, only interesting parts arisen from facial activity are left.

When designing the measurement set-up, the relax and movement periods' lengths in seconds are determined by this baseline removal step. The relax period is targeted to be long enough (8 s), and the movement period short enough (5 s) so that even when

computing the adaptive threshold for the middle sample of the movement there would be enough relax samples included. This is an important goal as the relax samples contain the in-theory-pure baseline whereas the movement part has facial activity in it. If the baseline is computed with too many movement samples as a reference in it, we might accidentally remove relevant data from the signal. The length of the filter is chosen to be 14 s; with the combination of 8 s relax period before and after each movement, 5 s movement and window size of 14 s, it is guaranteed that even in the middle of the movement there will be a majority of relax samples when computing the adaptive threshold. The sensitivity constant is set to 2 based on recommendation from the dissertation [6] studies.

After preprocessing, for the **data analysis** part the cross-correlation coefficients are computed as described in Subsection 4.4.3: a normalized coefficient for each test subject, movement type, channel, and repetition. The dynamic analysis continues by assessing the **temporal difference** by studying the difference of maximum amplitude locations; the **peak locations**. In order to state if there is temporal difference between the contralateral facial sides, the time in seconds between the channel pair signals is computed. The temporal resolution is 29 Hz as all the 22 channels are used in this preliminary step.

The static analysis is conducted by assessing the **spatial difference** between the contralateral sides. In more detail, the magnitudinal difference between the channel pairs is evaluated by comparing the **maximum amplitudes**. However, instead of direct comparison, the amplitudes are normalized before comparison. As mentioned in Subsection 4.5.1, the measurement begins by five repetitions of each movement at maximum intensity. These five repetitions are measured to provide **references**; channelwise average values for both used movements are computed. The measured repetitions are then proportioned to the suitable reference resulting in a percentage for each channel; how many percentages is this channel's maximum amplitude of the refence. The contralateral difference is then gained from the differential of these percentages.

In this manner, the magnitude is normalized against everyone's own reference before comparing the two different sides. The idea is, that for healthy participants the reference movements are done symmetrically and thus there is a reference available for both sides separately. In the patient case, the reference would be taken from the healthy side. The advantage of this reference system is that the contralateral comparison is more robust after normalization, and that a **personal reference** is used which considers the inter-person variation unlike for example House-Brackmann scale.

Table 4.2 **summarizes the data analysis** of the preliminary phase.

Table 4.2 *A summary for preliminary phase's data analysis methods.*

Quantity	**Parameter**	**Completed**
Spatial difference	Normalized maximum amplitude	Yes
Temporal difference	Maximum peak location	Yes
Vector of excursion	1. Direction; cross-correlation	Yes
	2. Speed; min and max derivative	No

Table 4.2 gathers together the data analysis phase's main steps. To assess the static asymmetry component, the maximum amplitudes are compared after normalization by references. The dynamic nature of symmetry perception is analyzed with two quantities; the temporal difference and vector of excursion. The temporal difference is simply computed by detecting the maximum peaks and comparing the time stamp between the contralateral sides. The vector of excursion has two key factor parameters; the direction and magnitude. The indicator for the direction is the cross-correlation coefficient, and the magnitude could have been assessed by derivatives. These parameters are computed for each repetition separately, and the corresponding channel pairs from contralateral sides are compared against one another.

A brief **statistical analysis** is conducted to be able to evaluate the test set-up, to provide insight from healthy participants to answer the research question, and to assess if the patient measurements should be carried forward. Firstly, statistical numbers such as **arithmetic average**, **standard deviation** and **mean** are computed on **intra-test participant level**. To easen the result intrepretation step, only the most important channel pairs are included into this statistical analysis phase. The decision which channel pairs to include is based on the dissertation work [6] and the observation of the data analysis results in graphical form. The following channel pairs are chosen:

- Channels 5 and 16, and 6 and 17 for symmetrical and asymmetrical smile
- Channels 2 and 13, and 1 and 12 for symmetrical and asymmetrical eyebrow lift

The statistical numbers are computed for the spatial difference and only the tasks with maximum intensity are used. The exclusion of medium intensity smile and eyebrow lift is made after observing the contralateral differences graphically; the

differences with medium intensity are negligible. This is pondered to be due to the difficulty of healthy people performing asymmetric movements on purpose. Once the asymmetric movement is supposed to be carried out with minimum intensity, it is possible that it becomes even harder. In addition to the three statistical numbers, **box plots** are used to gain the preliminary results.

In the **inter-person level**, **arithmetic average** and **mean/standard error** are computed for every movement channelwise.

4.5.3 Preliminary Results

The preliminary results are represented in this current section and discussed in the next one, Subsection 4.5.4. The results are computed in total for 18 volunteers. Thus, two test sets are excluded; the test data sets for subjects number 2 and 5. The two sets are ignored as they were only partially stored due to a software problem.

Figures 4.3-4.10 are example graphs that **visualize the contralateral spatial difference in percentages** for all used 11 channel pairs. The spatial difference quantity was explained in Subsection 4.5.2. The examples show movements symmetrical and asymmetrical eyebrow lift at max intensity, and symmetrical and asymmetrical smiling at max intensity. More example figures are located in Appendix C. The medium intensity movements are left out. However, to cover every aspect, few medium intensity examples are included in the prementioned appendix, to their own section.

As it comes to these example pictures, **every measured channel-pair is visualized** even though the statistical values are computed and shown only for selected channel pairs as explained at the end of Subsection 4.5.2. The choice to visualize every channel-pair at this point rises from not willing to limit the observation of channels too much before truly necessary at the statistical analysis and interpretation step. However, when choosing examples and observing the visualizations, the main attention is on the channel-pairs listed as significant in the previous section; It is not sensible to observe the lowermost sensors when studying eyebrow lift, neither is it reasonable to consider the most lateral sensors from the middle extension that are touching the participant's face when inspecting smile.

Boxplot is chosen to be the visualization method. Boxplots are a powerful tool to show multiple key values in a simple and condensed form that simultaneously allow the entity to be clear for the observer. The boxplots show the median of the data set as a horizontal line (here, the red line), the minimum and maximum values

are represented with horizontal lines at the end of the whiskers, the data values between the 25th and 75th percentiles are within the box, and possible outliers are marked as separate points (here with red plus-signs). A data point is considered here as an outlier, if it is more than 1.5 times the interquantile range away from the box edges. Here, the rows of the figures correspond to the extensions of the prototype; the sensor data originating from the uppermost extension are visualized on the uppermost rows and so forth. The boxplots of the central channels are shown on the left-hand side; the more lateral the sensor-pair is, the more to the right it is visualized. Each boxplot represents a channel-pair; the differences between the corresponding channels from contralateral sides are plotted into one boxplot. Thus, each boxplot visualizes the specific test person's certain channel-pair's all repetitions for the mentioned movement.

Figures 4.3 and 4.4 are **examples of typical smile results**. The interesting channel-pairs for smile repetitions are: 5 and 16, and 6 and 17. Figure 4.3 shows the boxplots for test person 13's symmetrical smile repetitions and Figure 4.4 the asymmetrical smile repetitions. For symmetrical movements the difference between the contralateral channel-pairs should ideally be very close to zero; for example in Figure 4.3 the boxplots of channels 5 and 16, and 6 and 17 are close to zero percentage difference. For asymmetrical movements the contralateral difference should deviate significantly from zero; this is illustrated well for example in Figure 4.4 where the channels 5 and 16, and 6 and 17 have contralatel difference around 100 percent. More typical smile results are available in Appendix C, in Section C.1.

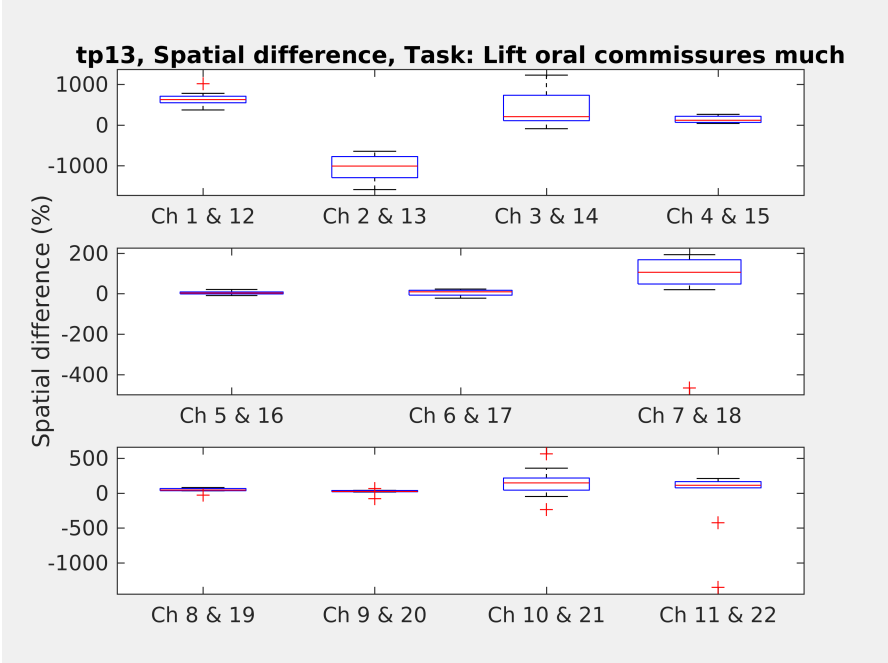

Figure 4.3 *Contralateral spatial difference for repetitions of symmetrical smile, test person 13.*

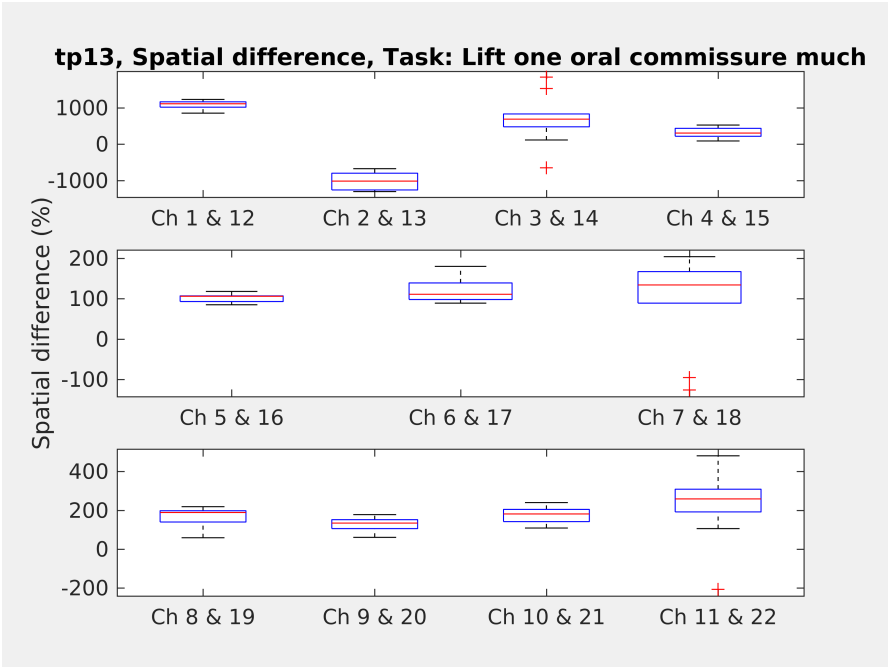

Figure 4.4 *Contralateral spatial difference for repetitions of asymmetrical smile, test person 13.*

Figures 4.5 and 4.6 are **examples of successful eyebrow lift results**; only few participants could conduct the asymmetrical eyebrow lift. This observation is also written in notes in Appendix B. More examples of successful eyebrow lift results are available in Appendix C, in Section C.2. The interesting channel-pairs for eyebrow lift repetitions are: 1 and 12, and 2 and 13.

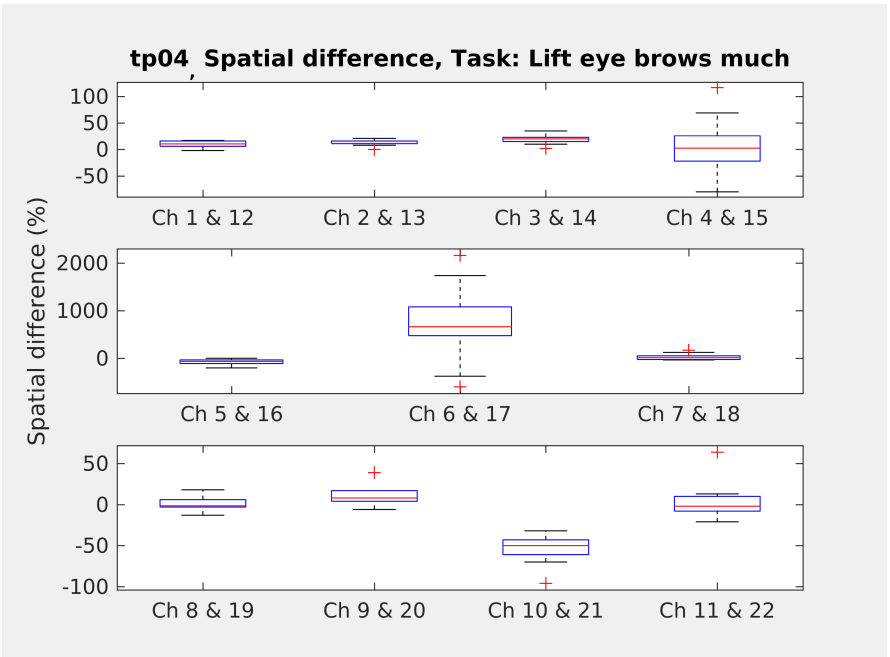

Figure 4.5 *Contralateral spatial difference for repetitions of symmetrical eyebrow lift, test person 4.*

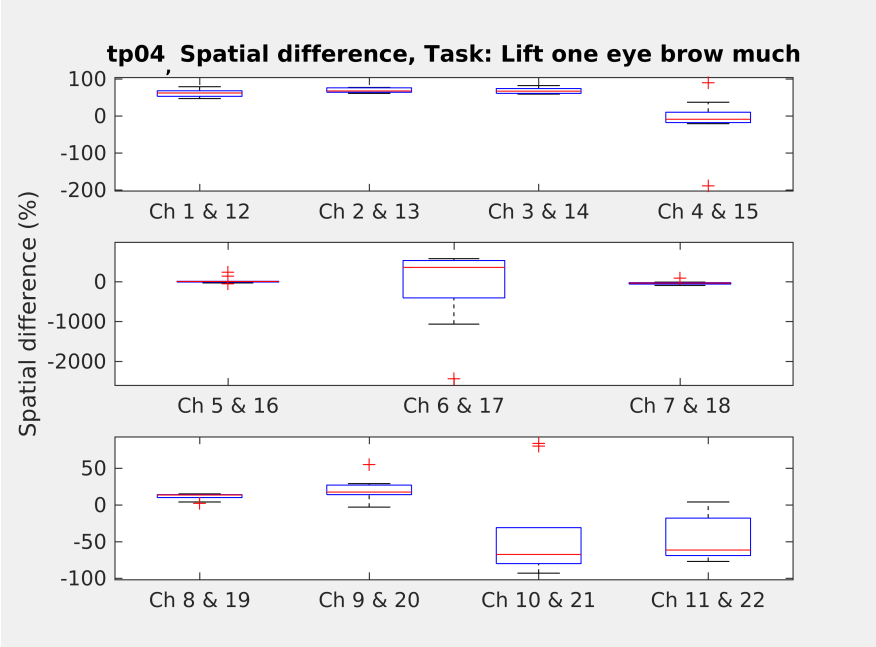

Figure 4.6 *Contralateral spatial difference for repetitions of asymmetrical eyebrow lift, test person 4.*

Figures 4.7 and 4.8 show **examples of unsuccessful eyebrow lifts**; here the participant lifted both eyebrows also in asymmetrical movement. This observation is written in notes in Appendix B, too. Another fail type of the eyebrow lift movement is to have scattered nonsense boxplots. An example of inconlusive eyebrow lift is given in Figures 4.9 and 4.10. More examples of unsuccessful eyebrow lift results are available in Section C.2.

Figure 4.7 *Contralateral spatial difference for repetitions of symmetrical eyebrow lift, test person 13.*

Figure 4.8 *Contralateral spatial difference for repetitions of supposedly asymmetrical eyebrow lift, test person 13.*

Figure 4.9 *Contralateral spatial difference for repetitions of symmetrical eyebrow lift, test person 20.*

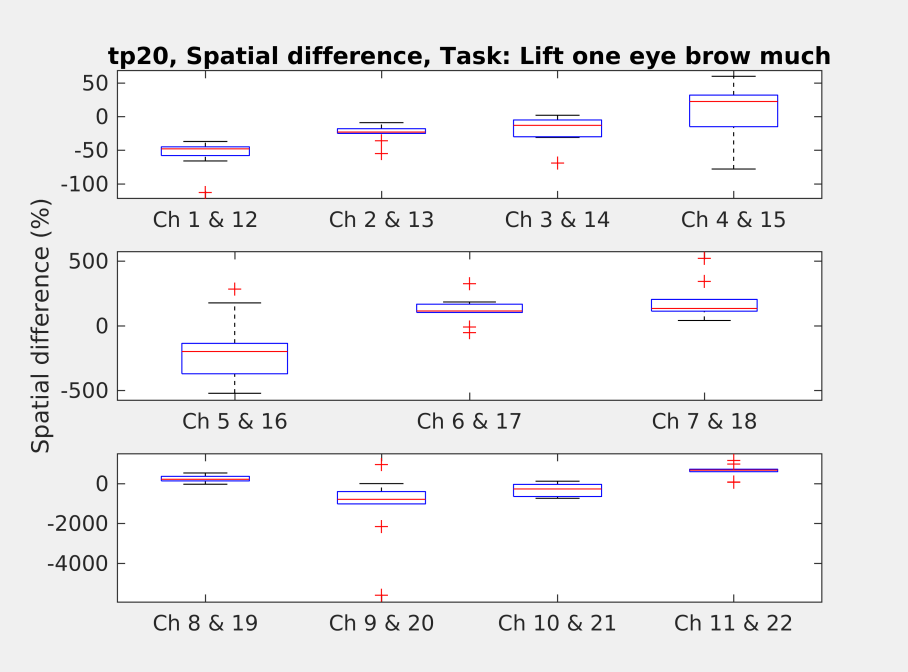

Figure 4.10 *Contralateral spatial difference for repetitions of inconclusive asymmetrical eyebrow lift, test person 20.*

The **preliminary statistical results** are shown in Figure 4.11. They visualize in **inter-person level** the arithmetic average of the contralateral spatial difference, and its standard error. These values are computed for maximum intensity smile and in more detail, for two channel-pairs. The channel-pairs are channels 5 and 6 (marked as Smile 1 or S1, and Asymmetric Smile 1 or AS1 to Figure 4.11), and 6 and 17 (S2, AS2) from the middle extension of the prototype. The corresponding maximum intensity eyebrow lift statistical results are represented with channel pairs 1 and 12 (marked as Eyebrow lift 1 or E1, and Asymmetric Eyebrow lift 1 or AE1 to Figure 4.11), and 2 and 13 (E2, AE2). The further reasoning behind the channel-pair choices is at the end of Subsection 4.5.2.

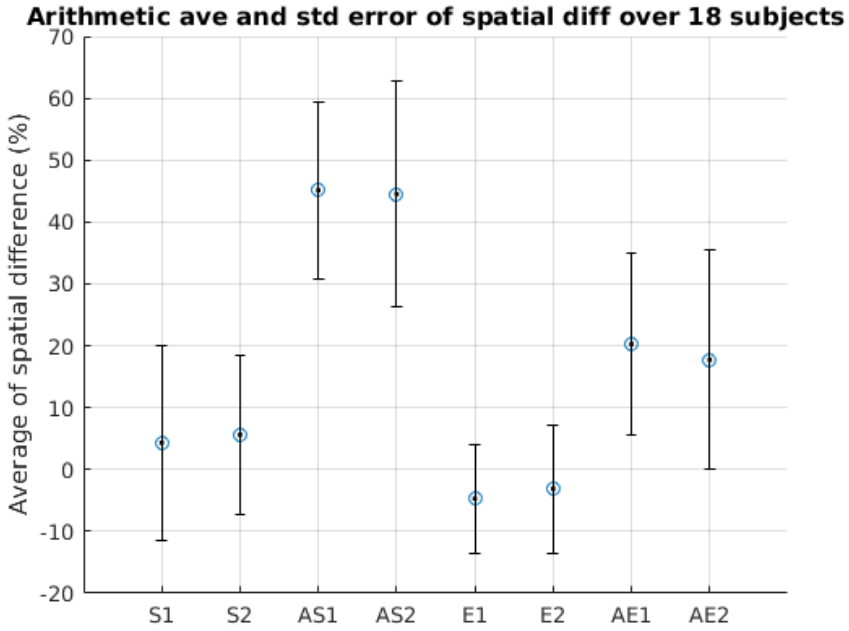

Figure 4.11 *Arithmetic average and its standard error for contralateral spatial difference in percentages computed over 18 healthy test subjects.*

Similarly to example graphs shown earlier in this current section, Figure 4.11 should be read as channel-pairs, too. Additionally, the channel-pairs can be compared against one another (which channel-pair can separate the symmetric and asymmetric movements from each other the best) and inspected as an entity (is it possible to make a distinction between symmetrical and asymmetrical movements with this data set and these channel-pairs).

4.5.4 Perseverence Point

This subsection is the **learning point** for the first cycle of the BML-feedback loop. Firstly, the preliminary results from Subsection 4.5.3 are discussed. Secondly, an assessment whether to **persevere or pivot** is made: Does the original approach seem correct based on the preliminary study? Can symmetrical and asymmetrical movements be distinguished with the prototype? Should the patient measurements be conducted? And if so, are there any recommended adjustments to the set-up, protocol, preprocessing or data analysis.

The analysis focused on inspecting the spatial difference from the three chosen and computed quantities (spatial difference, temporal difference, direction of vector of excursion). The results are represented in the previous section and in its Appendix C. The examples picked for smile (Figures 4.3 and 4.4, and C.1-C.6) show **typical**

smile results at maximum intensity. When observing the test subjects performing the smile-related movements, in general those movements were conducted correctly. Meaning that the symmetrical movement looked symmetrical and the asymmetrical was clearly asymmetrical to a human observer. In Tables B.1 and B.1 there are no notes concerning problems in smile movement execution for any participant. Thus, when inspecting the aforementioned figures, the specific test subject's symmetric and asymmetric boxplots differ. The midextensions' channel pairs 5 and 16, and 6 and 17 have medians very close to zero (4.3, C.5) whereas their asymmetric counterparts have medians in the proximity of 100 percent (4.4, C.6). The dispersion in those cases is relatively subtle, too.

The other picked examples demonstrate a less ideal situation where the symmetrical boxplots (C.1 and C.3) for channelpairs 5 and 16, and 6 and 17 have medians a bit further from zero, the exemplary case. However, their counterparts (C.2 and C.4) are still clearly distinguishable to be the asymmetric movement representatives. In other words, the distinction between symmetrical and asymmetrical movements can be made. In Section 2.4 it is explained that literature acknowledges normal variability in facial symmetry. Thus, normal variation between the facial sides might explain the less ideal results. In such cases, the benefit of this methodology is demonstrated; the usage of own reference and comparing to own facial side still allows detection between symmetrical or asymmetrical movement. However, this is speculative as no fundamental root-cause analysis is conducted. Therefore the reason may also be the prototype tilting during movement, or simply the method not working correctly.

The **eyebrow lift results** with maximum intensity, and thus examples, can be divided into three types. Based on observations, the test participants performed the asymmetrical eyebrow lift a) successfully, there were three participants who could do this, or b) by lifting both eyebrows and thus completing a symmetrical movement in place of the asymmetrical one, there were five participants behaving this way, and c) by conducting a movement that in most cases is hard to categorize either to be symmetrical or asymmetrical, the remaining cases are of this type. These three types show in results, too. The observations for the participant being able to do an asymmetrical eyebrow lift or lifting both eyebrows instead, are written in Tables B.1 and B.1.

Figures 4.5 and 4.6, and C.7 and C.8 show boxplots of the a)-case movements, **the successful eyebrow lift results**. The test participant number four's figures show an exemplary situation; the inspected channel-pairs 1 and 12, and 2 and 13, both from the uppermost extension, have medians close to zero with low dispersion in the symmetrical movement (Figure 4.5) and medians approaching 100 percent

with low dispersion in the asymmetrical movement (Figure 4.6). The other successful example (Figures C.7 and C.8) has asymmetrical median values well below 100 percent, but the symmetrical and asymmetrical movements can be distinguished from the boxplots based on the two chosen channel-pairs. This means, that the boxplots do not overlap. In practice it translates the latter test person being able to conduct the asymmetrical movement but with less intensity than the test person number four.

Figures 4.7 and 4.8, and C.9 and C.10 visualize some results of the b)-case movements, **the supposedly asymmetrical eyebrow lift results**. These figure pairs illustrate a situation, where the median is very close zero and the dispersion is minimal for both the symmetrical and the asymmetrical repetitions. In other words, there was no notable difference in spatial magnitude between the contralateral sides during the symmetrical or asymmetrical movements. These participants completed a symmetrical movement when they aimed for an asymmetrical one.

Figures 4.9 and 4.10, and C.11 and C.12 represent examples of the c)-case movements, **the inconclusive eyebrow lift results**. In such cases, the boxplots of the interesting channel-pairs (1 and 12, and 2 and 13 for eyebrow lifts) overlap for symmetrical and asymmetrical movements. This results in an inconclusive situation; the symmetrical and asymmetrical movements cannot be differentiated. The inconclusiveness is suspected to arise from the test participants' problems to conduct the asymmetrical eyebrow lift.

The example pictures discussed above provide details of the preliminary results. Figure 4.11 shows the bigger picture; **the statistical preliminary results**. For maximum intensity smiles (the bars with labels S1, S2, AS1, AS2), this set-up, protocol, and data analysis pipeline **can distinguish symmetrical and asymmetrical smiles** from this measured data set. This distinguishing capability is statistically valid for both chosen channel-pairs, 5 and 6, and 6 and 17, as the computed standard error bars of symmetrical and asymmetrical movements do not overlap. The arithmetic average for symmetrical spatial magnitude difference between the facial sides is below 10 %, and for asymmetrical movements the corresponding averages are significantly higher; above 40 %. The order and magnitudinal difference of these averages is sensible. The result is aligned with the observation that a typical (maximum intensity) smile was conducted as clearly symmetrical or asymmetrical. The examples support the result, too.

The **preliminary statistical results for maximum intensity eyebrow lifts** are visualized in the same Figure 4.11 as discussed above, with labels E1, E2, AE1, and AE2. It appears, that the **arithmetic averages for the symmetrical movements** and for both channel-pairs are between zero and minus ten percent. The negative averages reveal that for some test participants, there is contralateral difference in symmetrical movement. The spatial magnitude difference is computed as a subtraction; normalized-value-for-side-1 minus normalized-value-for-side-2. Which facial side is 1 and 2 depends on the participant; for both symmetrical and asymmetrical movements, the pretended paralysis side values are subtracted from the healthy side values. The negative average thus originates the chosen symmetrical side eyebrow lifts to be smaller in magnitude in comparison to its reference value than the pretended paralyzed side. Again, literature identifies normal variability in facial symmetry. The **average values of the asymmetric movements** are higher, around 20 %. The order and magnitudinal difference of the symmetric and asymmetric average values is aligned with the observation that only few could actually perform the asymmetric (maximum intensity) eyebrow lift. Thus, the asymmetric eyebrow lift averages are lower than the ones for asymmetric smiles. The inconclusive and symmetrical eyebrow lifts conducted in place of asymmetric ones decrease the asymmetric averages.

This set-up, protocol, and data analysis pipeline **can probably distinguish symmetrical and asymmetrical eyebrow lifts** from the measured data set. The distinguishing capability is statistically valid for channel-pair 1, that refers to channels 1 and 12; the error bars for that channel-pair do not overlap. However, the channel-pair number 2's, that means channels 2 and 13, error bars intersect. Thus, not for every repetition and/or for every test participant, could the differentiation be defined. This was expected based on the observations during the measurements. With that in mind, a conclusion that distinguishing could probably be made if the movements were really asymmetrical too, is reasonable.

The preliminary round evokes few **changes to the set-up**. Firstly, an **extra screen** to show the measurement instructions would be useful. Now, the instruction window opens when the measurements are started. It takes time to extend the instruction window to cover full screen and thus hide everything else from the screen that might disturb the participant. An extra screen that has a full screen instruction window before the actual measurements would easen the measurement start. Secondly, the **low intensity movements should be excluded** from the measurements. In the preliminary analysis they are represented only in Appendix C.3 to provide few examples. They are not included in the preliminary statistical analysis for two reasons; limited time and the observation that the execution was inconsistent. Thus, the idea of providing a threshold for asymmetric and symmetric

movement distinction was excluded for the sake of brevity. Also, there is no point of having medium intensity movements in patient measurements; that would not provide any information on the palsy level, only maximum expression will.

Thirdly, the **amount of measuring channels should be decreased**. As written as observations in Appendix B, the most lateral sensors of the two lower extensions touched the participants' face in some cases. Thus, no information could be gained from these sensors. As a result, those sensors could be muted from the measurement without losing useful data and improving the temporal resolution simultaneously. By dropping in total four channels from the measurement, the **temporal resolution would increase to 40 Hz** for the patient measurements.

The preliminary experiment induced **adjustments also to the preprocessing and analysis step**. To begin with, the baseline removal step (now completed with a CFAR filter) should have longer relax-periods than what were now available in the data. In other words, the duration of the relax period should be longer in relation to the movement part. There are two reasons behind this demand; the reaction time of the participant to stop the movement would be better considered with longer relax period, and also the correct functioning of the filter ensured. Additionally, the temporal analysis should not be based on peak detection and to the peak location difference but to a more robust method. The peak detection is sensitive to noise even after smoothing filter and there is no guarantee about the waveform the actually paralyzed side causes. Thus, the temporal analysis method should make use of the cross-correlation delays. Those values take the waveform shape into consideration. Another matter to note is the choice of tools. In order to not only answer the research question but to fulfill the objective of this thesis, the scripting approach should be replaced with heavier software development means.

To **summarize**, based on the preliminary results, asymmetric smiles can be distinguished from the asymmetric ones. This result is statistically valid. The symmetric eyebrow lifts can probably be distinguished from the asymmetric ones, the uncertainty arises from the insufficient performance of the asymmetric eyebrow lifts. Thus, based on this set-up, protocol, and data processing pipeline by the healthy subjects, **the experiment should be persevered** and patient measurements conducted. However, small adjustments should be made: an external screen added, temporal resolution improved by muting ineffectual sensors, low intensity movements excluded, temporal analysis changed to more robust one, and scripting replaced by other technological approaches.

4.6 Approach II: Patient Measurements and Software Development

This current section details the second cycle of the BML feedback loop. The aim of this cycle is to provide the means to answer the research question; can the prototype be applied for facial palsy severity rating purpose. After the first cycle (Section 4.5) concluded that the used set-up, protocol and data analysis pipeline can differentiate symmetrical smiles from asymmetrical ones, and very likely do the same for eyebrow lifts, the patient measurements are conducted. The adjustments suggested in the previous section are applied though. This current section addresses the objective set for this thesis, too; how to develop a software to enable further studies.

Subsection 4.6.1 explains the *build*-part. The basis for the *build*-part comes from the previous cycle and Subsection 4.4. Sections 4.6.2 and 4.6.3 correspond to the *measure*-part of the loop. Finally, Section 4.6.4 representes the *learn*-part; it holds the discussion whether to pivot or persevere. The objective is covered alongside answering the research question, and also in Appendices E-G.

4.6.1 Data Collection II

The basis for the patient experiment movements come from the preliminary measurements' movements that are explained in Subsection 4.5.1. The movements are simplified in the learning step in Section 4.5.4, as the medium intensity movements were suggested to be left out. Also, obviously there is no need to perform asymmetric and symmetric movements separately on patients. Thus, the **movements for patient experiment** are:

- Smile much (reference)
- Lift eyebrows much (reference)
- Smile
- Lift eyebrows
- (Relax)

The instruction of the movement that should be performed is shown to the test subject on a separate screen, as determined in the learning step of previous cycle (Subsection 4.5.4). The same **randomization approach** as described in Subsection 4.5.1 is applied to ensure that the results are independent on the movement order, again. The same software is used to show the instructions, and to control the prototype and record the data. The same Matlab script randomizes and generates the instruction text-files.

The **duration of the movements** are changed from the preliminary experiment; the relax period is now longer. The durations for the preliminary experiment are explained in Subsection 4.5.1, reasoned in Subsection 4.5.2, and discussed for change in Subsection 4.5.4. Again, the measurements begin with ten repetitions of the reference movements. Between each reference movement there is a relax period. After the references are recorded, a movement and relax instruction alternate until each movement is repeated ten times. The movement instruction is shown for five seconds as before, the **relax period** however experienced a change in comparison to the previous set-up and is now **ten seconds**. Thus, the **total duration** of the patient measurement is **10.5 minutes**.

Sunnybrook Facial Grading Scale is used as a reference to the prototype results. Thus, the patients need to perform the **Sunnybrook movements** as well when a doctor is observing. There are two plastic surgeons and a spelicialising physician who conduct the Sunnybrook evaluation. They are all trained to use the grading scale. The grading scale is introduced in Subsection 2.3.3 and the application of the scale is completed according to Figure 2.3.

The participating patients are volunteers from Helsinki and Uusimaa District Hospital. They are patients of the two plastic surgeons and a specializing physician involved in the wider MIMEFACE project. The patients did not gain any reward from participation. The patients were informed on the scientific purpose of the study, thus they are **nonnaive**, and explained that the test does not hold any diagnostic power.

The necessary anonymized details of the patients are represented in Appendix D, in Tables D.2 and D.3 to be specific. From those tables it can be read, that in total **17 patients**, 10 women and 7 men, participated the study. The patients were aged at the time between 18 and 90 years the average being 58 years and median 60 years. Also, the **aetiology** of the facial palsy varied, as did the **level** of the palsy. From the 17 patients, one had congential facial palsy, one idiopathic, and one had a palsy due to an ear infection. The remaining 14 patients had facial palsy onset as a result of a tumor or its operation. Test participants number 30 and 32 had recurrent conditions (parotic gland carcinoma and hemangioma tumor, respectively) that were operated successfully in the first place but after renewal the palsy either came back (tp30) or worsened (tp32). The Sunnybrook total sums varied from 8 to 87, the average being 40.8, and median 43.0. The corresponding numbers for Sunnybrook's voluntary movement part are: range between 24 and 92, average of 52.7 and median of 56.0.

The same prementioned tables contain a column for Operations and Weight, the **interventions** completed. These columns document the possible factors that might effect the measurement results, especially the Weight-column. The conducted operations include different kind of nerve transfers, soft tissue addressing in several ways, botox injections and placing a metallic weight to help with eye closing. All the operations target to improve symmetry. Especially the presence of the weight and its material is interesting because of the capacitive nature of the measurements.

The **fitting of the prototype** has improved after the preliminary measurements if the success is measured by how many measuring channels touch the face. From both sides of the prototype (right and left extensions), each extensions' most lateral channels are muted for the patient measurements. As based on the preliminary study, these are the channels that typically touch the face and thus provide unusable data. As written in Table D.1, in the patient measurements only one patient had a sensor touching; test subject number 30's left side's lowermost extension touched from the outermost recording sensor. In comparison, in preliminary experiments, 15 out of 20 test subjects had recording sensors touching (see Appendix B). Other observations related to patient measurements are good but airy fitting (tp33), prototype not reaching to central face (tp34), and sensors very close to the face (tp36). In general, the prototype fitted the patients suitably.

An **observation** made during the patient measurements was that there might be a problem in the reference movement data. The possible issue arises from the prototype moving during the reference movements with multiple patients. The reference movements were instructed to the patients to be conducted as big as possible, yet still natural. This seems to be misunderstood as several participants performed the references extremely extensively. However, the instruction was kept the same for the whole BML loop, here the patient measurements, to gain comparable inter-person results. The flicking needs to be taken into consideration in data analysis phase though.

4.6.2 Technology Choices

The objective of this thesis is to investigate how to develop a software that could be used in later studies. This current section lays out the **general overview** of the software development, and the **technology choices** behind it. Appendices E and F provide additional information. Subsection 4.6.3 concentrates on the analysis methods used for patient data, and how to test the analysis steps.

The current software enables conducting measurements with Face Hugger prototype. However, the **desired functionality** of a software to be used, would have an analysis mode in addition to the existing measurement mode. In other words, the measured data should be possible to be analyzed with the software. This would remove the need to write a separate script to conduct the analysis. The existing software was designed to be a general purpose software. Thus analysis inclusion was not timely or even possible during that development phase. Appendix E details some main parts of the functionality, and describes the fault tolerance.

The existing software is a general purpose program, and programmed and operated by engineers. Thus, little attention was put to **usability** and **graphic user interface** (GUI) design. In further studies, a major additional user group would be medical doctors and therefore the design of the GUI should be given attention. Appendix F details the GUI: Section F.1 overviews the central usability requirements, F.2 visualizes some wireframe-level examples, and F.3 describes usability testing.

The **choice of programming language** is the first technology decision made; C++ is chosen. It is a popular de facto language in PC applications where short latency and portability are required. It enables easy integration potential; for example, if the camera stream is wanted to be included in the program later. Another considered language is Python, but the prementioned performance and some perspectives such as static typing and compiling supports C++; this may be seen a question of opinion, but here the extra check provided by static typing and other compiler checks are thought as an extra safety net. Also, the programmer is comfortable with C++ and object-oriented programming (OOP). Thus, the choice of programming language underlines the need to develop a desktop application rather than an analysis script.

The choice of programming **framework** is Qt. It is native for C++, a cross-platform solution, and available with an open source LPGL license from the Qt web site. The cross-platform is a central requirement as the end-users (engineers and medical doctors) have both Linux and Windows computers. A cross-platform allows of having only one codebase. Qt comes with Qt Modeling Language (QML) that is a user interface markup language. Here, QML is used to build the GUI.

The transition from research prototyping into an application code requires ensuring that the scientific computation can be completed relatively painlesly and efficiently in the production environment. In other words, the choice of the scientific computation **library** should be conducted carefully. Thus, the factors being evaluated are: a) is the library up-to-date, is its development and community active b) how restrictive the library's license is c) does the library support the main analysis steps used during

the first BML, in other words, how suitable the library is to our purposes d) how is the performance, and e) how comprehensive is the documentation. From six options found and considered, Armadillo is chosen.

Armadillo is an open source C++ library designed for scientific computing and computationally intensive experiments. To put it briefly, Armadillo enables efficient linear algebra operations to data structures it provides (mainly matrices and cubes). [108] A new version was released just before the library comparison took place, and the web site had been active recently as well. The license is Mozilla Public License (MPL) [108] and thus not too restrictive. Armadillo holds services to conduct the facial palsy analysis other than cross-correlation and derivative functions. The performance comparisons provide that Armadillo is faster than its rivals: Matlab and Octave, and few previous C++ libraries [108]. The documentation appears reasonable in Armadillo's web page [109].

Another significant library choice to mention is the usage of C++ **standard template library (STL)**. STL is used for building and handling of data structures of the program. The decision to use STL is best described by a quotation from Bjarne Stroustrup's book about C++; *"A standard library facility is almost certainly better designed, better implemented, better documented, and more portable than anything you could design and implement in a hurry. - - - [It] will also make your code easier for others to understand."* [110, p. 1095] Other library and implementation choices are too minor to be considered here.

4.6.3 Test-Driven Development and Analysis

To answer the research question, the patient data needs to be analyzed. To address the objective of this thesis, and to follow the learnings from the preceding BML-loop, the patient data should be analyzed with technologies presented in Subsection 4.6.2 and alterations explained in Subsection 4.5.4. This evokes the need for comprehensive testing; the result determines the direction of following research and aims in future to have diagnostic power. This section states the principles used in low-level, or analysis-related, testing and details the analysis steps on necessary level.

Test-Driven Development (TDD) is a development method where feature implementation begins by writing its test cases. The test cases are then ran and checked that they fail; they should fail as the feature is not implemented yet. After that, the feature implementation is performed and followed by running the tests (should pass now). Necessary refactoring, and retesting, is the final step. Thus, the tests

are run several times and therefore in practice automating the test cases is required. There are several blends to apply TDD; in other words, how much the architecture is designed prior to TDD or at all. A popular variation, also used here, is to predesign the architecture and apply TDD in unit- and module level. [111, pp. 214-216]

TDD is chosen to be used as the programming is done by a single person. Thus, the psychological challenge of trying to break one's own code is circumvented by writing the tests prior to the feature. Also, the mathematical features are convenient for automated testing. A test module called **Qt Test**, provided by Qt for testing Qt applications, is used to build and run the tests. The testing is **data-driven**, and Matlab scripts are used to generate the oracle values. The scripts are a credible oracle for two reasons. Firstly, several features developed in C++ are available as ready or almost-ready services in Matlab. Secondly, if same result is obtained with two different level languages the possibility of random error is relatively low.

The test-case design aims to gain high **code coverage**. In practice that means applying test-design techniques of equivalence partitioning and boundary value analysis. **Equivalence partitioning** is a method to limit the number of test cases; an input and its passing or failing is assumed to represent the whole partitioned group [111, p. 209]. Thus, the amount of test cases is diminished as the partitioned group can be tested with few representatitive cases [111, p. 209]. Also, equivalence partitioning enables effortless concentration on the negative test cases. Here, the targets for testing are the interfaces, limiting values, handling of error conditions, data structures, and execution paths. There are multiple parameters to be partitioned for a single feature. A ready script is used to form combinations from the test parameters in such a way that the total test case amount stays limited.

In **boundary value analysis** the borders of a value range and their environment is tested. The borders are typically borders of equivalence partitioned groups or other logical groups. The border value analysis is well reasoned testing approach as the borders are a very typical location for errors. One way to see border value analysis is that it complements the equivalence partitioning. [111, pp. 209-210]

The application is developed using **object-oriented programming**. The class diagram of the application is shown in Appendix G. OOP brings its own flavor to the testing: the state of the object, encapsulation, inheritance, possible dynamic binding, and exceptions need to be considered in testing. The three first factors are handled in the test case construction. Each test begins by constructing the object in suitable state, and then calling the methods to the constructed object. This approach is an option as the main emphasis on the testing is in logic due to the analytical

nature of the application. In case of inheritance, the base class is used here simply to distinguish different class types, and for future purposes. Thus, in testing the base class is constructed. Dynamic binding is not used in the application. Testing the exceptions are written as their own test cases.

The **preprocessing steps** are implemented as methods into the Movement-class (see Appendix G) of the application. For each preprocessing step - offset removal, distance conversion, baseline removal, and filtering - there is a public interface method to be called. In addition, there is a method to verify the sanity of the raw data. Also, a single public interface method *performPreprocessing* exists that calls the preprocessing methods one by one. The individual preprocessing methods are left in the public interface for unit testing, and thus there is an **encapsulation cost** to keep the application testable.

In preprocessing, the difference to the preliminary phase solution is the **simplification of the baseline removal** approach. CFAR filter is used in the preliminary study to compute the baseline; a Matlab function developed in the dissertation work is available. However, no suitable library implementation is found for the C++ implementation and its development would be laborious. The baseline calculation is reduced to median filter with adaptable length. The default length is set to two times the movement duration. The existing Matlab script is used to verify the suitability of the simplification.

In **analysis**, the **temporal analysis is changed to more robust** one. In the preliminary study, the temporal difference is computed from the maximum amplitude locations on the contralateral side. This method, however, is sensitive to noise and odd waveforms. Thus, the delay of the best cross-correlation coefficient is used instead. In other words, the temporal difference analysis goes from peak difference to maximum waveform match delay.

The second significant change to the analysis is in the usage of **spatial references**. In the preliminary study, reference movements are used to compute a personal reference for each participant. However, an observation from the patient data collection is that the prototype shifted notably on the patients' head during the reference recording. The instruction to perform as big movement as possible for the reference guided the patients to do exaggerated expressions. Thus, an alternate method to normalize the spatial quantities is used. The median over every repetition for the spatial quantity value is used as a reference. Thus, each patient has their own reference. The healthy side's values are used for both sides.

4.6.4 Pivot Point

This subsection is the **learning point** for the second cycle of the BML-feedback loop. The purpose of the second loop is to answer the research question (can the prototype be used to rate severity of facial palsy) and address the objective (determine how to develop a software to enable further studies). However, as described in Section 4.3, minimizing the time around the loop and producing a MVP to be learned from, is at the core of the method chosen. Thus, to enable validated learning, a point to steer the process is used as a MVP exists; the classes that implement the analysis are completed with the TDD approach, and an initial GUI without result visualization is done. Questions such as *is this the way to answer the research question* and *is the time through the BML-feedback loop adequate* are asked.

The **adjustments** made prior to this second loop to the **movements, set-up and protocol** are suitable; the patient measurements go smoothly. However, an exception is the instruction to smile or lift eyebrows much for reference that causes multiple patients to overdo the expression. Thus, neither the movement is natural nor the data usable as the prototype shifts in the test persons' heads. Therefore, for future studies, if reference movements are used they should be collected with more care.

The need for an adjustment in the reference data collection gets emphasized due to the **technology choices** and the usage of TDD. The software development is advanced in parallel to the patient data collection based on the learnings from the loop one. Also, the patient data is not explored prior to analysis with the newly chosen technologies and method; the demo phase is considered to be passed with the preliminary measurements and analysis. These two factors result in laborious change in the analysis as the overall architecture is designed for the initially chosen normalization method. Additionally, the workload in general is a bit much: there is a significant difference to do a PoC and to implement production level code.

Eric Ries lists *"the general feeling that product development should be more productive"* [106, p. 164] to be one central **telltale signal for pivoting**. Here indeed the two questions given at the beginning of this section are answered no and no. The current approach does not seem to be way to answer the research question efficiently for two reasons. Firstly, the **perks specific for patient data** are not known. To the best of knowledge, this is the first time facial palsy patients' facial expressions are measured capacitively. Thus, inter alia the typical waveforms of the facial palsy patient data are not known. In general, the first step in analytical solutions should be data exploration in order to make educated choices. Secondly,

the **necessity of the application** is not verified. The development is founded on the hypothesis that there would be need for further studies with the Face Hugger prototype on palsy patients. In other words, the objective holds a presumption that the research question could be answered positively; that the prototype could provide a potential solution. If the research question was answered "no", there would be no need for a software as no further measurements would be needed. A **new strategic hypothesis** [106, p. 176] is that the research question should be answered prior to further addressing the objective of this thesis.

To conclude, a **technology pivot** [106, p. 176] is made. The main emphasis is on answering the research question. The objective is already mostly fulfilled, however it should not be finalized prior to answering the research question; firstly, to avoid unnecessary work, and secondly, to make use of the patient data as much as possible. This requires data exploration and fast prototyping. Thus, the third BML-loop should have patient data demo approach.

4.7 Approach III: Patient Data and Python Scripting

This current section details the third cycle of the BML feedback loop. The aim of this cycle is to provide the means to answer the research question. This loop uses demo approach, as suggested in the technology pivot point in Subsection 4.6.4. The data collected in the second loop (Subsection 4.6.1) is used, as is the preprocessing method described and updated earlier. The data analysis approaches from Subsection 4.6.3 are applied. In short, the technology choices are changed to support the change of the general approach to demo scripting; a **patient data PoC** is needed.

Python is used as it has several suitable libraries to provide the services that are needed. In more detail, the library for scientific computing called NumPy is used. NumPy has support for multi-dimensional arrays and wide selection of mathematical functions [112]. The results computed with NumPy are visualized with 2D plotting library Matplotlib [113] and visualization library Seaborn [114]. A data science platform Anaconda [115] is used inter alia to manage the library version with a virtual environment. The **advantage of the technology choice** is the enabled data exploration and fast prototyping. Unit tests are not written but the correcteness of the computation results is evaluated from the dimensions and few samples typically from the beginning and end of the array in question. This sanity check is completed after each midstep.

The **analysis** concentrates on the robust quantities: the waveform (cross-correlation coefficients) and temporal difference from the delay of the maximum waveform correlation. Thus, of quantities listed in Table 4.2, spatial difference and derivatives are excluded. A **statistical analysis** is conducted to the waveform coefficient and delay results. The statistical analysis is focused on the **interpatient** waveform and delay trend, and their correlation to the control data; the Sunnybrook values. The comparison of the prototype measurement results to the existing Sunnybrook scale values is what allows the discussion of the suitability of Face Hugger to rate facial palsy level.

The key results are visualized and briefly explained in Chapter 5. That chapter thus functions as the *measure*-phase of the third BML feedback loop. Chapter 6 serves as the *learn*-step. Finally, the last pivot or persevere point occurs in Chapter 7.

5. RESULTS

This chapter offers the results obtained in the third and final BML feedback loop; the analyzed patient data. The results concentrate on representing the **interpatient trend** of the measured data against the Sunnybrook values. This provides the means **to answer the research question** as the prototype measured data is compared to an existing grading scale. In other words, the *measure*-step of the third BML feedback loop is represented in this chapter. It is noteworthy that addressing the objective of this thesis is left to the background since the pivot point (Subsection 4.6.4), and is thus excluded from this chapter. The objective is addressed in Section 4.6 and in its Appendices E-G.

Section 5.1 begins the chapter by laying out the choices behind the visualizations. The following Section 5.2 considers the measured waveform results against Sunnybrook control values. Finally, Section 5.3 juxtaposes the delay results with the gathered Sunnybrook data.

The sources of error, deviations from the expected results, reliability of the results, the result interpretation, and future aspects are all discussed in the Chapter 6. This chapter concentrates on showing the results.

5.1 What to Visualize

This current section discusses the **choices behind the visualizations** of Sections 5.2 and 5.3. Firstly, typical signal pair examples are shown. Secondly, the channel-pairs used to visualize the waveform results are selected for both movements, smile and eyebrow lift. Moreover, the same channel-pair selection is completed for the delay results.

Figure 5.1 illustrates the **different behavior types observed between the contralateral channel-pairs**. In the figure, the green background is used for the healthy side, and red for the paralyzed side. Thus, the first row represents the healthy side of three different patients, and the second row shows the paralyzed

sides for the same patients. Smile movement and the channel number four are chosen to work as an example. In each plot, all ten repetitions are shown; each with their own color. Concentration should now be paid on the overall trends, not to the details. To put it briefly, the channel-pair behavior may be divided into three groups: same direction, opposite direction, and shape difference. To begin with the first group; the movement waves are pointing to the **same direction**. This is the case with the patient number 15 in the figure. There the difference between the healthy and paralyzed sides is in the magnitude and details of the shape. The patient number five's plots are shown in the middle, and illustrate another type of channel-pair behavior; the waves are to **opposite directions**. Also, the magnitude may vary and slight differences in the shape be present. Finally, in the rightmost plot, patient number 12 serves as an example of the case where the wave **shapes differ significantly**. On the paralyzed side, the signal is fluctuating very close to zero and the shape is on most repetitions far away from the healthy side's shape. The patient data fits into different levels of these three groups.

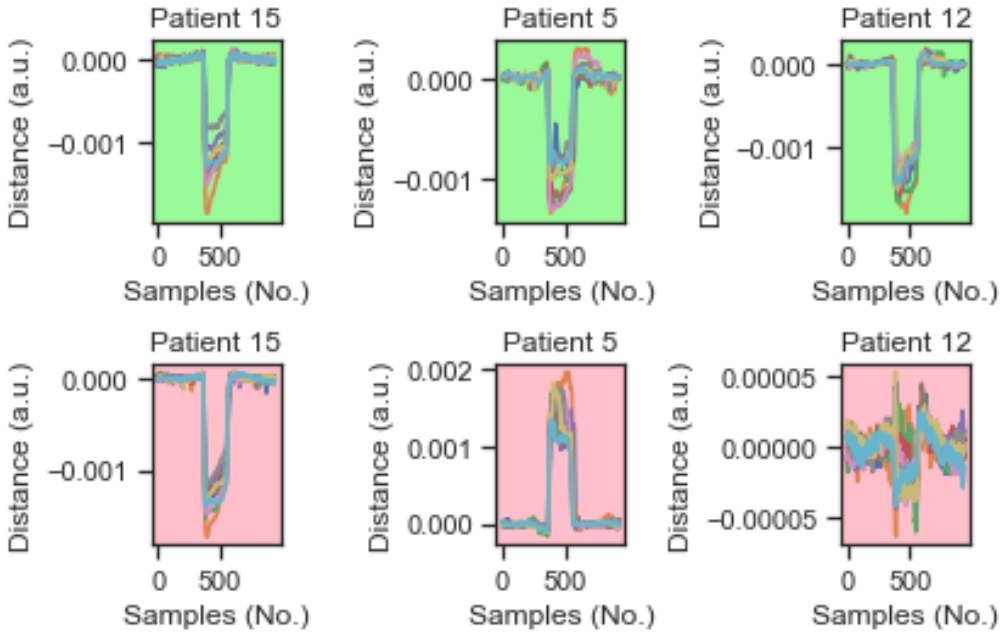

Figure 5.1 *Different channel-pair behavior types illustrated. The green background signals healthy side, and the red one the paralyzed side. The smile movement and channel 4 have been chosen to provide an example.*

The patient data is preprocessed, chopped into movements, and then plotted for **data exploration** purposes. The data is visualized before analysis to gain overview and understanding of it; are the analysis approaches described during previous BML feedback loops suitable now, or should adjustments be made - in the pivot point it was decided that this cycle is the patient data PoC. Table 5.1 summarizes the completed data analysis steps of the patient data PoC. In short, the robust parameters are analyzed. In more detail, the cross-correlation is computed and its coefficients used

to assess the direction of the vector of excursion (Section 5.2), and the delay to investigate the temporal difference (Section 5.3). The spatial difference or the speed of the vector of excursion are not evaluated due to the presence of significantly differing channel-pair signals.

Table 5.1 *A summary for patient data analysis phase's data analysis methods.*

Quantity	**Parameter**	**Completed**
Spatial difference	Normalized maximum amplitude	No
Temporal difference	Delay of maximum cross-correlation coefficient	Yes
Vector of excursion	1. Direction; cross-correlation	Yes
	2. Speed; min and max derivative	No

Appendix H contains similar plots than Figure 5.1 for each patient. Section H.1 shows the smile movement signals, and Section H.2 the eyebrow lift data. These plots are added to provide insight in data exploration sense and to offer additional information for evaluating the results. Because to the best of knowledge this is the first time facial palsy patients' facial movements are measured capacitively.

To show the movement signals in the appendices for each patient and for both movements, and to visualize the computed results on the interpatient study, there is a need to limit the amount of data. In other words, an informed choice is needed about **which channels to visualize**. In the preliminary part this same topic was visited (see Subsection 4.5.2) and two channel pairs chosen for both movements. Here, in addition to the preliminary study's decision, Figures 5.2-5.5 are taken into account.

Figures 5.2 and 5.3 show the channelwise **boxplots over all patients of the best cross-correlation coefficients** of smile and eyebrow lift (EBL), respectively. Each of the 17 patients repeated the movement types 10 times. From those ten repetitions, cross-correlation between the contralateral sides is computed. The highest value of the cross-correlation function for each channel-pair is chosen. Thus, each patient has ten sets of best cross-correlation coefficients for the channelpairs. An arithmetic average is computed for each patient over the ten sets producing an average value for each channel-pair. Therefore, each patient has a channelwise average value computed from the highest cross-correlation coefficients. These average values are then boxplotted over the patient set thus producing Figures 5.2 and 5.3.

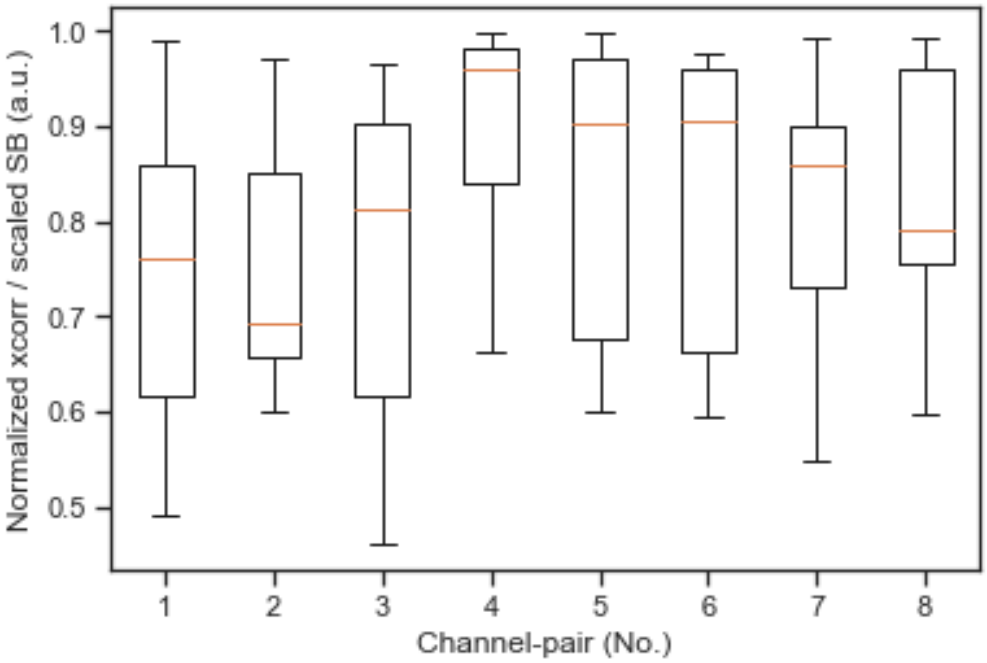

Figure 5.2 *Best smile cross-correlation coefficient averages per channel over all patients.*

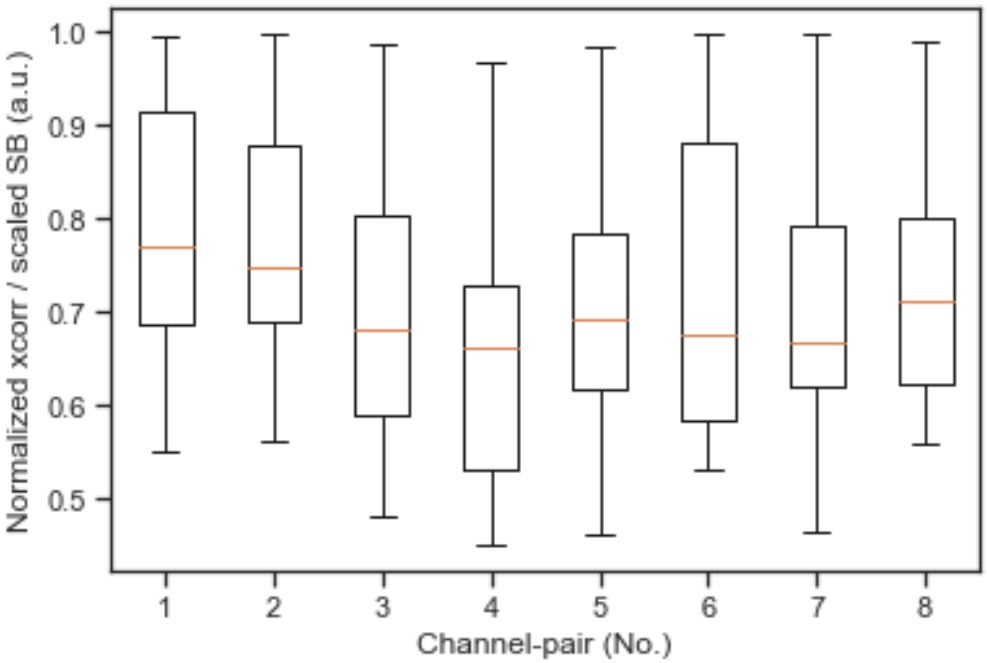

Figure 5.3 *Best EBL cross-correlation coefficient averages per channel over all patients.*

Figures 5.4 and 5.5 are constructed in similar manner than the two previous figures. The difference is, that Figures 5.4 and 5.5 show the **boxplots of the highest cross-correlation coefficients' delays**. In other words, the delays corresponding the highest cross-correlation coefficient are picked, averaged, and plotted over all patients. Figure 5.4 has the delays for smile movement, and Figure 5.5 for eyebrow lift.

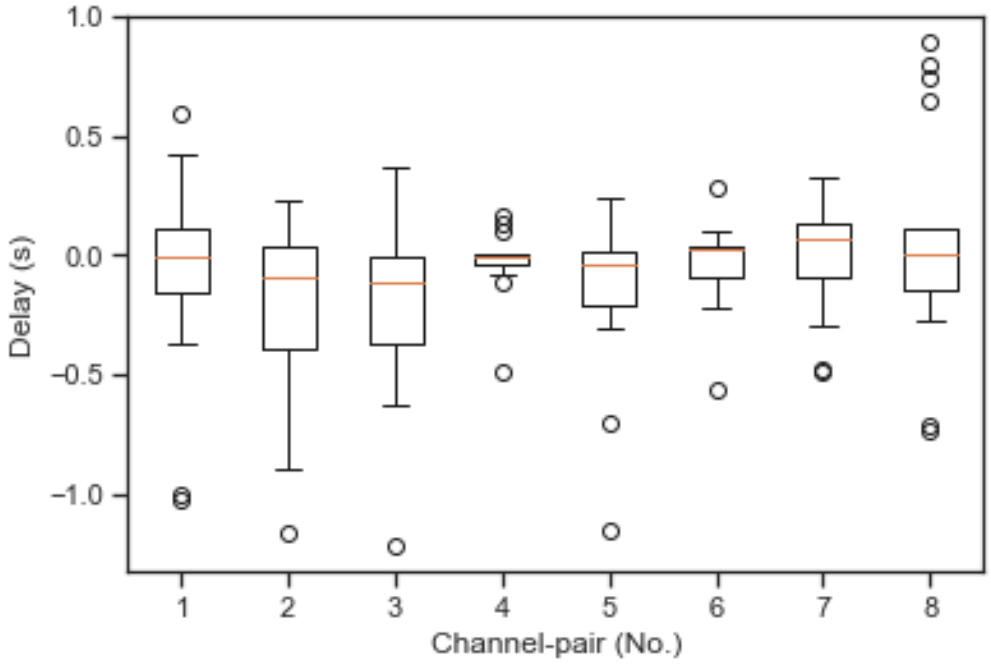

Figure 5.4 *Best smile delay averages per channel over all patients.*

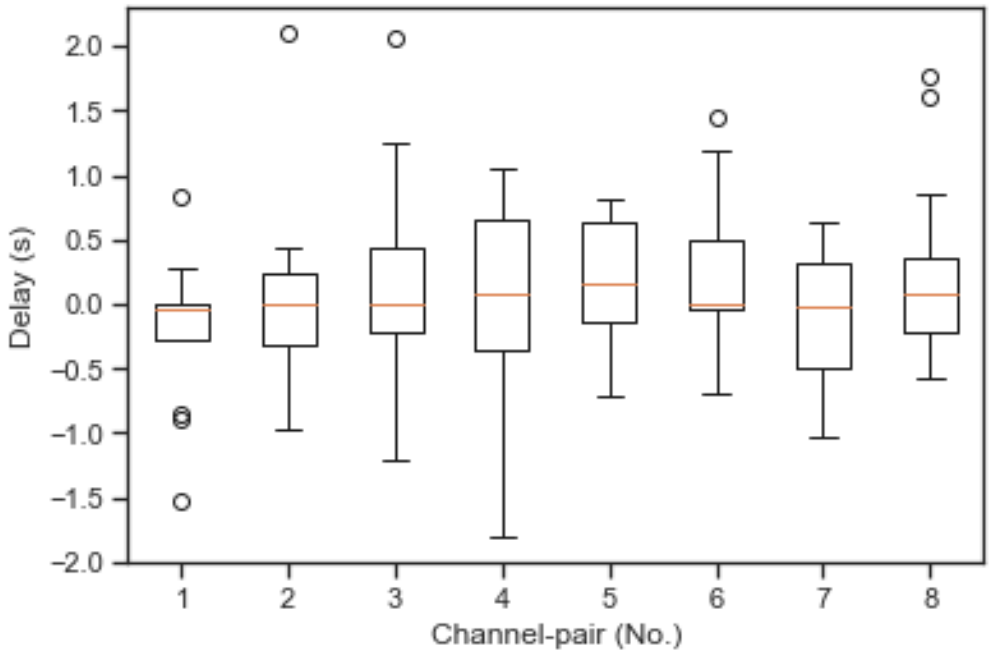

Figure 5.5 *Best EBL delay averages per channel over all patients.*

The four previous figures are shown here a) **to provide an overview** to cross-correlation coefficient and delay results over all patients, and b) to form **a basis for brief discussion which channels to visualize** in the coming sections. Figure 5.2 shows that cross-correlation coefficients' medians are above 0.7 and a bit below 1.0 for smile. The interquantile ranges (IQR) are above 0.6 and a bit below 1.0. The whiskers are limited to maximum and minumum of 1.5 IQR. No outliers are present. The medians and IQRs are highest on channel-pairs 4-6; these are the most central midextension channels. This is the area where the smile is expected to be visible in the data. Based on this figure, also the most central sensor from the midextension, channel four, is to be visualized.

Figure 5.3 shows that the cross-correlation coefficient medians are above 0.65 and below 0.8 for eyebrow lift. The IQRs are above 0.5 and below 0.95. Again, no outliers are present. The medians and IQRs are highest on channel-pairs 1 and 2; these are the most central upper extension channels. This is the area where the eyebrow lift is expected to be visible in the data. Based on this figure, the same channels chosen for preliminary study's visualization are used.

The both boxplots for delays have medians fluctuating around zero. In other words, the best cross-correlation coefficients for both movements are gained close to zero delay. For smile, the IQRs for the channels 4-6 are the smallest. For other channels the midranges are between -2 and 1 seconds. For eyebrow lift, the IQRs for the channels 1-2 are the smallest, too. The other channels have midspreads between -2 and 2 seconds. There are outliers present for both movements, in other words the 1.5 IQR is exceeded several times. The whiskers extent from almost -4 to 2 seconds for smile, and from -6 to 6 seconds for eyebrow lift when observing all the channels. If limiting to the channels 4-6 for smile and 1-2 for eyebrow lift, the whiskers are approximately -1 to 1 seconds and -4 to 2 seconds, respectively.

The argument can be made that using the existing results in Figures 5.2-5.5 to choose which channels to further analyze is having **circular reasoning**, or cherry-picking for the least. However, the choice is made acknowledging the problem, and with the relevance in mind; the symmetry of eyebrow lift should be measured close to the eyebrows as that is where the muscles lifting them and the movement is. The same is true for the smile; due to the muscles used the movement is concentrated on the prototype's midextension area. The **most relevant locations** for the movements are being focused on, recognizing the problems and limitations of the approach.

To conclude, the channels 4-6 are used for analysing smile and channels 1-2 for computing results for eyebrow lift in the following sections. The measured data is compared to an existing system, the Sunnybrook method, next.

5.2 Vector of Excursion: Waveform vs Sunnybrook

This section represents the **waveform -parameter results** for both movements, smile and eyebrow lift. As shown in Table 5.1 and discussed already in Subsection 4.4.3, the waveform investigates the direction of the vector of excursion. The vector of excursion is found to be a crucial factor when a human is detecting asymmetry from another human's face, as detailed in Subsection 2.4.2. The waveform results are compared against the current clinical measure, Sunnybrook Facial Grading Scale, described in Subsection 2.3.3.

Figure 5.6 shows the **smile movement's cross-correlation results** over all patients and against Sunnybrook results. The figure has three lines of different pink shades. Each shade represents a channel from the prototype's midextension; channel 4 is the most central one, channel 5 is lateral to the channel 4, and channel 6 the most lateral sensor of the mentioned ones. Each line is computed by comparing the contralateral channel-pairs by computing cross-correlation for every repetition, and over all 17 patients. For every repetition, the highest coefficient is extracted. The value shown in Figure 5.6 is the average value of the specific patient's ten repetitions' highest coefficients. Every value is visualized; in other words, no outliers are excluded.

In the same Figure 5.6 there are three lines or dashed lines representing the **Sunnybrook values** for the test patients. The figure is organized according to increasing Sunnybrook total score, that translates to increased health in respect to facial palsy. The Sunnybrook scores are scaled between 0 and 1 for visualization purposes; the score is originally between 0 and 100. Sunnybrook is an unweighted regional scale, and the total score is visualized with continuous blue line with label "SB total". The total score contains the resting symmetry and the dynamic movements' symmetry by regions, and the synkinesis evaluation. As only dynamic movements are measured by the prototype, and synkinesis may be detected depending on the location and intensity, the "SB voluntary" score is plotted as well. It is the score of the dynamic voluntary movements. Finally, as the channels picked for this figure represent smile, the "SB smile" curve visualizes the smile score that is embedded in the "SB voluntary" score. To summarize, the level of generality decreases from SB total to SB vountary, and again to SB smile. It is acknowledged, that the Sunnybrook scale is designed to be observed via the total score. The partial scores are visualized here as a curiosity.

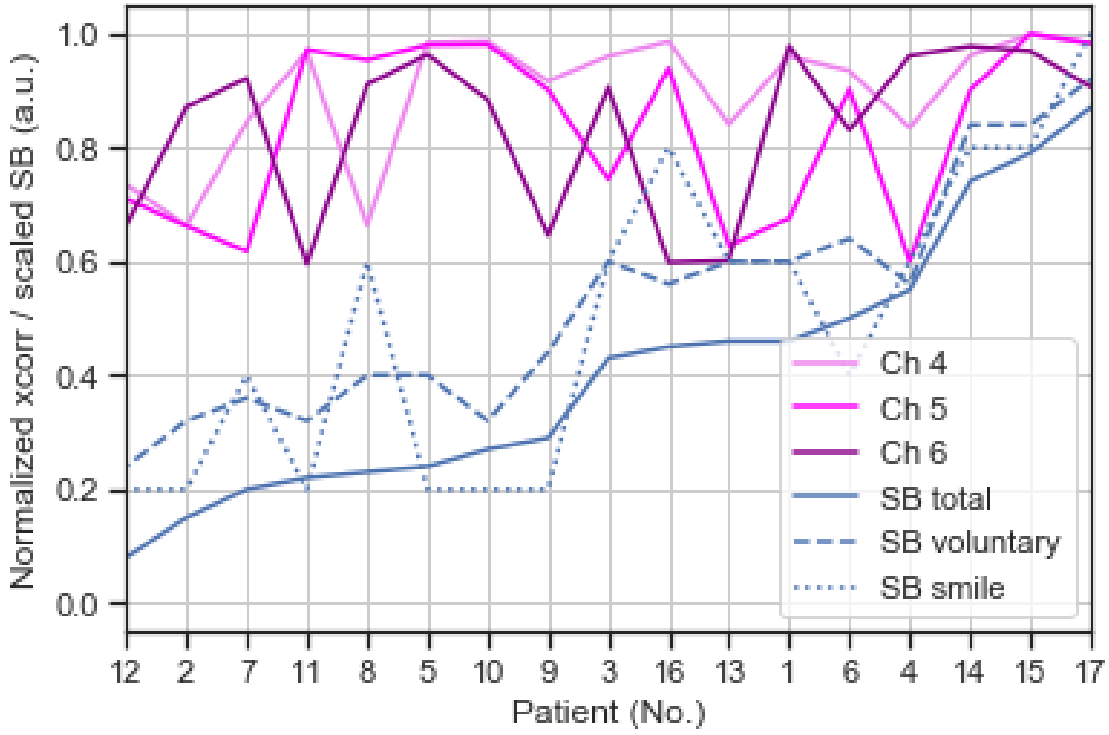

Figure 5.6 *Highest cross-correlation coefficients for smile movement over all patients.*

Similarly, Figure 5.7 visualizes the **eyebrow lift movement's cross-correlation results** over all patients and against Sunnybrook results. The results are computed in the same manner as smile results in Figure 5.6 except that the channels are now from the upmost extension; channel 1 is the most central channel and channel 2 is its lateral neighbor. Also, one Sunnybrook result differs from smile results; the "SB EBL" which is the Sunnybrook score for the eyebrow lift movement. That value corresponds to "SB smile" in Figure 5.6.

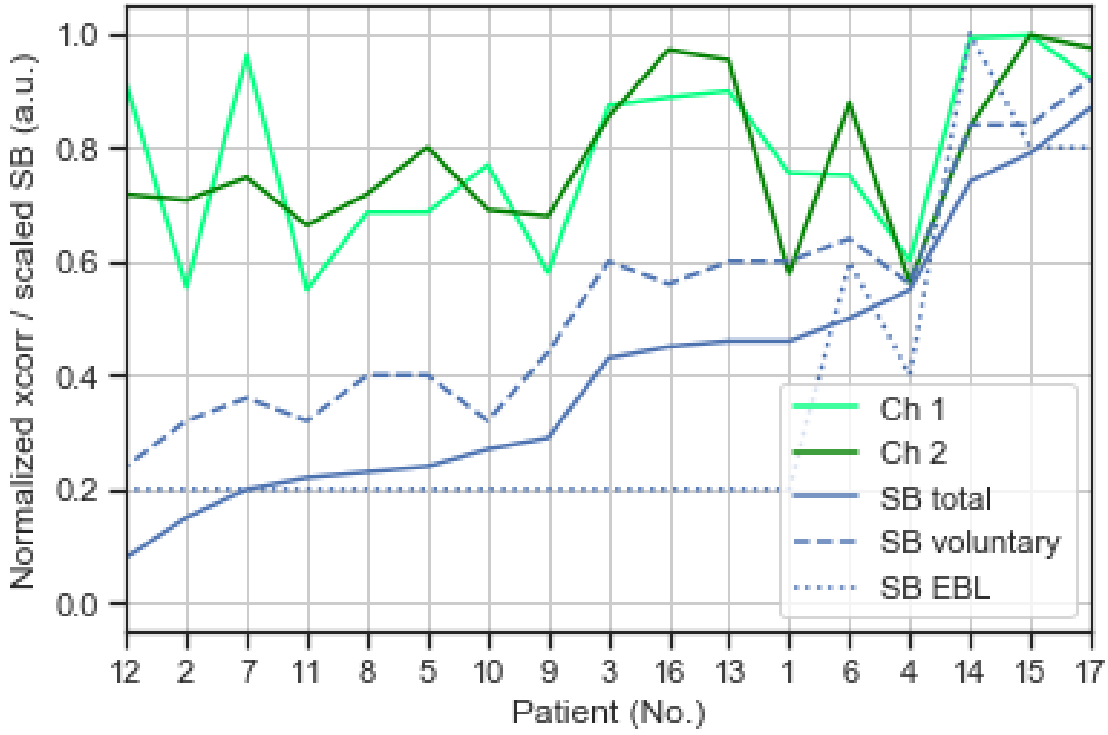

Figure 5.7 *Highest cross-correlation coefficients for eyebrow lift movement over all patients.*

Figures 5.6 and 5.7 show the prototype measurement results againt the relevant Sunnybrook scores over all patients and in the decreasing order of palsy level. In order to evaluate **how well the prototype results correlate with the current clinical method** we can inspect the two aforementioned figures. Figure 5.6 shows that the patient with the most severe palsy grading with Sunnybrook (patient number 12) has lower prototype results in all channels than the healthiest patient (patient number 17). However, the patients between these extremities have fluctuating prototype results. This is true for the eyebrow lift figure, too. Therefore, with naked eye it is challenging to evaluate the entity; how well do the prototype results and Sunnybrook scores correspond? To offer necessary insight to compare the prototype and Sunnybrook results, Figures 5.8-5.11 are provided.

Figure 5.8 visualizes the prototype **smile results** for the channels 4-6 against Sunnybrook scores. The data in the figure is same as in Figure 5.6 but in different format; here, a **pair plot** is used. On x-axis, there is a slot for each relevant channel, and on y-axis there is a position for the picked Sunnybrook scores. The most general Sunnybrook score, the total, is on the upmost row whereas the most specific Sunnybrook value is at the lowermost row; the generality level decreases when going down the y-axis. The pair plot then scatters the prototype results against Sunnybrook values; each channel against every Sunnybrook score.

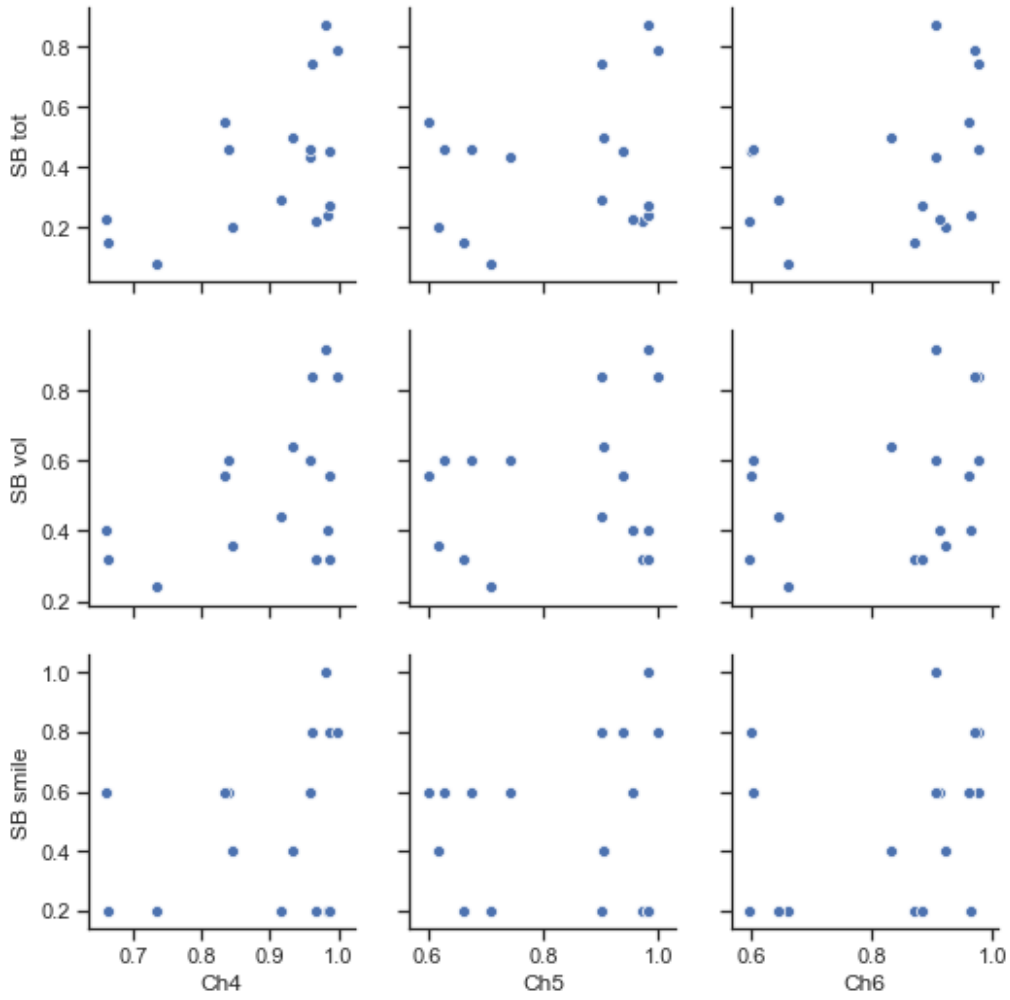

Figure 5.8 *Pair plot of highest smile cross-correlation coefficients against Sunnybrook values.*

Figure 5.9 offers even higher level representation of the data; it lays out the correlation between the prototype measurements and Sunnybrook data. The figure is a **heatmap** where each element of the matrix shape format correspond to the previous pair plot. The darker color indicates increased correlation level. The correlation coefficient is also numerically visible in the heatmap. If inspected channelwise, the channel 4 data shows weak to moderate correlation with the Sunnybrook data. Channel 5 measured values do not correlate, or correlate moderately. Channel 6 measured values correlate on moderate level. Each channel correlates stronger with the most general Sunnybrook score, the SB total score. When the generality of the Sunnybrook score decreases, so does the correlation coefficient. In general, channel 4 has the highest correlation with Sunnybrook values, and its neighbor channel 5 the smallest.

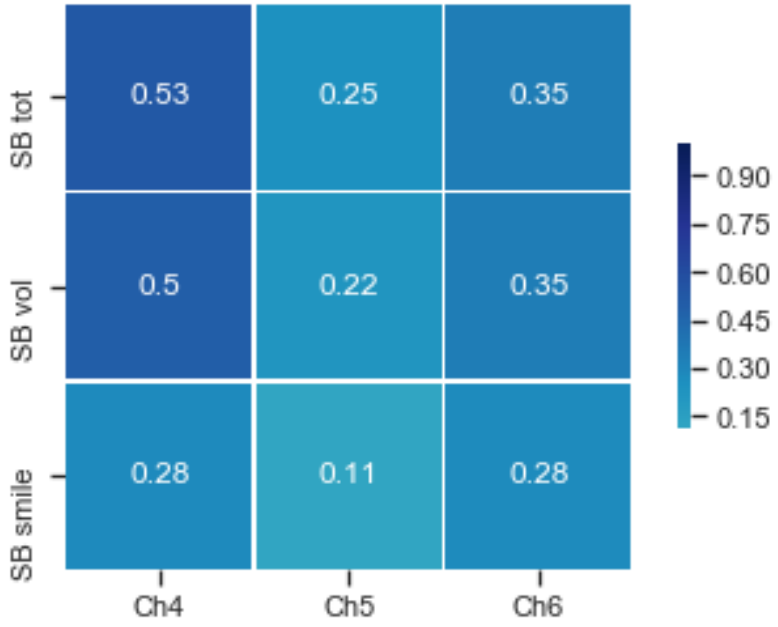

Figure 5.9 *Heatmap showing the correlation of prototype results and the Sunnybrook values for smile movement.*

Similarly, Figures 5.10 and 5.11 visualize the **correlation of prototype eyebrow lift results against Sunnybrook scores**. The following figures are ordered in the same manner as the two previous ones for smile results. The differences are the choice of channels, for eyebrow lift the channels 1 and 2 are used, and the most detailed Sunnybrook score is the eyebrow lift score instead of the smile score.

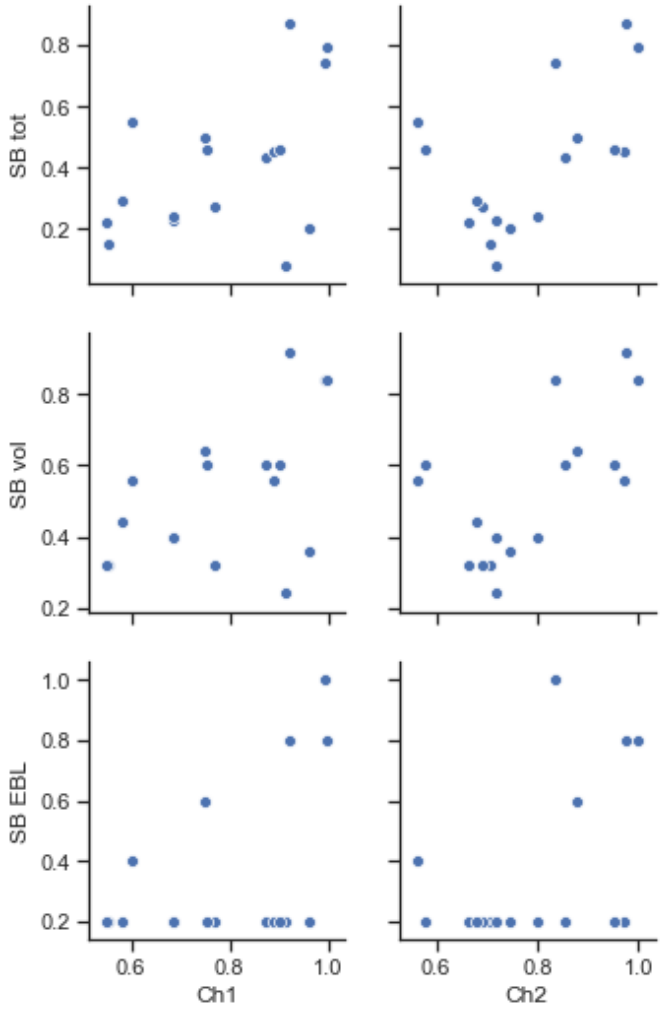

Figure 5.10 *Pair plot of highest eyebrow lift cross-correlation coefficients against Sunnybrook values.*

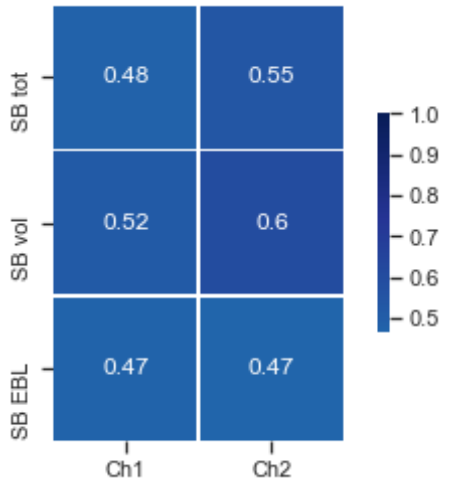

Figure 5.11 *Heatmap showing the correlation of prototype results and the Sunnybrook values for eyebrow lift movement.*

Reading the **eyebrow lift heatmap** from Figure 5.11 channelwise, it can be seen that for channel 1 the correlation to Sunnybrook scores are moderate. For channel 2 the correlation between the prototype data and Sunnybrook values are moderate, or even close to strong. Thus, channel 2 has slightly higher correlation with the Sunnybrook values than channel 1. The **pair plot** in Figure 5.10 reveals that for several patients (12 out of 17) the eyebrow lift score is the same despite the different Sunnybrook total score. This flat line shows in the correlation heat map too; the lowermost row of the heatmap has the weakest correlation. The more general Sunnybrook values, the voluntary movement and total score, have higher correlation with the protype data.

To **summarize** the waveform or direction of vector of excursion results over all patients, there is weak to moderate correlation between the measured data's waveform and Sunnybrook values for smile movements. The correlation increases with the generality of the Sunnybrook scale. The most central channel, channel number 4, has the highest correlation. For eyebrow lift movements the correlation is moderate. The most specific Sunnybrook value has flat line in the eyebrow lift pair plot that shows as the lowest correlation in the heatmap as well.

5.3 Temporal Difference: Delay vs Sunnybrook

This section provides the **results of the contralateral temporal difference** for both movements, smile and eyebrow lift. The used parameter to investigate the temporal difference is the delay of the cross-correlation coefficients. This is shown in Table 5.1 and discussed in Subsections 4.5.4 and 4.6.3. The cross-correlation coefficient, or waveform, results are described in the previous Section 5.2 and thus that section is closely related to the current one. To avoid redundance, the previous section is referred on several occasions in this section. The temporal contralateral difference is a notable factor when the symmetry of human facial expressions is evaluated, this is explained further in Subsection 2.4.2. Similarly to the previous section, the delay results are compared against Sunnybrook (see Subsection 2.3.3) values.

Figures 5.12 and 5.13 visualize the **smile and eyebrow lift movements' temporal difference results** over all patients and against Sunnybrook results, respectively. The figures are created in the same way as Figures 5.6 and 5.7 in the previous section. The only alternations made to this section is the usage of the best correlation coefficients' *delays*, not the coefficients themselves.

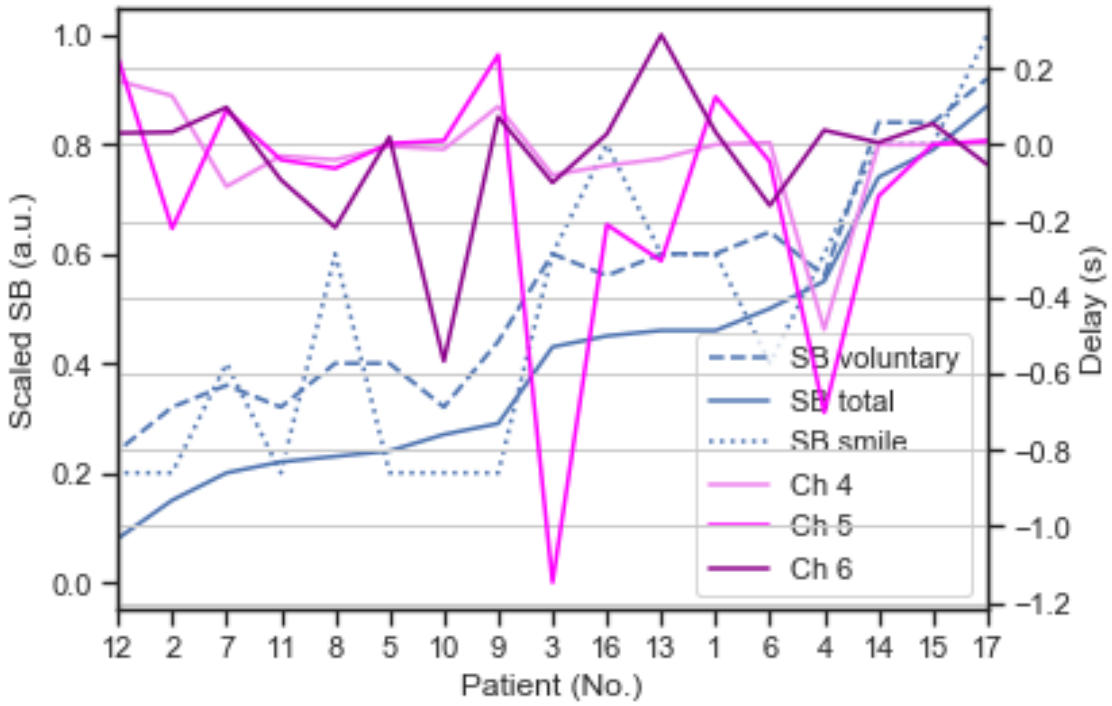

Figure 5.12 *Contralateral delays for smile movement over all patients.*

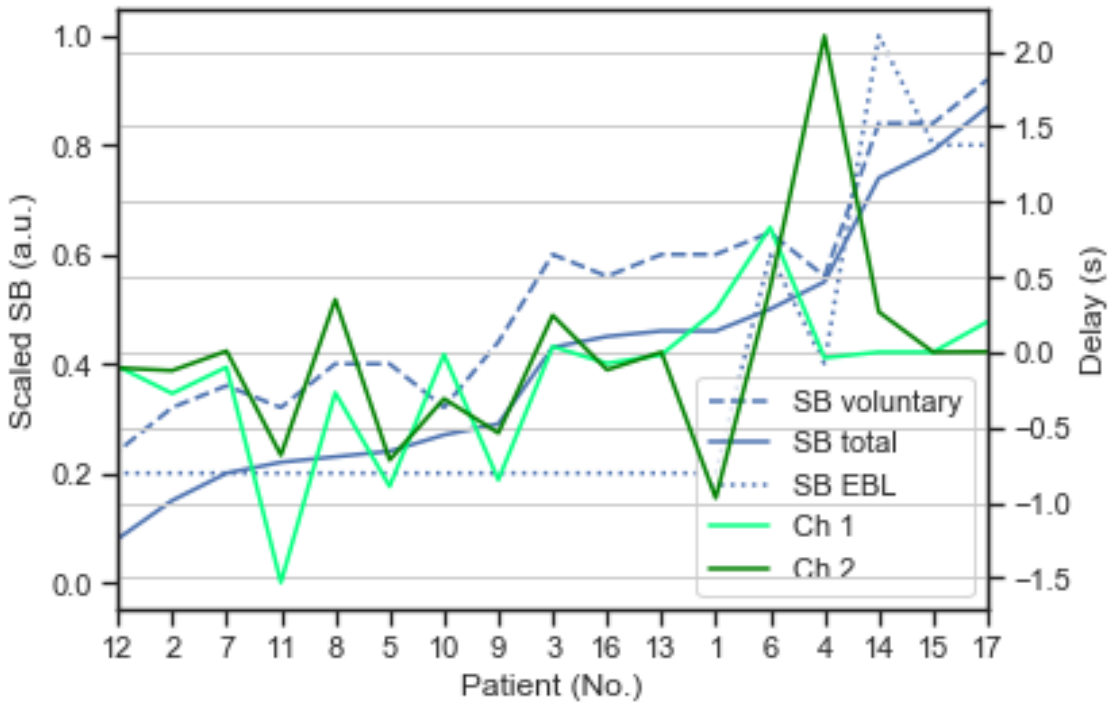

Figure 5.13 *Contralateral delays for eyebrow lift movement over all patients.*

Again, every value is visualized and included in the analysis; no outliers are excluded. This decision is made despite the outliers visible in Figures 5.4 and 5.5. For both movements, the contralateral delays fluctuate around zero seconds, roughly speaking. For smiles the fluctuation range is between approximately -1.2 seconds and 0.3 seconds. For eyebrow lifts the range is wider; it is between -1.5 seconds to 2.0 seconds. Positive delay translates to the palsy side being ahead of the heatlhy side at the measurement site, and negative delay to the contrary. Similarly to the previous section, in order to evaluate **how well the prototype delay results correspond to the current clinical method**, the Sunnybrook scale, further Figures 5.14-5.17 are provided. The pair plots and heatmaps are generated in the same manner as in the previous section with the exception of using delays.

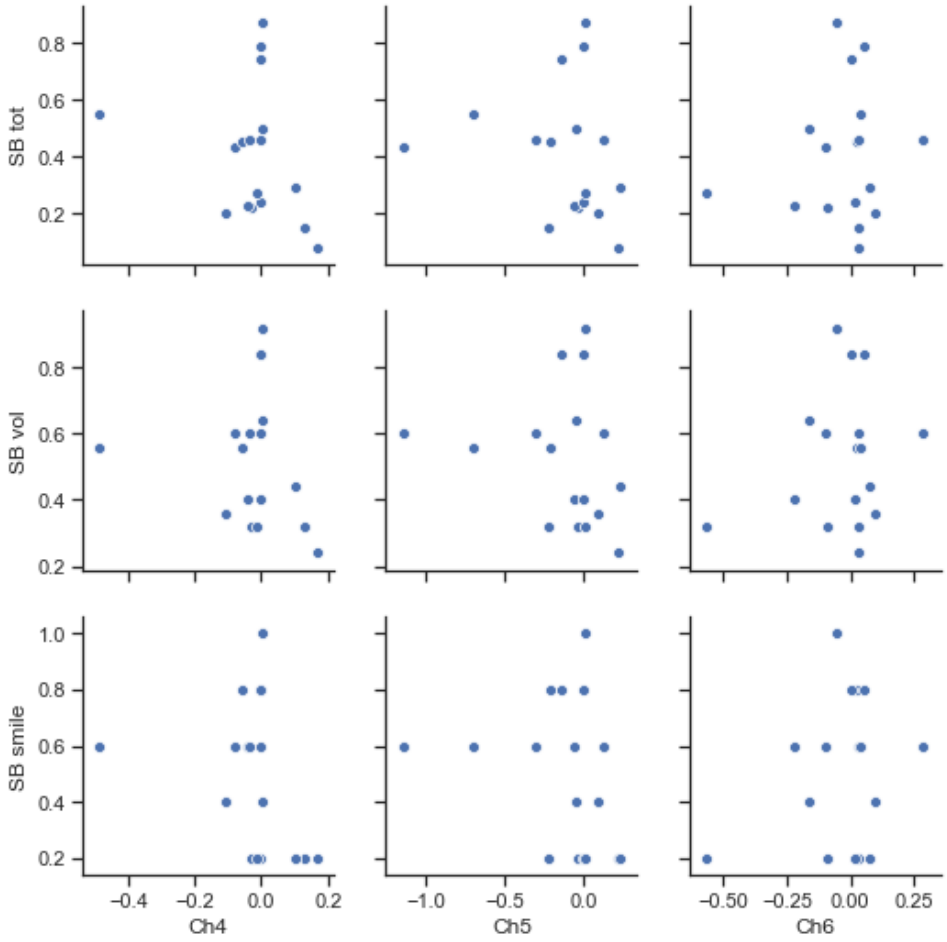

Figure 5.14 *Pair plot of smile delays against Sunnybrook values.*

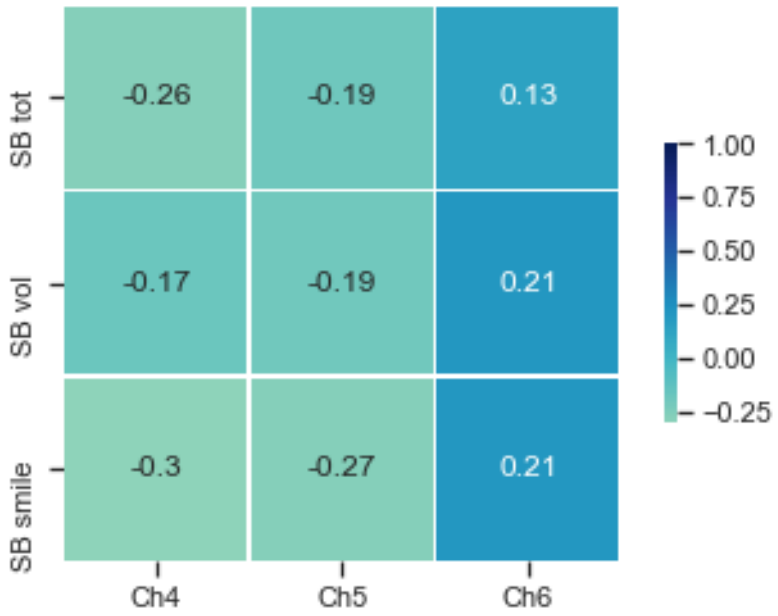

Figure 5.15 *Heatmap showing the correlation of prototype results and the Sunnybrook values for smile lift delay.*

The **pair plot of smile** movement's delays against Sunnybrook scores, Figure 5.14, illustrate the concentration of delay values close to zero. In other words, most of the measured delays between the contralateral channel-pairs have close to zero temporal difference irrespective of the palsy level. The points are clustered to the proximity of zero seconds, and extend vertically depending on the measured Sunnybrook value. The outliers are visible; for example, the plot of channel 4 against different Sunnybrook values have a point beyond -0.4 seconds that clearly differ from the cluster. The prototype measured values' centering around zero seconds is visible from the **heatmap**, Figure 5.15, as well. In general, there is no correlation between measured delay values and Sunnybrook scores; the correlation coefficients of the heatmap are so close to zero. More specifically, channel 4 has almost negative weak correlation with the different Sunnybrook scores. Also, the most specific Sunnybrook value, the smile score, has weak negative correlation for channels 4 and 5.

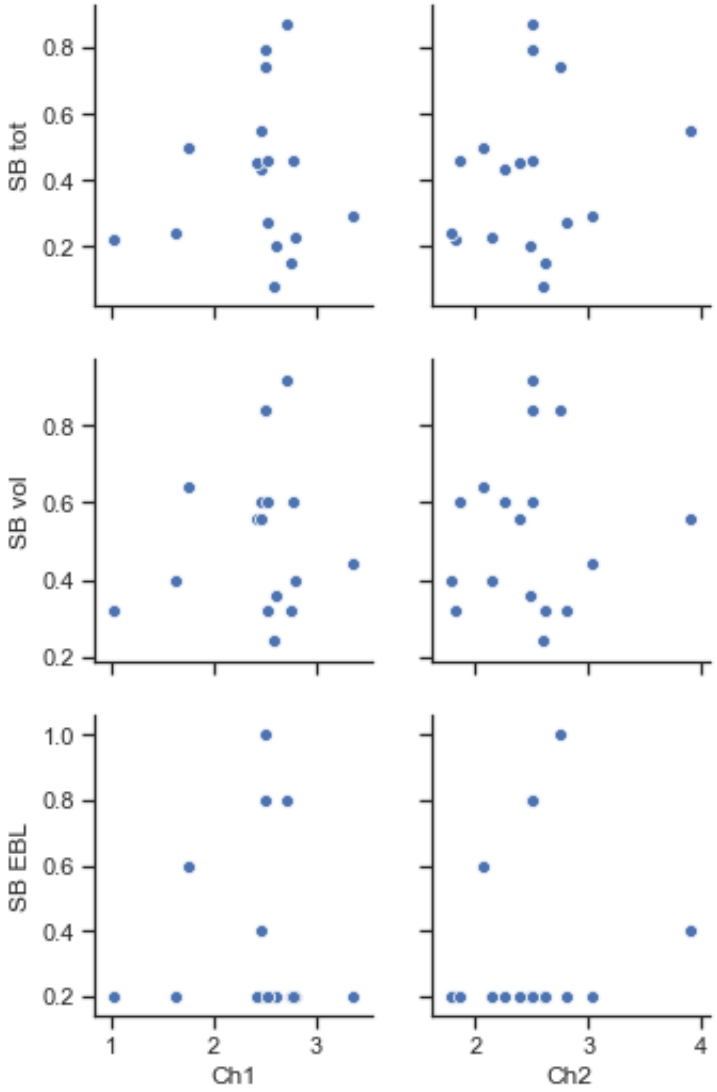

Figure 5.16 *Pair plot of eyebrow lift delays against Sunnybrook values.*

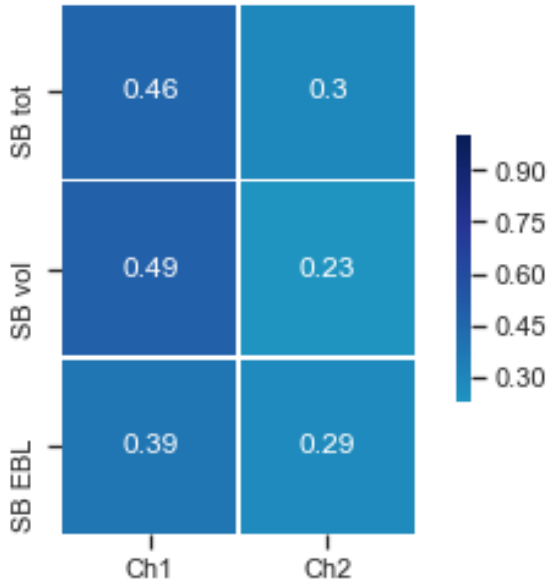

Figure 5.17 *Heatmap showing the correlation of prototype results and the Sunnybrook values for eyebrow lift delay.*

The eyebrow lift movement's delay values are more scattered than the smile movement's. This is shown in the **pair plot for eyebrow lift**, Figure 5.16. The **heat map**, Figure 5.17, then further illustrates the correlation between the measured delays and Sunnybrook controls. The measured values for channel 1 have moderate positive correlation with the Sunnybrook scores. The channel 2 data has weak positive correlation with the Sunnybrook values.

To **summarize** the delay results over all patients, there is no correlation in general found between the prototype data for smile movement and Sunnybrook scores. For eyebrow lift movements there is moderate or weak positive correlation found between the measured data and Sunnybrook controls. The results are discussed in the next chapter.

6. DISCUSSION

The current chapter provides the discussion of this thesis. The chapter is divided into three main sections; to begin with, the results are considered in Section 6.1. Next, in Section 6.2, the limitations of the used approach and sources of error are evaluated. Finally, the future steps are contemplated in Section 6.3.

6.1 Findings of Results

The starting point for this thesis was the **research gap**; there is a need for an objective standardized rating scale that would enable repeatable and comparable palsy level results. The gap is discussed more profoundly in Section 4.1 and the demands set for an ideal system in literature are collected together in Table 4.1. Currently, the lack of a universal objective system is holding the scientific development back; the myriad of methods cause problems in comparing the results, and the subjectivity evokes intra- and interobserver variability. Thus, the gap is the basis for setting the research question of studying if the prototype could be used to determine the facial palsy level. To summarize, the results to be discussed next arise from a **real-world need** and are gained by **a novel method**.

The **approach** to answer the research question utilizes the **build-measure-learn feedback loop** described by Eric Ries. **The first cycle** demonstrates that it is possible to statistically distinguish between symmetrical and asymmetrical smiles and eyebrow lifts. The first loop involves healthy participants and behaves as a very general PoC to avoid unnecessary patient measurements, and to test out the measurement protocol and set-up. **The second cycle** includes patient measurements, and developing a software to provide the means to analyze and later measure patient data. The overall goal of the work is towards an objective rating system. However, during the second loop it becomes clear that the solution strategy needs to concentrate on the data analysis solution instead of being software development centered. Thus, the technical pivot point steers the work away from the objective of this thesis and prioritizes the research question. The following **third feedback loop** then aims to analyze the patient data with agile PoC approach to provide the

required information to answer the research question. The results of the final loop are represented in Chapter 5. This current section focuses on discussing those third loop patient data analysis results.

Due to the novelty of this method, a **literature review on how symmetry is detected** and perceived is necessary. Section 2.4 covers the factors which the existing research identify to affect asymmetry detection. These factors form the basis for the data analysis choices; what parameters to analyze and why. In addition, the mentioned section gathers together the static and dynamic thresholds (see Tables 2.8 and 2.9, respectively). In other words, what are the limits for different facial parts to be observed as asymmetrical. The dynamic thresholds and the temporal protype results are discussed a bit later in this section.

Subsection 4.5.2 details the chosen parameters and their purpose. During the BML feedback loops, the analyzed parameters describing the chosen quantities change to more robust ones. These changes are discussed in the perseverence point (see Subsection 4.5.4) and the pivot point (see Subsection 4.6.4). To recap, the patient data analysis employes the waveform comparison to describe the magnitude of the **vector of excursion**. The waveform comparison is completed by inspecting the **cross-correlation coefficients**. Studying the direction of the vector of excursion is neglected. The importance of vector of excursion in perception of asymmetry is dicussed in literature and summarized in Section 2.4. The **temporal difference**, or the dynamic threshold described in Section 2.4, is computed from the **maximum cross-correlation coeffiecient's delays**. This is more robust method than comparing the contralateral maximum amplitude's temporal locations; the paralyzed signals may have very irregular shapes, as seen in Figure 5.1. For this same reason, the spatial maximum, or static threshold from Section 2.4, is not computed for the patients from the maximum amplitudes eventhough it was computed for healthy participants.

To conclude, the basis for the analyzed quantities is based on literature; how are asymmetries perceived. The quantities are investigated by certain parameters that needed to be changed to more robust ones along the BML feedback loops. However, the quantities behind the parameters remained constant during the study. The patient results concentrate on the waveform (magnitude of vector of excursion) by analyzing the cross-correlation coefficients, and the temporal difference (dynamic difference) by inspecting the delays of the maximum cross-correlation coefficients. The details of those parameters' results are discussed next.

The obtained Sunnybrook data is used as a reference; the measured data is compared against the Sunnybrook values. Vrabec et al. [20] advise that *"[d]evelopment of any facial movement scale is incomplete without some form of validation"*. The completed steps are by no means enough to *validate* the current approach, but provide the means to answer the research question; the measured values need a point of comparison to evaluate the prototype's capabilities for the intended purpose. Thus, the research question is addressed by comparing the prototype results to the Sunnybrook values. This can be considered as a step towards validation; if the conclusion was encouraging, further studies would be suggested to validate the approach. As a note, in that case the demands represented in Table 4.1 should all be studied. The Sunnybrook scale is chosen for reference, as it is currently the most viable scale option (see Section 2.3) and the choice commonly used in Finland.

With this set of data, measured from 17 patients, the parameter describing the **waveform** of the measured data has a weak or moderate positive correlation with the Sunnybrook data. In other words, the prototype data corresponds to the current de facto method in a weak to moderate manner. The **smile** movement's correlation values are the greater the more general (the generality increases when moving from single movement's score to voluntary movements' score and finally to the composite score) the Sunnybrook score is, at the highest the correlation is moderate. Also, different channels have different correlation levels; channel four has the best correlation level, the next best is channel six, and finally channel five. Thus, the most central sensor has the highest correlation, there is a drop when coming to channel five, and an increase when considering the lateral sensor number six. This clear drop in the middle sensor five is a peculiar one. It would be expected that the sensors very close to each other would provide similar results; smile movement includes wide areas of the face and it is surprising to see a major change in such a short distance.

The **eyebrow lift** movement's correlation values are consistent with each other; in general, the eyebrow lift waveform correlates with the reference Sunnybrook values in a moderate sense. The channel number two correlates a bit better than channel number one. The eyebrow lift movement has a stronger correlation with the Sunnybrook values than the smile and less variance. A possible reason could be that the eyebrow lift is a more **localized movement** in the face. Therefore, it is easier to place the sensors to the corresponding locations. Also eyebrows helped the positioning the uppermost extensions. The **fitting of the prototype** is a factor, too. There is only one prototype of a fixed size. The observations made by naked eye suggest that there was significant variance in the fitting, especially with the middle extensions. For some patients the extensions reached much more centrally than for others. This translates to not measuring from the same location for every patient -

and yet comparing the values as if they were. It is possible that the fitting of the prototype is visible with the eyebrow lift waveform results as well; neither was the fitting of the uppermost extension identical between the patients. However, this is speculative.

Moving on to the **delay** results, for this measured data set there is either no correlation (smile) or a weak to moderate positive correlation (eyebrow lift). Inspecting the pair plot of the **smile** delays (Figure 5.14) reveals the reason for smile movement delay's poor correlation to the Sunnybrook results. The reasons are already stated alongside the mentioned figure; the smile delays seem to fluctuate very close to zero. Thus, if excluding the outliers the pattern of the smile delays is approximately a vertical line whereas the palsy level, or the Sunnybrook values, increase from left to right. To mention, the outliers were left into the analysis to keep the data set big enough for a conclusive analysis. The outliers are not identified to be originate from an experimental error, rather the variability is just high.

Considering the moderate *positive* correlation that the **eyebrow lift** delays have with the Sunnybrook results, the outcome is to the contrary what was expected. Increasing Sunnybrook score indicates healthier facial status, or improved symmetry. Thus, when the Sunnybrook score increases, one would expect the delay to decrease. This would yield a *negative* correlation between the Sunnybrook scores and delay. However, the results suggest the opposite. It seems contradictory that when the palsy level decreases the contralateral temporal difference increases. The only reason found to describe this phenomena in the available data, is the greater variance of the eyebrow lift delays than when compared to the smile delays. However, that does not fully explain the event.

The delay results for both smile and eyebrow lift can also be compared to the **dynamic thesholds** gathered into Table 2.9 in Subsection 2.4.2. The threshold for detecting an asymmetry for a slow eyebrow lift is 100 ms, for fast eyebrow lift 67 ms, and for smile 67 ms, according to the table. For **smile**, by just observing the graphs (see Figure 5.12) it can be seen that for multiple patients the threshold of 67 ms is exceeded. The same holds for the eyebrow lift (see Figure 5.13) and the thresholds of 67 ms and 100 ms. However, a further analysis concentrating on comparing the dynamic thresholds to the gained results would be needed to state a concrete outcome instead of just commenting on that the threshold seems to be exceeded on multiple cases.

When **interpreting the results**, both waveform and delay, it is important to note that the used method is based on **linear** correlation; the analysis is restricted to study only the linear dependency between the signals. In other words, if the results suggest that there is no notable correlation it may be due to a) the method not working or b) that there is no linear correlation as a phenomenon present in the first place. It is possible, that for some other model would provide higher correlation with the measured data. In Subsection 2.3.3 it is stated based on literature that the Sunnybrook scale has the most significant variability in the mid areas of the composite score range. In other words, the mid areas of the total Sunnybrook score are most prone to variability according to the used reference of [29]. This is important to acknowledge when interpreting the results; neither is the comparison point, or **oracle**, free of error. Finally, the correlation study employes a simple form of correlation, the linear one, as the de facto approach to data analysis to begin with a plain method.

Another factor to note when interpreting the results is the **inter-patient approach**; no intra-patient study enabling measurements are made as each test participant is measured only once. Also, the prototype narrows the inter-patient study to only to limited points and channels. The limitations and sources of error are further discussed in the next section.

6.2 Limitations and Sources of Error

This section evaluates the **sources of error**. Few error sources are mentioned already along discussing the results in the previous section. Those factors are extended and covered more throughly in the current section. Also, the **limitations** of the method and this research are considered here; how are the demands listed in Table 4.1 met.

As noted in the previous section, **placing the extensions and sensors** is a source of error. To begin with, the extensions and sensors should to be placed completely symmetrically as the data analysis assumes identical contralateral locations in the single patient-level. The result for each patient is gained by comparing the right and left side. If the measurement points are not the same, a distortion is caused to the results. Secondly, the inter-patient analysis expects the data to be measured from corresponding locations. To summarize, placing the extensions and sensors to identical contralateral locations is needed in both, the patient and inter-patient, levels.

There are multiple factors that hamper fulfilling the requirement of identical contralateral sensor-positioning. Firstly, the sensors and extensions have three degrees of freedom (DoF). Secondly, severe facial palsy may cause the affected side to appear

deformed, and thirdly, each person's facial structure varies a bit. Finally, a human adjusts the sensors and extensions to their places manually. These factors, especially the last one, introduce **subjectivity and human-based error** in intra-adjuster and inter-adjuster level. In other words, the intra-adjuster error arises from a person's incapability to fix the prototype to exactly same locations between different gos, whereas the inter-adjuster error is due to differences in adjustment between various adjusters.

Another source of error mentioned in the previous section is the **fitting of prototype**. It affects the results at least in three levels. As discussed in the previous section, one of the mechanisms is the different level of sensor extending between the patients. This is closely related to the pondering above concerning the same measurement spots within a patient and multiple patients. In addition, the prototype shifted in some of the patients head during strong movements. This causes directly error to the results. Finally, a major factor with the fitting of the prototype is the touching of the lateral sensors and the skin. In other words, for some healthy test participants the prototype could not be adjusted in a manner to allow enough space between the face and the sensors. In the perseverence point (see Subsection 4.5.4) this problem is addressed by shutting down the most lateral sensors for the patient measurements. This action also allowed higher sampling frequency. However, there were still a patient that had a recording sensor touching and couple of patients whose prototype fitting had something to report (see Appendix D).

There are also other sources of error, limitations, and factors that affect the results than the ones mentioned in the previous section. One of the other factors is the relationship between **the direction of the measurement** and movement; the capacitance measurements are based on the change in distance *perpendicular* to the face whereas the actual movement is a 3D movement *along* the facial surface. In the technical background section it is summarized from literature [48–50] that 2D systems should be excluded from facial palsy level assessment (see Section 3.1). To briefly recap from the theory, the exclusion is based on lost information when a 3D movement is projected to 2D plane; especially the mouth's anteriorposterior movement is ignored with 2D systems [48, 50]. Thus, the 2D amplitudes are measured to be significantly smaller than 3D motion amplitudes [48, 50]. The 3D results also correlate better with the used clinical scales than the 2D systems [50].

In the current study the movement is not projected to 2D as the extensions are set to follow the facial surface. However, due to the novelty of this research - applying capacitive approach to measure symmetry and facial palsy level - there is no knowing if this method loses information due to the measurement direction and/or produces

smaller amplitudes than it should. Also, there is no telling how much the relationship between the direction of the measurement and movement affects the correlation; to the best of knowledge, capacitive approach has not been compared to a clinical grading scale such as Sunnybrook before. Therefore, the goodness of the capacitive approach in comparison to these mentioned factors, and for example to the 3D systems, remains an open question.

Another limiting factor of this study is the **limiting the measurements to few points**. This is also a source of error; are the points on corresponding contralateral locations within a patient, and are the locations same when conducting inter-patient analysis. However, that is already discussed above and the focus is now in limiting the amount of data. For the sake of this discussion we could assume (eventhough it is unjustified) that the points are perfectly identical within a patient and multiple patients. The prototype can be equated with landmark-based methods as the measurements are concentrated on pre-defined locations; points. Thus, in this current study, the very same arguments aimed towards landmark-based approaches apply. In the technical background chapter (see Section 3.2 and specifically Subsection 3.2.3) the criticism towards the landmark-based approaches is collected from literature. The baseline of the criticism is limiting the analysis to certain individual points [47, 51]; the approach might provide enough information, but modern technologies could provide more insight [51, 54]. Also, some areas such as nose are too complex to be analyzed with few landmarks [51]. This is a relevant consideration as one of the clinically used movements to assess the facial nerve's different branches' functioning is wrinkling the nose (see Section 2.2).

To discuss the limitation and this study in more detail, in Section 2.2 six movements are listed to be performed to evaluate the differential movement of facial nerve branches. **Two movements, smile and eyebrow lift, are used** in this research. If the more complex movements such as wrinkling the nose and lip puckering were to be investigated, problems could possibly arise. To begin with, the extensions in general would not reach over these areas thoroughly. If this could be fixed, the nose for example could still be too complex to be analyzed based on arguments that Alqattan et al. exposed in [51]. Additionally, in the dissertation [6] in which the prototype was developed, the eye closure movement appeared to be a bit problematic in the classification, detection and locating of the movements (see Section 3.6). To conclude, whether there are movements that are too complex to be analyzed with the prototype cannot be concluded for certainty based on this research. However, limiting the measurements to certain locations, or points, could be a factor affecting this study's results significantly.

The **effects and limitations of the research approach** should be discussed as well. In retrospect, it was a well-grounded choice to begin with a preliminary study with healthy participants (Section 4.5). Perseverence point offered multiple improvements to be considered before the patient measurements. However, as the scientific knowledge does not exist yet, the work flow chosen for the second BML feedback loop (Section 4.6) was too productization-oriented. In other words, having both an objective and a research question for a thesis complicated getting results. The objective directed the second BML loop towards TDD, comprehensive unit tests, UI design et cetera. This resulted in concentrating on implementing a software instead of focusing on answering the research question. The priority should have been on the research question due to the novelty of the measurement method; there is no use for the software unless the research question can be answered 'yes'. The pivot point (Subsection 4.6.4) addresses this issue. The third and final loop (Section 4.7) applies agile techniques that allow fast and easy changes to the analysis, and data exploration in general. The third BML loop is the PoC with the patient data. In the final loop, the focus shifted from fulfilling the objective to answering the research question as a priority. To conclude, the lean method worked well in this thesis, but the starting point of having an objective and a novel research question caused challenges.

Finally, Table 4.1 collects the **demands for a facial palsy grading system** from literature. In Section 4.2 six out of the thirteen demands are left outside the scope of this thesis, one stated to be the limitation of the method, two demands analyzed mathematically, and five factors left for discussion. The two analyzed demands are the capability of dynamic measurements and quantitativity. Those demands are covered in the previous Chapter 5 and Section 6.1.

The five demands left to be discussed here are minimal invasiveness (demand 7), clinical convenience (demand 8), cost effectiveness (demand 11), fast to use (demand 12), and the requirement not to utilize plenty of equipment (demand 13). These demands are not included into the limitations and sources of error, but are covered here alongside the other demands.

The requirement of **minimal invasiveness** (requirement 7) is expected to be fulfilled. The measurement only requires wearing a cleanable headset, without the branches touching the face, and performing facial expressions. For a patient's point of view, the measurement can be perceived as wearing headphones and making facial expressions. The approach can thus be claimed to be minimally invasive.

The requirements of **clinical convenience** (demand 8), **fast usage** (demand 12), and **not requiring much equipment** (demand 13) are closely related to each other. The clinical convenience consists of factors such as the amount of equipment and the time needed for the measurements. Other factors are for example the level of needed education prior to conducting the measurement, and if special conditions are needed. To begin with demand 13, the extent of equipment is small: only the prototype, a computer with a bluetooth radio, a chair, a table, and an external screen are needed. Basically the only special equipment is the prototype itself, the other means are regular accoutrement of a doctor's room. Neither are special lighting or setting of mirrors and cameras to exact locations needed, to compare to methods covered in Chapter 3.

Demand 12, fast usage, is also a relative term. During the research the actual measuring phase took 20 minutes and the preparations took another 20 minutes that included taking the reference Sunnybrook, explaining and practicing the protocol - and adjusting the prototype. The total duration is significantly more than running a Sunnybrook test (couple of minutes) but not much more than taking for example an MRI. However, it should be noted that this study is the first investigation with this method and the optimization of the measurements timewise would be a future task. Finally, the amount of education is also dependent on the system's level of maturity. The existing software (Section 3.6) is not specialized for the purpose and requires thus more knowledge that what to adjust than the designed UI (Appendix F). Adjusting the sensor extensions should be straightforward to physicians with anatomy knowledge. To conclude, the system does not require much equipment or special conditions, can be used already now with acceptable speed and the required time could be probably optimized in future, and thus can be argued to be clinically convenient.

The requirement of **cost-effectiveness** (demand 11) is left to discuss. In general, medical imaging costs much. However, a multi-purpose system such as MRI or X-ray can be more expensive than a single-purpose facial grading system. Based on this study, the moderate level of equipment and no special environment requirements support the argument of probable low cost. Answering the question of cost-effectiveness in detail is a question for future and productization.

6.3 Future Aspects

This section builds on the two previous sections that discussed the results (Section 6.1), and the sources of error and limitations (Section 6.2). Also the entire preceding thesis as a basis, it is possible to **discuss the research question, objective, and next steps**.

The research gap exists because no system can answer the demands set for facial grading system (Section 3.5 and Table 4.1). The purpose of this thesis is to investigate if the capacitive prototype could provide the means to measure the severity level of facial palsy, if the prototype could fulfill the demands laid out in literature. The results presented in Chapter 5, and discussed in Section 6.1 provide a *probable maybe* as an answer. However, when considering the limitations and sources of error given in Section 6.2, the outcome becomes *not as such*. In other words, the solution requires changes so that even further studies would be well-advised.

The prototype has similar problems as the existing approaches. As discussed in previous section, subjectivity is introduced to the prototype measurements. The traditional scales such as Sunnybrook employ **subjectivity** in the estimation of the palsy level itself, the prototype involves subjective assessment in the sensor positioning. Subjectivity results in intra- and interobserver variance. The issue with the prototype extension adjustment is also the need for **symmetrical contralateral positioning**. This is required in the patient-level as the result computation assumes identical positions on different facial sides. Also, if the same patient was measured again after certain period of time, the results should be gained from the same location for comparability. In the interpatient level, the locations used for multiple patients need to be equivalent for the same reason.

In addition to subjectivity, and the variance and error followed by it, the prototype has the issue of **limiting the measurement to scarce amount of points**. This is typically the problem of landmark-based systems (Section 3.2). Whether or not the prototype is actually limiting necessary information is a debatable question. However, for example Figure 5.9 visualizes a peculiar drop in the correlation coefficients in the channel five, as discussed in Section 6.1. In other words, the results are not that excellent that there would be support for the argument of still recording enough information even though concentrating on certain points.

Thirdly, it can be speculated if the prototype's **direction of the measurement** is problematic. The capacitive measurement is conducted perpendicular to the facial surface whereas the actual movement is a 3D movement along the facial surface. This

may potentially lead to lost information as happens with 2D systems. The 2D systems are excluded from facial palsy studies because they underestimate especially the anteriorposterior-component. However, unlike with the limiting the measurement to scarce amount of points, there is no definite indication in the results that information would be lost due to the measurement direction. Therefore it is just stated, that this factor should be further studied.

Finally, the **fitting** of the prototype needs to be addressed. As discussed in the previous section, the prototype could be moving in relation to the patient's head producing error. Also, as there is only one-size-fits-all type of solution, the extensions may reach the needed spots for all patients or produce data from different locations when comparing over patients. The *reaching to the necessary facial structures*-issue may become pronounced if all movements used in clinical evaluation (Section 2.2) would be involved in a study.

To conclude, it is plausible that the prototype has error and variance originating from subjective extension adjustment, and fitting. In addition, the measuring is likely limited to too few points. The effects of the direction of the measurement should be further studied. Thus, changes to the system would be required.

In order to overcome the subjectivity of placing the sensors, the **adjustment should be automated**. For previous objectives presented in dissertation [6] the manual adjustment accuracy may have been sufficient, but for medical diagnostic purposes the placement should be deterministic and repeatable. Eventually, the grading scale would be used to determine if an invasive intervention should take place, or to assess how the effects of an operation to compare methods within a possibly global research community. Eventhough just a PoC has been implemented, the view should be extended to product phase when considering the viability of the approach.

Another matter to address is the fitting of the prototype. To intervene with the prototype shifting along a movement, not extending to same extent with all patients, and having too few measuring points that sometimes touch the face, **the prototype should be modified**. Concurrently, all of the six movements used in clinical assessment (Section 2.2) should be enabled to be evaluated. A possibility could be similar head-supporting structure that oculists' microscopes utilize. The patient places their chin to a support and lean their foreheads against another. The height of the supports can be adjusted. This stabilizes the head for the oculist to inspect the eyes. The improved prototype could employ the chin-rest as such, but in order

not to limit the forehead movement the upper support could extend from top instead from side-to-side. The supports could be adjusted from the sides (chin) or from above (forehead) and a setting value noted for the patient for future measurements.

The automated sensor adjustment could be combined with the head-supporting structure. Once the patient's head is stabilized to a position, the sensors could be placed to close proximity in front of the face. The number of sensors should be increased to approach a surface-based analysis rather than point-based. The amount of sensors obscure and would require its own study, but should allow the measurement of the complex structures such as nose (discussed in [51]) and the six movements used in clinical assessment (Section 2.2). The basic structure of the sensors could be a net instead of extensions to allow more sensors and higher coverage, or the amount of extensions and their length could be increased. The automated adjustment technology and approach would require its own research to fit people of different facial size and structure.

The improved prototype and its related design and research projects are a strategy of a longer term. The first step should be an **additional analysis** with the existing data to deduce if the prototype should be modifed to specialize to this medical diagnostic purpose. In other words, more information is needed to decide if this approach should be further developed. To begin with, the parameter which is not computed in the scope of this thesis, the derivatives to assess the size of the vector of excursion, could be calculated. Also, the correlation between the Sunnybrook results and the measured data could be tried with other than linear model; the analytical solutions could be developed and researched in general. Other demands from the requirements (Table 4.1) for grading system could be investigated from the data as well. However, many additional analysis feature would require further patient measurements, for example demand number four; reproducibility. Also, some requirements such as demand three, synkinesis evaluation, would very likely need the improved prototype; the likelihood of the synkinesis movement hitting a recording sensor with the current layout is small. To conclude, some further analysis could be done with the existing data but it is unclear how much extra information that would provide.

Another aspect to consider is the objective of this thesis; determining how to develop a software to enable further studies. In the pivot point (Subsection 4.6.4) that took place after the second BML feedback loop, the priority shifted from the objective to the research question. This is dicussed in Section 6.2, too. However, if an improved prototype was set to be the goal, the objective should be fully addressed. The **software should be finalized** (Appendices E and F lay out specifications and

design, respectively), and extra features such as combining the video data into the application could be deliberated. One very clear advantage the mature one-purpose application could provide is the **handling of outliers**. If the problematic repetitions could be marked already in the measuring phase (for example, conducted a wrong movement) and an overview of the results shown at the end of the measurement, more repetitions could be measured in the first place.

An alternative to the prototype improvement and further analysis is **a technical pivot**; to change the approach to measure the severity of facial palsy. That is an option that should be carefully considered; adding another compromised method to the myriad of existing methods would not do any good or enhance the medical development to understand and treat facial palsy. A possibility could be to utilize deep maching learning techniques such as CNNs to detect the palsy level. Before beginning the work of this thesis, those approaches were excluded as a boundary condition; a deterministic approach was a must. However, let us take a moment to discuss the alternatives.

Section 3.4 covers methods of CNNs and FACS and lays out examples of both. The **CNNs** have been used to classify facial palsy from 2D images employing the House-Brackmann range [67], as covered more thoroughly in Subsection 3.4.1. In brief, a strength that a deep learning method introduces is the automated feature selection. Sajid et al. [67] suggested that hand-crafted methods could choose wrong features to the analysis and thus decrease the accuracy. It is an interesting thought; in this thesis the **parameters are hand-picked** based on literature about how does a human perceive and detect symmetry. Thus, as the correlation between the reference method and the measured data remained moderate at its best, one can list hand-picked parameters to be a possible result impairing factor.

Automated FACS is covered in Subsection 3.4.2. To the best of knowledge, no solution that could be utilized to grade facial palsy or seen as an MVP towards it, is available. Martinez et al. [103] concluded that detection of AU segements and their intensities still remain an open challenge. Romero et al. [92] have applied a novel CNN-based solution that they named AUNets to detect AUs from multi-view videos. The method employed optical flow layer, and several different detectors that provided a probability for 12 different AUs. They listed a large number of parameters and a long training time to be the limitation. Martinez et al. [103] have also highlighted the need for open-source databases.

Automated neural network-based facial palsy grading system would require an adequate (open-source) **database** of 3D faces. The **data source** is an important matter; Liu et al. [91] used a geometric model, but as there are commercial IR cameras widely available (Subsection 3.3.3) the IR-data might be an option for the input. Also, computational power to train the 3D NNs cannot be overlooked. Another central question is that **what is the neural network taught**. The question should be studied and given a well-reasoned answer. The options include inter alia utilizing the FACS system (intensity, dynamics and symmetry of AUs), using the existing grading scales such as Sunnybrook, detection and perception threshold approach, or something else. Finally, the matters of privacy and ethics come to question when considering a facial database - especially if a pathological condition such as facial palsy is present. However, artificially generated faces might provide a solution to that.

The discussion above does not aim to be a comprehensive review to deep learning approaches and their utilization to grade facial palsy level. The goal is to provide a brief overview for a future aspect suggestion, give alternatives to the prototype solution, and to support the conclusion of this thesis. The baseline is that the current deep learning approaches are not 3D solutions and the lack of a database developed for 3D measurement solutions and thus the lack of adequate data is hindering the development, to the best of knowledge. Neither is the question that what should the neural network be taught researched, as far as we can tell. Finally, the question if the approach should be deterministic should be addressed; the system would not support a human in their decision making but instead provide diagnostic value by grading the facial palsy. That grade is then used in inter alia deciding interventions. The IR camera that is suggested above to be a possible input for the database, could offer a stand-alone solution for palsy grading, too. Finally, an option is to improve the capacitive approach; even though the current system as such is not an ideal form to utilize capacitive method to facial palsy grading, the capacitive method is still a viable option.

Whether the further development of the automatic facial palsy grading system is based on the capacitive approach or to a more modern solution such as deep learning, and possibly involving IR cameras, it is certain that there is a need for such system. The demands listed in Table 4.1 need to be answered to speed up the scientific development to understand and treat facial palsy.

7. CONCLUSIONS

In this thesis the suitability of a capacitive prototype to grade facial palsy level automatically was investigated. Validated learning was applied in the research to proceed from healthy test participant PoC of 20 volunteers to measure 17 facial palsy patients. The BML feedback loop enabled steering and improving the approach along the way. To the best of knowledge, this was the first time capacitive measurements were used to grade facial palsy level.

The research question of this thesis was to study if the capacitive prototype developed in [6–9] could be applied for the facial palsy severity rating system purpose. The research question was addressed by reviewing existing literature on how a human detects and perceives asymmetry. Those parameters were then computed from the measured data by comparing the contralateral channels, and then compared against a reference scale, Sunnybrook, that is used in clinical assessment. The measured data correlated with the Sunnybrook values in waveform from weak to moderate (smile) and moderately (eyebrow lift). In delay, the correlation between the measured data and Sunnybrook was non-existent (smile) and from weak to moderate (eyebrow lift). To answer the research question; these results do not provide evidence that the investigated capacitive prototype method would be suitable as such to be applied to rate facial palsy severity.

The objective of this thesis was to determine how to develop a software to enable further studies. The focus shifted from the objective to the research question in the pivot point after the second BML-loop. At the pivot point, it was acknowledged that the objective holds a presumption that the research question could be answered positively. In other words, if the research question was answered "no" there would be no need for a software. Thus, after the pivot point the emphasis of the work became to answer the research question; the objective is mainly targeted in the appendices. The technology pivot towards fast prototyping and a patient PoC also enabled data exploration for the patient data that is in general a useful approach in solving data-oriented problems. However, the technologies and approaches used to address the objective are viable to be used later on if this capacitive approach is further studied and developed.

Currently, the lack of an automatic grading system hinders the scientific research. The fundamental problems of the existing methods are subjectivity that results in intra- and interobserver variability, and complicates the comparability of studies. The results of this thesis contain problems from the existing approaches. The sensor adjustment is manual and thus introduces a subjective error to the measurement. Also, the prototype limits the measurement to scarce amount of points. These problems could be potentially solved by an improved prototype that would have head-support, automated adjustment, and more sensors. However, as the prototype has a constant limitation of not enabling static assessment, other approaches should be considered. CNNs and automated FACS are alternative analytical solutions, and IR cameras such as Kinect II are an alternative data collection technique. The lack of 3D database hinders the development of deep learning approaches. The deep learning approach would also require answering what exactly the neural network should be taught; the FACS system, the existing grading scales such as Sunnybrook, detection and perception threshold, or something else.

The next step - whether it is the further development of the capacitive approach or a more modern solution such as deep learning, and possibly involving IR cameras - should be carefully considered. The field is already suffering from a myriad of methods, and creating another compromised system does not solve the fundamental need. On the contrary, it adds to the issue. After 35 years of House and Brackmann's article's opening sentence *"[t]he major problem in assessing the results of facial nerve surgery or medical treatment lies in the subjective methods of assessment and reporting"* the problem still remains to be solved.

REFERENCES

[1] Mervi Kanerva and Anne Pitkäranta. "Perifeerinen kasvohalvaus". In: *Duodecim* 122.18 (2007), pp. 2267–2274.

[2] John W House. "Facial nerve grading systems." In: *The Laryngoscope* 93.8 (1983), pp. 1056–1069.

[3] John W House and Derald E Brackmann. "Facial nerve grading system." In: *Otolaryngology–Head and Neck Surgery* 93.2 (1985), p. 146.

[4] Brenda G Ross, Gaeton Fradet, and Julian M Nedzelski. "Development of a sensitive clinical facial grading system". In: *Otolaryngology–Head and Neck Surgery* 114.3 (1996), pp. 380–386.

[5] Adel Y Fattah, Javier Gavilan, Tessa A Hadlock, et al. "Survey of methods of facial palsy documentation in use by members of the Sir Charles Bell Society". In: *The Laryngoscope* 124.10 (2014), pp. 2247–2251.

[6] Ville Rantanen. "Capacitive Facial Activity Measurement". 64 p. Doctor of Science in Technology thesis. Tampere, Finland: Tampere University of Technology, 2014.

[7] Ville Rantanen, Pekka Kumpulainen, Hanna Venesvirta, et al. "Capacitive facial activity measurement". In: *Proceedings of the XX IMEKO World Congress*. Busan, Republic of Korea, 2012.

[8] Ville Rantanen, Pekka Kumpulainen, Hanna Venesvirta, et al. "Capacitive facial activity measurement". In: *ACTA IMEKO* 2.2 (2013), pp. 78–85.

[9] Ville Rantanen, Hanna Venesvirta, Oleg Špakov, et al. "Capacitive measurement of facial activity intensity". In: *IEEE Sensors Journal* 13.11 (2013), pp. 4329–4338.

[10] Seppo Soinila, Markku Kaste, and Hannu Somer. *Neurologia*. 2nd ed. Duodecim, 2006, pp. 185, 197–200.

[11] Nelson Murray Gantz. *Manual of Clinical Problems in Infectious Disease*. Philadelphia: Wolters Kluwer, 2005. Chap. VII Nervous System, pp. 183–190.

[12] Susan E Mackinnon. *Nerve Surgery*. New York: Thieme, 2015. Chap. 16 Facial Nerve Injury, pp. 481–503.

[13] Greet Gevers and Peter Lemkens. "Bilateral simultaneous facial paralysis–differential diagnosis and treatment options. A case report and review of literature." In: *Acta Otorhinolaryngologica Belgica* 57.2 (2003), pp. 139–146.

[14] Joel N Bleicher, Steve Hamiel, Jon S Gengler, et al. "A survey of facial paralysis: etiology and incidence." In: *Ear, Nose, & Throat Journal* 75.6 (1996), pp. 355–358.

[15] Jeffrey D Tiemstra and Nandini Khatkhate. "Bell's palsy: diagnosis and management". In: *American Academy of Family Physicians* 76.7 (2007), pp. 997–1002.

[16] Frederick J McCoy and R Cole Goodman. “The crocodile tear syndrome.” In: *Plastic and Reconstructive Surgery* 63.1 (1979), pp. 58–62.

[17] Adel Y Fattah, Anthony DR Gurusinghe, Javier Gavilan, et al. “Facial nerve grading instruments: systematic review of the literature and suggestion for uniformity”. In: *Plastic and Reconstructive Surgery* 135.2 (2015), pp. 569–579.

[18] Rafal Niziol, Francis P Henry, Jonathan I Leckenby, et al. “Is there an ideal outcome scoring system for facial reanimation surgery? A review of current methods and suggestions for future publications”. In: *Journal of Plastic, Reconstructive & Aesthetic Surgery* 68.4 (2015), pp. 447–456.

[19] Katsuaki Mishima and Toshio Sugahara. “Analysis methods for facial motion”. In: *Japanese Dental Science Review* 45.1 (2009), pp. 4–13.

[20] Jeffrey T Vrabec, Douglas D Backous, Hamid R Djalilian, et al. “Facial nerve grading system 2.0”. In: *Otolaryngology-Head and Neck Surgery* 140.4 (2009), pp. 445–450.

[21] Brent I Lewis and Kedar K Adour. “An analysis of the Adour-Swanson and House-Brackmann grading systems for facial nerve recovery”. In: *European Archives of Otorhinolaryngology* 252.5 (1995), pp. 265–269.

[22] Susan E Coulson, Glen R Croxson, Roger D Adams, et al. “Reliability of the “Sydney,”“Sunnybrook,” and “House Brackmann” facial grading systems to assess voluntary movement and synkinesis after facial nerve paralysis”. In: *Otolaryngology-Head and Neck Surgery* 132.4 (2005), pp. 543–549.

[23] George E. Murty, Joseph P. Diver, Peter J. Kelly, et al. “The Nottingham System: Objective assessment of facial nerve function in the clinic”. In: *Otolaryngology-Head and Neck Surgery* 110.2 (1994), pp. 156–161.

[24] Jurg Rickenmann, Claude Jaquenod, D. Cerenko, et al. “Comparative value of facial nerve grading systems”. In: *Otolaryngology-Head and Neck Surgery* 117.4 (1997), pp. 322–325.

[25] Thomas L Yen, Colin LW Driscoll, and Anil K Lalwani. “Significance of House-Brackmann facial nerve grading global score in the setting of differential facial nerve function”. In: *Otology & Neurotology* 24.1 (2003), pp. 118–122.

[26] Shari D Reitzen, James S Babb, and Anil K Lalwani. “Significance and reliability of the House-Brackmann grading system for regional facial nerve function”. In: *Otolaryngology-Head and Neck Surgery* 140.2 (2009), pp. 154–158.

[27] Paul W Gidley, Bruce J Gantz, and Jay T Rubinstein. “Facial nerve grafts: from cerebellopontine angle and beyond.” In: *The American Journal of Otology* 20.6 (1999), pp. 781–788.

[28] Thomas S Kang, Jeffrey T Vrabec, Neil Giddings, et al. “Facial nerve grading systems (1985–2002): beyond the House-Brackmann scale”. In: *Otology & Neurotology* 23.5 (2002), pp. 767–771.

[29] J Gail Neely, Nevin G Cherian, Cody B Dickerson, et al. “Sunnybrook facial grading system: reliability and criteria for grading”. In: *The Laryngoscope* 120.5 (2010), pp. 1038–1045.

[30] Gerard H Chee and Julian M Nedzelski. “Facial nerve grading systems”. In: *Facial Plastic Surgery* 16.04 (2000), pp. 315–324.

[31] Mervi Kanerva, Tuija Poussa, and Anne Pitkäranta. “Sunnybrook and House-Brackmann facial grading systems: intrarater repeatability and interrater agreement”. In: *Otolaryngology-Head and Neck Surgery* 135.6 (2006), pp. 865–871.

[32] Fatma Tulin Kayhan, David Zurakowski, and Steven D. Rauch. “Toronto facial grading system: interobserver reliability”. In: *Otolaryngology - Head and Neck Surgery* 122.2 (2000), pp. 212–215.

[33] Tim T Wang, Louis Wessels, Gazi Hussain, et al. “Discriminative thresholds in facial asymmetry: a review of the literature”. In: *Aesthetic Surgery Journal* 37.4 (2017), pp. 375–385.

[34] Marc H Hohman, Sang W Kim, Elizabeth S Heller, et al. “Determining the threshold for asymmetry detection in facial expressions”. In: *The Laryngoscope* 124.4 (2014), pp. 860–865.

[35] Philipp Meyer-Marcotty, Angelika Stellzig-Eisenhauer, Ute Bareis, et al. “Three-dimensional perception of facial asymmetry”. In: *The European Journal of Orthodontics* 33.6 (2011), pp. 647–653.

[36] Eugene A Chu, Tarik Y Farrag, Lisa E Ishii, et al. “Threshold of visual perception of facial asymmetry in a facial paralysis model”. In: *Archives of Facial Plastic Surgery* 13.1 (2011), pp. 14–19.

[37] Sang W Kim, Elizabeth S Heller, Marc H Hohman, et al. “Detection and perceptual impact of side-to-side facial movement asymmetry”. In: *JAMA Facial Plastic Surgery* 15.6 (2013), pp. 411–416.

[38] Bruno Pereira Silva, Emilio Jimenez-Castellanos, Rafael Martinez-de-Fuentes, et al. “Laypersons’ perception of facial and dental asymmetries.” In: *International Journal of Periodontics & Restorative Dentistry* 33.6 (2013).

[39] Kyoung-Ho Kwak, Yong-Il Kim, Hyung-Jin Nam, et al. “Differences among deviations, genders, and observers in the perception of eye and nose asymmetry”. In: *Journal of Oral and Maxillofacial Surgery* 73.8 (2015), pp. 1606–1614.

[40] Farhad B Naini, Ana Nora A Donaldson, Fraser McDonald, et al. “Assessing the influence of asymmeftry affecting the mandible and chin point on perceived attractiveness in the orthognathic patient, clinician, and layperson”. In: *Journal of Oral and Maxillofacial Surgery* 70.1 (2012), pp. 192–206.

[41] Justin L Paletz, Ralph T Manktelow, and Roman Chaban. “The shape of a normal smile: implications for facial paralysis reconstruction.” In: *Plastic and Reconstructive Surgery* 93.4 (1994), pp. 784–9.

[42] Lisa Ishii, John Carey, Patrick Byrne, et al. “Measuring attentional bias to peripheral facial deformities”. In: *The Laryngoscope* 119.3 (2009), pp. 459–465.

[43] Manuel R Rodriguez and Eduardo Méndez. "Quantal processing of visual information in the brain". In: *Neuroscience* 84.3 (1998), pp. 641–644.

[44] Alice Frigerio, Paolo Cavallari, Marta Frigeni, et al. "Surface electromyographic mapping of the orbicularis oculi muscle for real-time blink detection". In: *JAMA Facial Plastic Surgery* 16.5 (2014), pp. 335–342.

[45] Hashmat Popat, Stephen Richmond, Lanthao Benedikt, et al. "Quantitative analysis of facial movement—A review of three-dimensional imaging techniques". In: *Computerized Medical Imaging and Graphics* 33.5 (2009), pp. 377–383.

[46] Peter C Revenaugh, Ryan M Smith, Max A Plitt, et al. "Use of objective metrics in dynamic facial reanimation". In: *JAMA Facial Plastic Surgery* (2018), pp. 501–508.

[47] Galen S Wachtman, Jeffrey F Cohn, Jessie M VanSwearingen, et al. "Automated tracking of facial features in patients with facial neuromuscular dysfunction". In: *Plastic and Reconstructive Surgery* 107.5 (2001), pp. 1124–1133.

[48] M Melissa Gross, Carroll-Ann Trotman, and Kelly S Moffatt. "A comparison of three-dimensional and two-dimensional analyses of facial motion". In: *The Angle Orthodontist* 66.3 (1996), pp. 189–194.

[49] Ritvik P Mehta, Song Zhang, and Tessa A Hadlock. "Novel 3-D video for quantification of facial movement". In: *Otolaryngology–Head and Neck Surgery* 138.4 (2008), pp. 468–472.

[50] Sachiyo Katsumi, Shinichi Esaki, Koosuke Hattori, et al. "Quantitative analysis of facial palsy using a three-dimensional facial motion measurement system". In: *Auris Nasus Larynx* 42.4 (2015), pp. 275–283.

[51] Mohammad Alqattan, Jelena Djordjevic, Alexei I Zhurov, et al. "Comparison between landmark and surface-based three-dimensional analyses of facial asymmetry in adults". In: *The European Journal of Orthodontics* 37.1 (2015), pp. 1–12.

[52] Chiarella Sforza, Virgilio F Ferrario, et al. "Soft-tissue facial anthropometry in three dimensions: from anatomical landmarks to digital morphology in research, clinics and forensic anthropology". In: *Journal of Anthropological Sciences* 84 (2006), pp. 97–124.

[53] Yue Wu and Qiang Ji. "Facial landmark detection: a literature survey". In: *International Journal of Computer Vision* 127.2 (2017), pp. 1–28.

[54] Stanislav Katina, Kathryn McNeil, Ashraf Ayoub, et al. "The definitions of three-dimensional landmarks on the human face: an interdisciplinary view". In: *Journal of Anatomy* 228.3 (2016), pp. 355–365.

[55] Leslie G Farkas. *Anthropometry of the Head and Face*. Raven Pr, 1994.

[56] Joan T Richtsmeier, Valerie Burke Deleon, and Subhash R Lele. "The promise of geometric morphometrics". In: *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists* 119.S35 (2002), pp. 63–91.

[57] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, et al. "The first facial landmark tracking in-the-wild challenge: Benchmark and results". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2015, pp. 50–58.

[58] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. "Automatic analysis of facial affect: A survey of registration, representation, and recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.6 (2015), pp. 1113–1133.

[59] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, et al. "DeepFace: closing the gap to human-level performance in face verification". In: *IEEE Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 1701–1708.

[60] Robin J Hennessy, Patrizia A Baldwin, David J Browne, et al. "Three-dimensional laser surface imaging and geometric morphometrics resolve frontonasal dysmorphology in schizophrenia". In: *Biological Psychiatry* 61.10 (2007), pp. 1187–1194.

[61] Arshed M Toma, A Zhurov, R Playle, et al. "Reproducibility of facial soft tissue landmarks on 3D laser-scanned facial images". In: *Orthodontics & Craniofacial Research* 12.1 (2009), pp. 33–42.

[62] Jelena Djordjevic, Pertti Pirttiniemi, Virpi Harila, et al. "Three-dimensional longitudinal assessment of facial symmetry in adolescents". In: *The European Journal of Orthodontics* 35.2 (2011), pp. 143–151.

[63] Bernardo Hontanilla and Cristina Aubá. "Automatic three-dimensional quantitative analysis for evaluation of facial movement". In: *Journal of Plastic, Reconstructive & Aesthetic Surgery* 61.1 (2008), pp. 18–30.

[64] Matthew Shreve, Neeha Jain, Dmitry Goldgof, et al. "Evaluation of facial reconstructive surgery on patients with facial palsy using optical strain". In: *Computer Analysis of Images and Patterns*. Springer. 2011, pp. 512–519.

[65] Mohammad Y Hajeer, Ashraf F Ayoub, Declan T Millett, et al. "Three-dimensional imaging in orthognathic surgery: the clinical application of a new method." In: *The International Journal of Adult Orthodontics and Orthognathic Surgery* 17.4 (2002), pp. 318–330.

[66] Jamie R Gwilliam, Susan J Cunningham, and Tim Hutton. "Reproducibility of soft tissue landmarks on three-dimensional facial scans". In: *The European Journal of Orthodontics* 28.5 (2006), pp. 408–415.

[67] Muhammad Sajid, Tamoor Shafique, Mirza Baig, et al. "Automatic grading of palsy using asymmetrical facial features: a study complemented by new solutions". In: *Symmetry* 10.7 (2018), p. 242.

[68] Sergey Tulyakov and Nicu Sebe. "Regressing a 3D face shape from a single image". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 3748–3755.

[69] Chao Gou, Yue Wu, Fei-Yue Wang, et al. "Shape augmented regression for 3d face alignment". In: *European Conference on Computer Vision*. Springer. 2016, pp. 604–615.

[70] László A Jeni, Jeffrey F Cohn, and Takeo Kanade. "Dense 3D face alignment from 2D videos in real-time". In: *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. Vol. 1. IEEE. 2015, pp. 1–8.

[71] Amin Jourabloo and Xiaoming Liu. "Pose-invariant 3D face alignment". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 3694–3702.

[72] Chavdar Papazov, Tim K Marks, and Michael Jones. "Real-time 3d head pose and facial landmark estimation from depth images using triangular surface patch features". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 4722–4730.

[73] Shu Liang, Jia Wu, Seth M Weinberg, et al. "Improved detection of landmarks on 3d human face data". In: *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*. IEEE. 2013, pp. 6482–6485.

[74] Daniele Gibelli, Valentina Pucciarelli, Zuzana Caplova, et al. "Validation of a low-cost laser scanner device for the assessment of three-dimensional facial anatomy in living subjects". In: *Journal of Cranio-Maxillofacial Surgery* 46.9 (2018), pp. 1493–1499.

[75] Daniele Gibelli, Valentina Pucciarelli, Annalisa Cappella, et al. "Are portable stereophotogrammetric devices reliable in facial imaging? A validation study of Vectra H1 device". In: *Journal of Oral and Maxillofacial Surgery* 76.8 (2018), pp. 1772–1784.

[76] Daniele Gibelli, Valentina Pucciarelli, Pasquale Poppa, et al. "Three-dimensional facial anatomy evaluation: Reliability of laser scanner consecutive scans procedure in comparison with stereophotogrammetry". In: *Journal of Cranio-Maxillofacial Surgery* 46.10 (2018), pp. 1807–1813.

[77] Chung How Kau, Stephen Richmond, Alexei I Zhurov, et al. "Reliability of measuring facial morphology with a 3-dimensional laser scanning system". In: *American Journal of Orthodontics and Dentofacial Orthopedics* 128.4 (2005), pp. 424–430.

[78] L Kovacs, A Zimmermann, G Brockmann, et al. "Three-dimensional recording of the human face with a 3D laser scanner". In: *Journal of Plastic, Reconstructive & Aesthetic Surgery* 59.11 (2006), pp. 1193–1202.

[79] Michael W Vannier, Tom Pilgram, Gulab Bhatia, et al. "Facial surface scanner". In: *IEEE Computer Graphics and Applications* 11.6 (1991), pp. 72–80.

[80] Chieh-Han John Tzou, Nicole M Artner, Igor Pona, et al. "Comparison of three-dimensional surface-imaging systems". In: *Journal of Plastic, Reconstructive & Aesthetic Surgery* 67.4 (2014), pp. 489–497.

[81] Mahmoud Amir Alagha, Xiangyang Ju, Stephen Morley, et al. "Reproducibility of the dynamics of facial expressions in unilateral facial palsy". In: *International Journal of Oral and Maxillofacial Surgery* 47.2 (2018), pp. 268–275.

[82] Marina Codari, Valentina Pucciarelli, Fabiano Stangoni, et al. "Facial thirds–based evaluation of facial asymmetry using stereophotogrammetric devices: Application to facial palsy subjects". In: *Journal of Cranio-Maxillofacial Surgery* 45.1 (2017), pp. 76–81.

[83] Filareti Tsalakanidou, Frank Forster, Sotiris Malassiotis, et al. "Real-time acquisition of depth and color images using structured light and its application to 3D face recognition". In: *Real-Time Imaging* 11.5-6 (2005), pp. 358–369.

[84] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, et al. "Static and dynamic 3D facial expression recognition: A comprehensive survey". In: *Image and Vision Computing* 30.10 (2012), pp. 683–697.

[85] Timen C Harkel, Caroline M Speksnijder, Ferdinand Heijden, et al. "Depth accuracy of the RealSense F200: Low-cost 4D facial imaging". In: *Scientific Reports* 7.16263 (2017), pp. 1–8.

[86] Dugan Um, Dongseok Ryu, and Myungjoon Kal. "Multiple intensity differentiation for 3-d surface reconstruction with mono-vision infrared proximity array sensor". In: *IEEE Sensors Journal* 11.12 (2011), pp. 3352–3358.

[87] Stefanie TL Pöhlmann, Elaine F Harkness, Christopher J Taylor, et al. "Evaluation of Kinect 3D sensor for healthcare imaging". In: *Journal of Medical and Biological Engineering* 36.6 (2016), pp. 857–870.

[88] Supriya Sathyanarayana, Ravi Kumar Satzoda, Suchitra Sathyanarayana, et al. "Vision-based patient monitoring: a comprehensive review of algorithms and technologies". In: *Journal of Ambient Intelligence and Humanized Computing* 9.2 (2018), pp. 225–251.

[89] Amira Gaber, Mona F Taher, and Manal Abdel Wahed. "Quantifying facial paralysis using the kinect v2". In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2015, pp. 2497–2501.

[90] Pawel Mazurek, Jakub Wagner, and Roman Z Morawski. "Acquisition and preprocessing of data from infrared depth sensors to be applied for patients monitoring." In: *IDAACS*. 2015, pp. 705–710.

[91] Zhilei Liu, Guoxian Song, Jianfei Cai, et al. "Conditional adversarial synthesis of 3D facial action units". In: *Neurocomputing* 355 (2019), pp. 200–208.

[92] Andrés Romero, Juan León, and Pablo Arbeláez. "Multi-view dynamic facial action unit detection". In: *Image and Vision Computing* (recommended for acceptance 2018).

[93] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[94] Jia Deng, Wei Dong, Richard Socher, et al. "Imagenet: A large-scale hierarchical image database". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255.

[95] Ting Wang, Junyu Dong, Xin Sun, et al. "Automatic recognition of facial movement for paralyzed face". In: *Bio-medical Materials and Engineering* 24.6 (2014), pp. 2751–2760.

[96] Shu He, John J Soraghan, Brian F O'Reilly, et al. "Quantitative analysis of facial paralysis using local binary patterns in biomedical videos". In: *IEEE Transactions on Biomedical Engineering* 56.7 (2009), pp. 1864–1870.

[97] Jane Reilly Delannoy and Tomás E Ward. "A preliminary investigation into the use of machine vision techniques for automating facial paralysis rehabilitation therapy". In: *IET Irish Signals and Systems Conference*. 2010, pp. 228–232.

[98] Stewart McGrenary, Brian F O'Reilly, and John J Soraghan. "Objective grading of facial paralysis using artificial intelligence analysis of video data". In: *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*. IEEE. 2005, pp. 587–592.

[99] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. "Generative adversarial nets". In: *Advances in Neural Information Processing Systems*. 2014, pp. 2672–2680.

[100] Anselm Brachmann and Christoph Redies. "Using convolutional neural network filters to measure left-right mirror symmetry in images". In: *Symmetry* 8.12 (2016), p. 144.

[101] Olga Russakovsky, Jia Deng, Hao Su, et al. "Imagenet large scale visual recognition challenge". In: *International Journal of Computer Vision* 115.3 (2015), pp. 211–252.

[102] Paul Ekman, Wallace V Friesen, and Joseph C Hager. *Facial Action Coding System (FACS)*. Manual. 2002.

[103] Brais Martinez, Michel F Valstar, Bihan Jiang, et al. "Automatic analysis of facial actions: A survey". In: *IEEE Transactions on Affective Computing* 10.3 (2017), pp. 325–347.

[104] Gianluca Donato, Marian Stewart Bartlett, Joseph C Hager, et al. "Classifying Facial Actions". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21.10 (1999), pp. 974–989.

[105] Manfred Frey, Andreas Jenny, Pietro Giovanoli, et al. "Development of a new documentation system for facial movements as a basis for the international registry for neuromuscular reconstruction in the face". In: *Plastic and Reconstructive Surgery* 93 (7 1994), pp. 1334–1349.

[106] Eric Ries. *The Lean Startup*. 1st ed. Crown Business, 2011.

[107] Bracewell R.N. *The Fourier Transform and its Applications*. New York: McGraw-Hill, 1978.

[108] Conrad Sanderson. *Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments*. Technical Report. National ICT Australia (NICTA), 2010.

[109] *Armadillo C++ library for linear algebra & scientific computing*. `http://Arma.sourceforge.net`. Accessed: 2019-03-26.

[110] Bjarne Stroustrup. *Programming Principles and Practice Using C++*. English. 4th. Upper Saddle River (N.J.): Addison-Wesley, 2013.

[111] Tommi Mikkonen Ilkka Haikala. *Ohjelmistotuotannon käytännöt*. Finnish. 12th. Talentum, 2011. Chap. 13th, 16th.

[112] *Python Library NumPy*. `https://www.numpy.org/`. Accessed: 2019-05-12.

[113] *Python Plotting Library Matplotlib*. `https://matplotlib.org/`. Accessed: 2019-05-12.

[114] *Python Data Visualization Library Seaborn*. `https://seaborn.pydata.org/`. Accessed: 2019-05-12.

[115] *Data Science Platform Anaconda*. `https://www.anaconda.com/`. Accessed: 2019-05-12.

[116] Susan Weinschenk UX Magazine. *The Secret to Designing an Intuitive UX*. `http://uxmag.com/articles/the-secret-to-designing-an-intuitive-user-experience`. Accessed: 2016-09-22. 2011.

[117] Carolyn Chandler Russ Unger. *A Project Guide to UX Design: For User Experience Designers in the Field or in the Making*. English. 2nd. New Riders, 2012. Chap. 4th.

[118] UX Apprentice. `http://www.uxapprentice.com/`. Accessed: 2016-09-22.

[119] Jared M. Spool, User Interface Engineering. *Starting Your User Research*. `https://articles.uie.com/starting_user_research/`. Accessed: 2016-09-22. 2012.

APPENDIX A: SUMMARY OF FACIAL ASYMMETRY THRESHOLD STUDIES

The following Table A.1 summarizes the most important studies presented in the theory Section 2.4. In the theory part, the studies are briefly explained when the article is referenced; for example, how many observers were used and what was the progressive model's interval in the study. Table A.1 summarizes the studies into a comparable form.

In Table A.1, the leftmost column gives the reference, the Observers-column lays out the number of observers, their type and the observation time if that was available. The Model-column shows the type of model used: 2D or 3D, and the gender of the model. The Interval-part contains the information about the progressive asymmetry model used: what was the interval, was it the same for every movement/area, and what was the range. The rightmost column describes the area or the expression the research studied.

Table A.1 *A summary of facial subunit asymmetry studies used in this thesis. Table is loosely adapted from [33].*

Study	**Observers**	**Model**	**Interval**	**Area / expression**
Chu et al. [36]	90 naive laypeople, 2 s and 10 s observation time	2D photography, digitally modifed, male	OC: 0-10 mm (1 mm increments) B: 0-5 mm (0.5 mm increments) OC+B: 0-5 mm (1 mm increments)	Brow, oral commissure, brow + oral commissure
Meyer-Marcotty et al. [35]	Informed 60 clinicians and 30 laypeople	3D facial model, male	2 mm, 4 mm, 6 mm, 8 mm	Nose, chin
Naini et al. [40]	Laypeople, clinicians, patients	2D model, female and male	0-25 mm with 5 mm increments	Chin
Kim et al. [37]	58 informed laypeople	Video, digitally manipulated, female	33 ms, 67 ms, 100 ms, 133 ms, 200 ms, 267 ms	Blink, slow eyebrow lift, fats eyebrow lift, lip depression, smile
Silva et al. [38]	100 laypeople	2D photograph, digitally manipulated, female	Nose: 1-4 mm with 1 mm increments Chin: 1-6 mm with 1 mm increments	Inter alia nose and chin
Hohman et al. [34]	145 informed laypeople and clinicians, 5 s observation time	2D photography, digitally modified, female	0 mm, 1 mm, 2 mm, 3 mm, 4 mm, 5 mm, 6 mm	Eyebrow elevation, eyelid closure, smile
Kwak et al. [39]	40 informed laypeople, 120 dental professionals and students	2D photograph, digitally manipulated, female	Nose, eyes: 1°-5° canting with 1° increments to both right and left (nose), and clockwise and counterclockwise (eyes)	Canting of eyes and nose

APPENDIX B: PRELIMINARY STUDY'S TEST PARTICIPANTS

Tables B.1 and B.2 contain the preliminary study's test participants' details. In methodology Subsection 4.5.1, a general overview of the participants is given.

In Tables B.1 and B.2 the leftmost column gives the abbreviation used for the participant. The abbreviation is in the form of 'tpX' where tp stands for test participant, and where X is a consecutive running number starting from 1. Abbreviation is used in the data gathering and analysis as the test participants agreed on giving right to use their capacitive data and the video recordings in data analysis in an anonymous manner. Thus only gender (second column from the left) and age (third column from the left) can be further linked with the data. The column 'Per 2b' lays out the information if the Permission 2b of using the video data for scientific publications and presentations is given. Also, notes made during data collection about fitting of the prototype (column 'Fitting') and ability to perform especially the asymmetric movements (the rightmost column) are included in the tables.

Table B.1 *Necessary details and anonymization of the healthy test participants 1-11.*

No.	**Gender**	**Age**	**Per 2b**	**Fitting**	**Notes**
tp01	Male	23	Yes	Prototype fitted well	Was able to lift only one eyebrow very well
tp02	Male	27	Yes	Did not seem to touch, said after measurements that might have touched the cheeks during movement (middle extensions)	-
tp03	Male	29	Yes	The lowest extensions: both 4th sensors touched, the middle extensions: both 3rd sensors touched	-
tp04	Female	25	No	The lowest extensions: both 4th sensors touched	Was able to lift only one eyebrow very well
tp05	Female	51	Yes	The lowest extensions: both 4th sensors touched, the middle extensions: both 3rd sensors touched	-
tp06	Male	35	No	The lowest extensions: both 4th sensors touched	-
tp07	Male	46	No	The lowest extensions: both 4th sensors touched, the middle extensions: both 3rd sensors touched	-
tp08	Female	50	Yes	The right middle extension: the 3rd sensor touched, the protype moved on the head during movements	-
tp09	Male	35	Yes	Prototype fitted well	-
tp10	Male	22	Yes	The left lowest extension: the 4th sensor touched	-
tp11	Male	26	Yes	The lowest extensions: both 4th sensors touched	-

Table B.2 *Necessary details and anonymization of the healthy test participants 12-20.*

No.	Gender	Age	Per 2b	Fitting	Notes
tp12	Female	35	No	The lowest extensions: both 4th sensors touched, the middle right extension: the third might have touched possibly	Was not able to lift only one eyebrow; lifted both
tp13	Female	34	No	The lowest extensions: both 4th sensors touched, the middle right extension: the third might have touched possibly	Was not able to lift only one eyebrow; lifted both
tp14	Male	33	Yes	No obvious touching, but in the middle extensions the 3rd sensors were very close	Was not able to lift only one eyebrow; lifted both
tp15	Female	25	Yes	The lowest extensions: both 4th sensors touched, the middle extensions: both 3rd sensors touched	-
tp16	Female	35	No	The lowest right extension: the 4th sensor may have touched	-
tp17	Male	35	Yes	The lowest extensions: both 4th sensors touched, the middle extensions: the 3rd sensors very close and may have touched	Was not able to lift only one eyebrow; lifted both
tp18	Female	42	No	The lowest extensions: both 4th sensors touched, the middle extensions: the 3rd sensors very close and may have touched	Was not able to lift only one eyebrow; lifted both
tp19	Male	28	Yes	Prototype fitted well	Was able to lift only one eyebrow very well
tp20	Female	29	Yes	The lowest extensions: both 4th sensors touched, the middle extensions: the 3rd sensors very close and may have touched	-

APPENDIX C: PRELIMINARY STUDY'S RESULTS

This appendix contains complementary pictures for preliminary results presented in Subsection 4.5.3. Section C.1 shows example figures for smiles at maximum intensity, and Section C.2 for eyebrow lifts at maximum intensity. Section C.3 lays out few examples of medium intensity repetitions, despite the fact that in Subsection 4.5.2 the data analysis focus is on maximum intensity movements. It should be noted, that figures should be read as pairs; for example, Figure C.1 shows the symmetrical movement results, and the following Figure C.2 the corresponding asymmetrical movement outcome.

C.1 Smile Results

The following figures give additional examples to typical smile at max intensity movements.

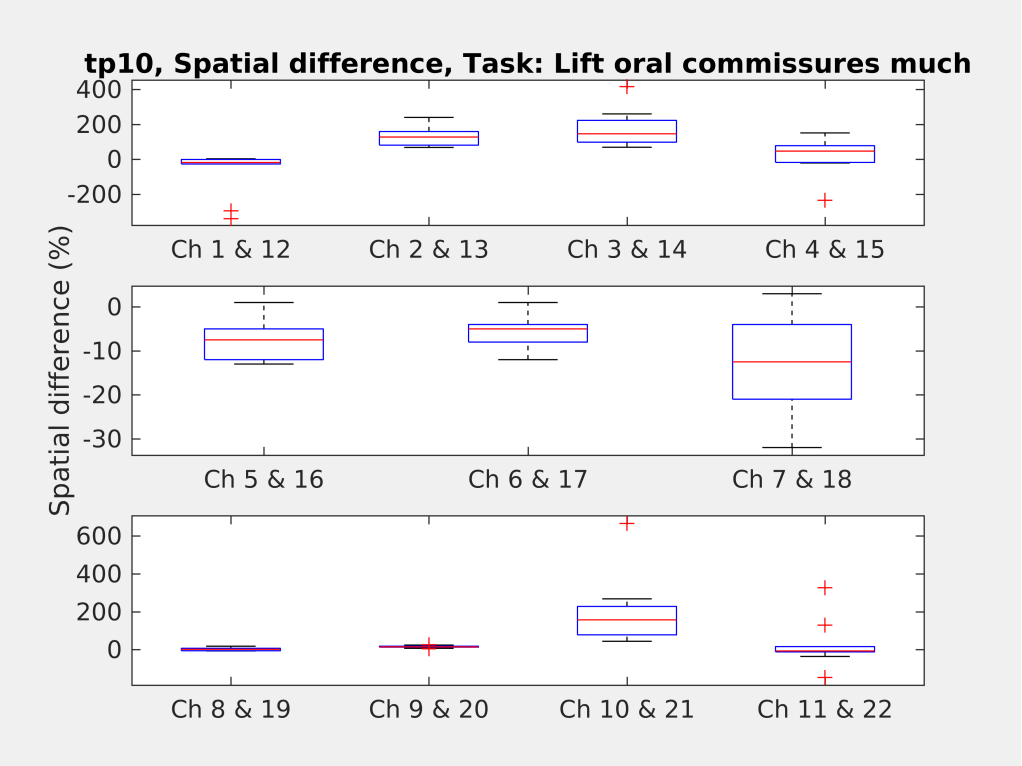

Figure C.1 *Contralateral spatial difference for repetitions of symmetrical smile, test person 10.*

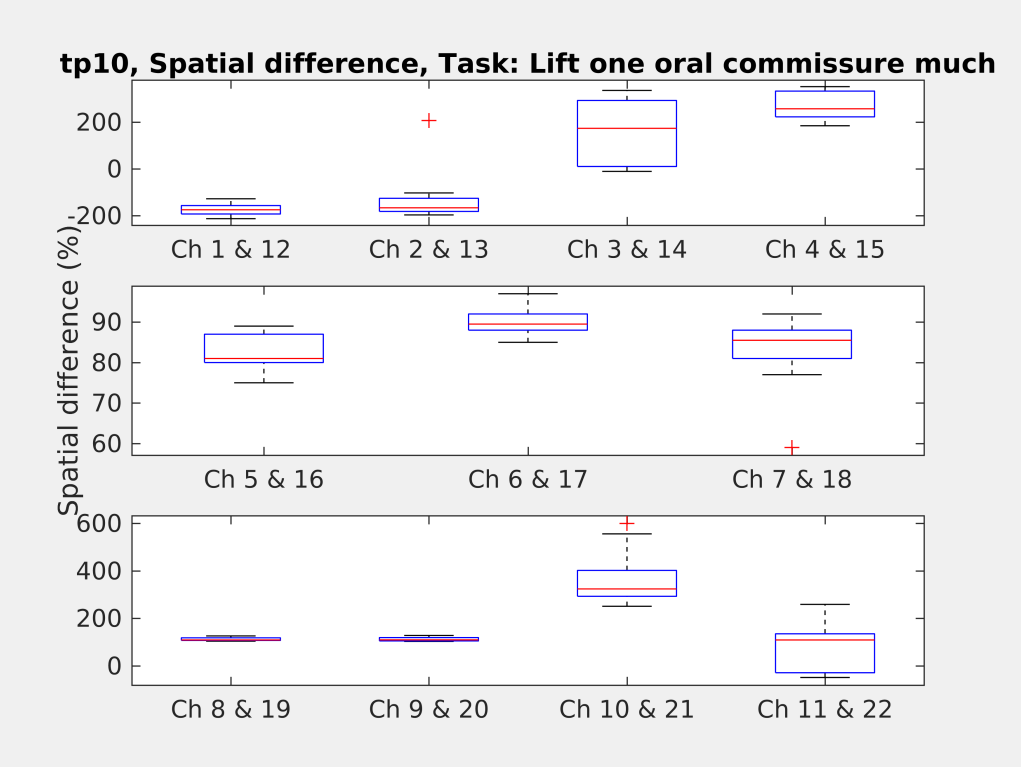

Figure C.2 *Contralateral spatial difference for repetitions of asymmetrical smile, test person 10.*

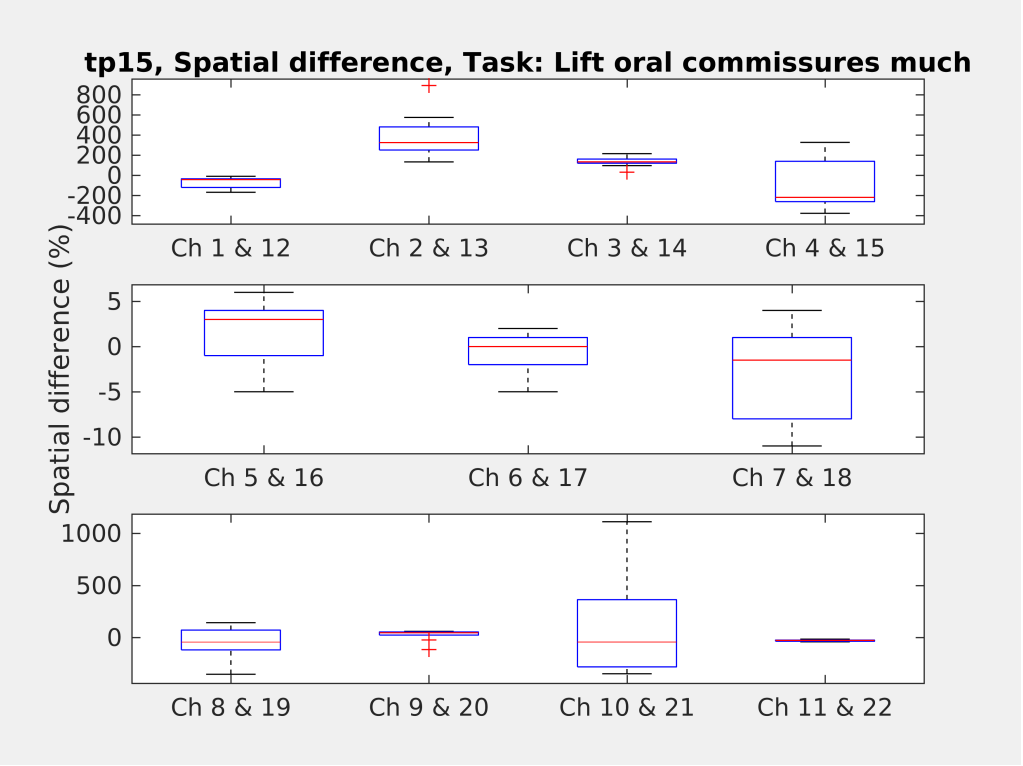

Figure C.3 *Contralateral spatial difference for repetitions of symmetrical smile, test person 15.*

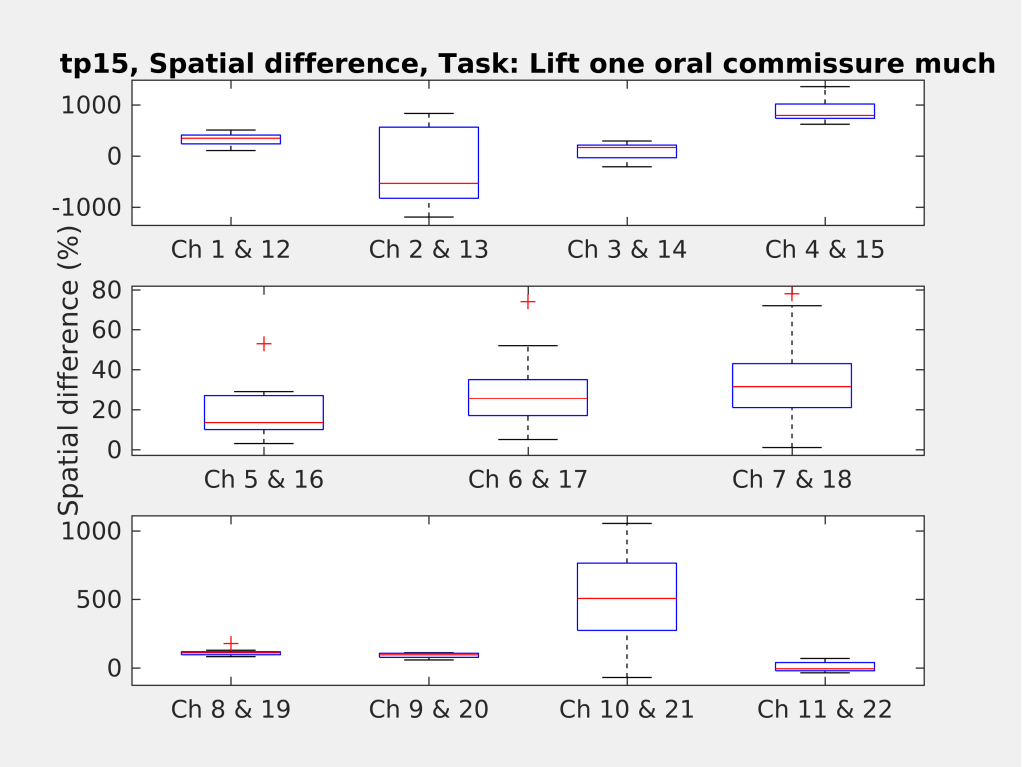

Figure C.4 *Contralateral spatial difference for repetitions of asymmetrical smile, test person 15.*

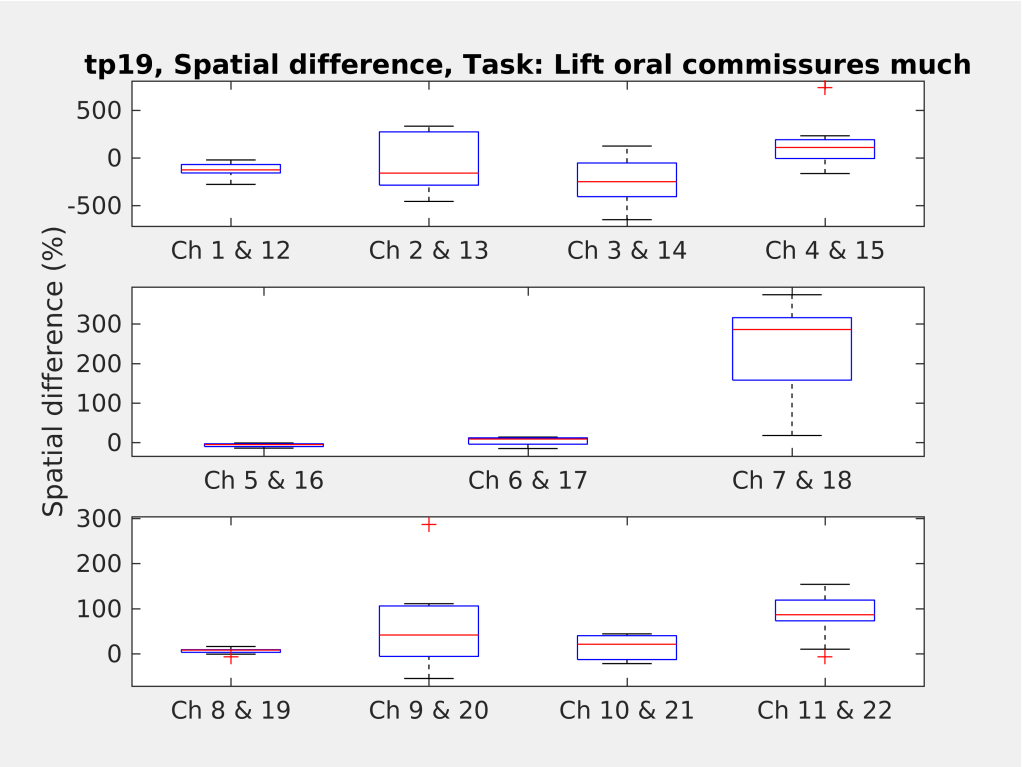

Figure C.5 *Contralateral spatial difference for repetitions of symmetrical smile, test person 19.*

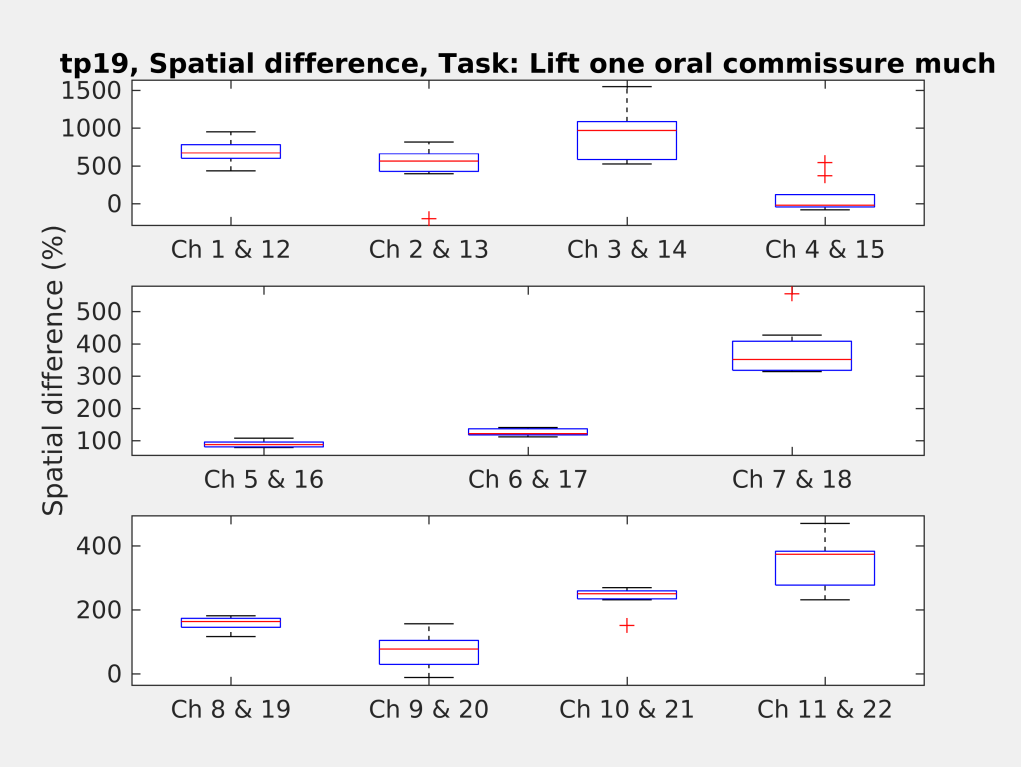

Figure C.6 *Contralateral spatial difference for repetitions of asymmetrical smile, test person 19.*

C.2 Eyebrow Lift Results

The following Figures C.7 and C.8 give additional example of successful eyebrow lift at max intensity movement. Figures C.9 and C.10 represent a failed case where the participant lifted both eyebrows also in the asymmetrical movement, and Figures C.11 and C.12 visualize an inconclusive data set. All figures show movements with maximum intensity.

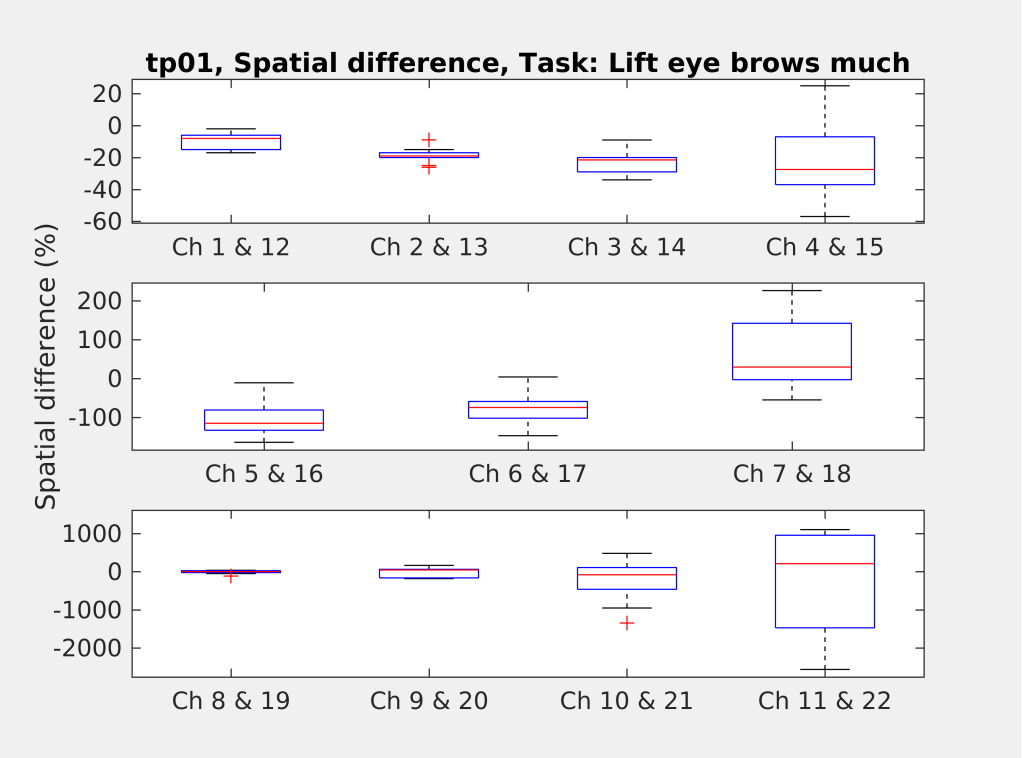

Figure C.7 *Contralateral spatial difference for repetitions of symmetrical eyebrow lift, test person 1.*

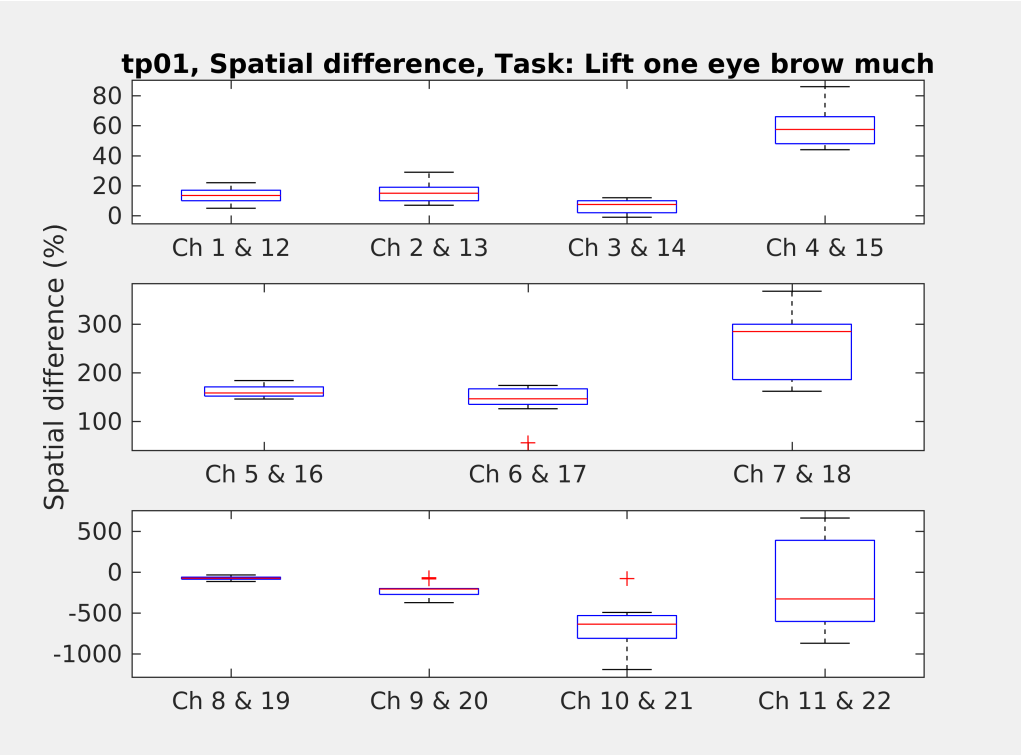

Figure C.8 *Contralateral spatial difference for repetitions of asymmetrical eyebrow lift, test person 1.*

Figure C.9 *Contralateral spatial difference for repetitions of symmetrical eyebrow lift, test person 18.*

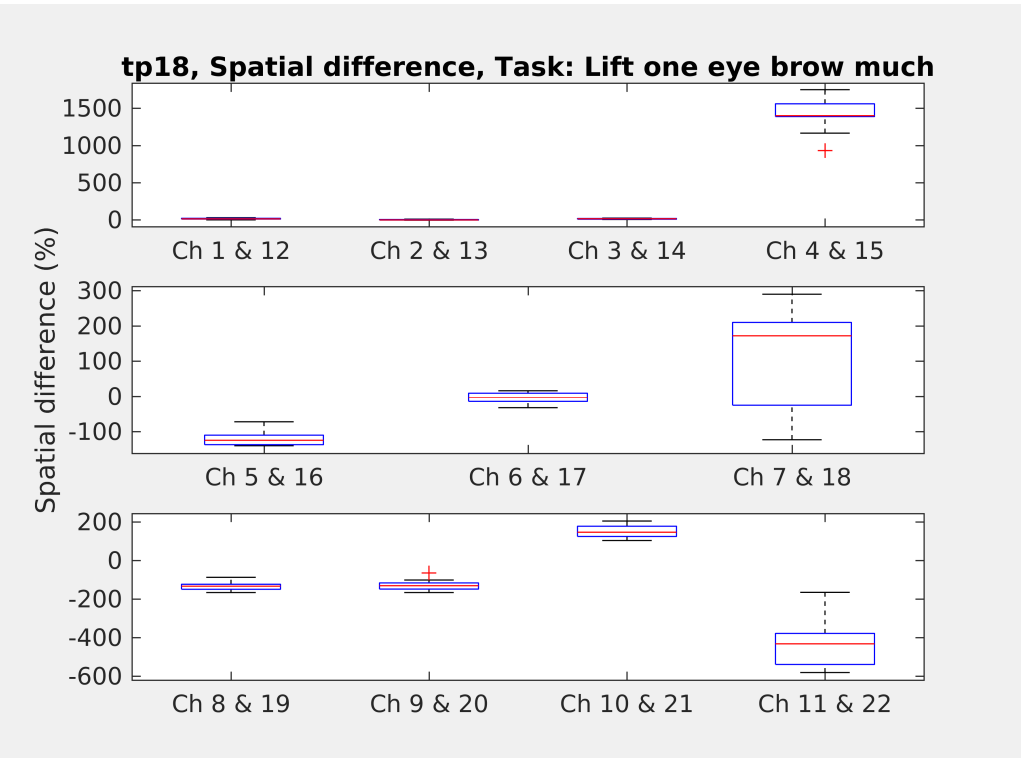

Figure C.10 *Contralateral spatial difference for repetitions of supposedly asymmetrical eyebrow lift, test person 18.*

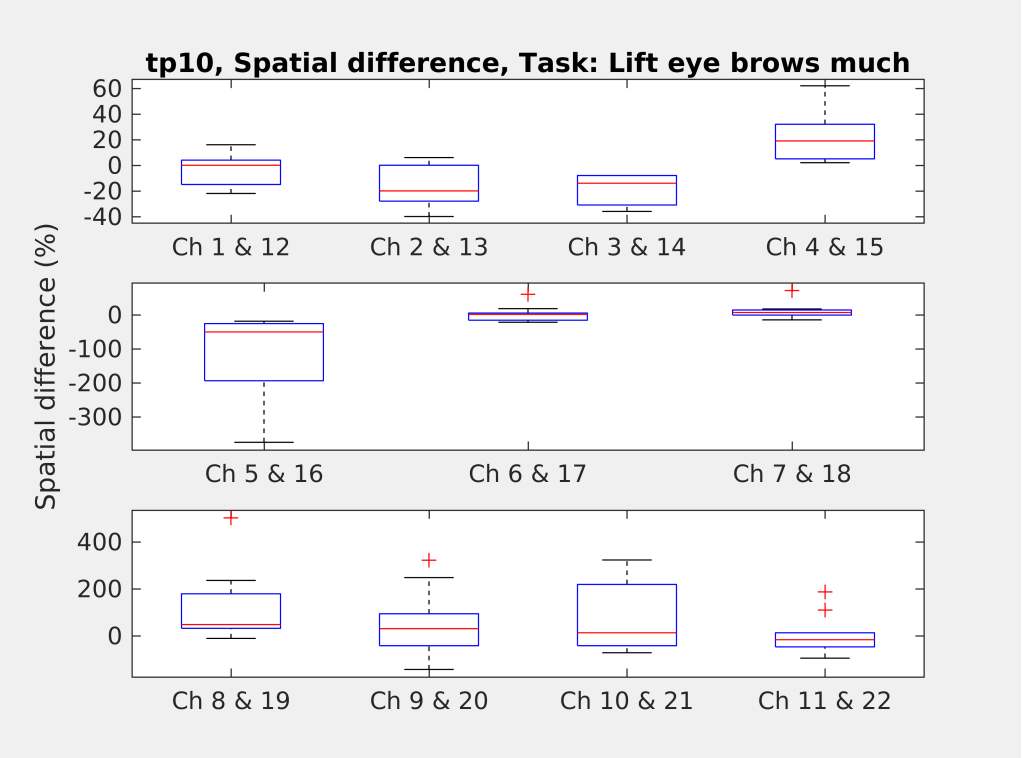

Figure C.11 *Contralateral spatial difference for repetitions of symmetrical eyebrow lift, test person 10.*

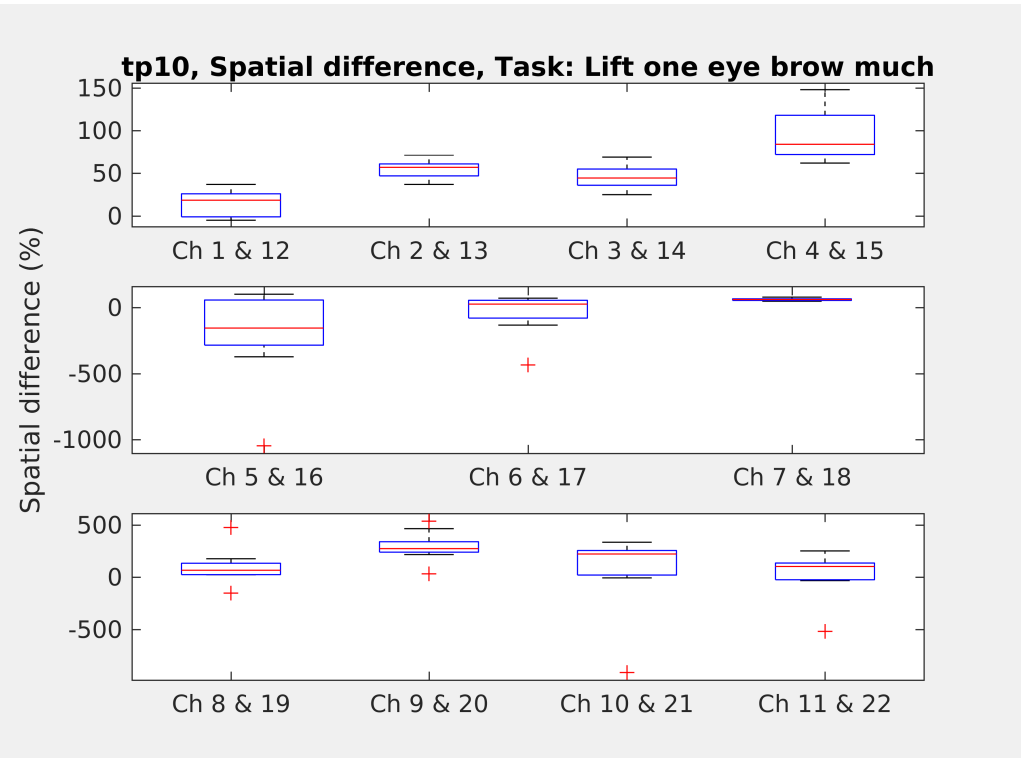

Figure C.12 *Contralateral spatial difference for repetitions of inconclusive asymmetrical eyebrow lift, test person 10.*

C.3 Results with Medium Intensity

The following Figures C.13-C.22 give examples of medium intensity movements, both eyebrow lift and smile movements. Medium intensity is excluded from further analysis, as stated in Subsection 4.5.2, but for completeness, some examples are given here. Figures C.13-C.16 visualize smile repetitions, Figures C.17 and C.18 successful eyebrow lifts (that test participant was observed to be able to complete asymmetric eyebrow lift), and Figures C.19-C.22 unsuccessful eyebrow lifts.

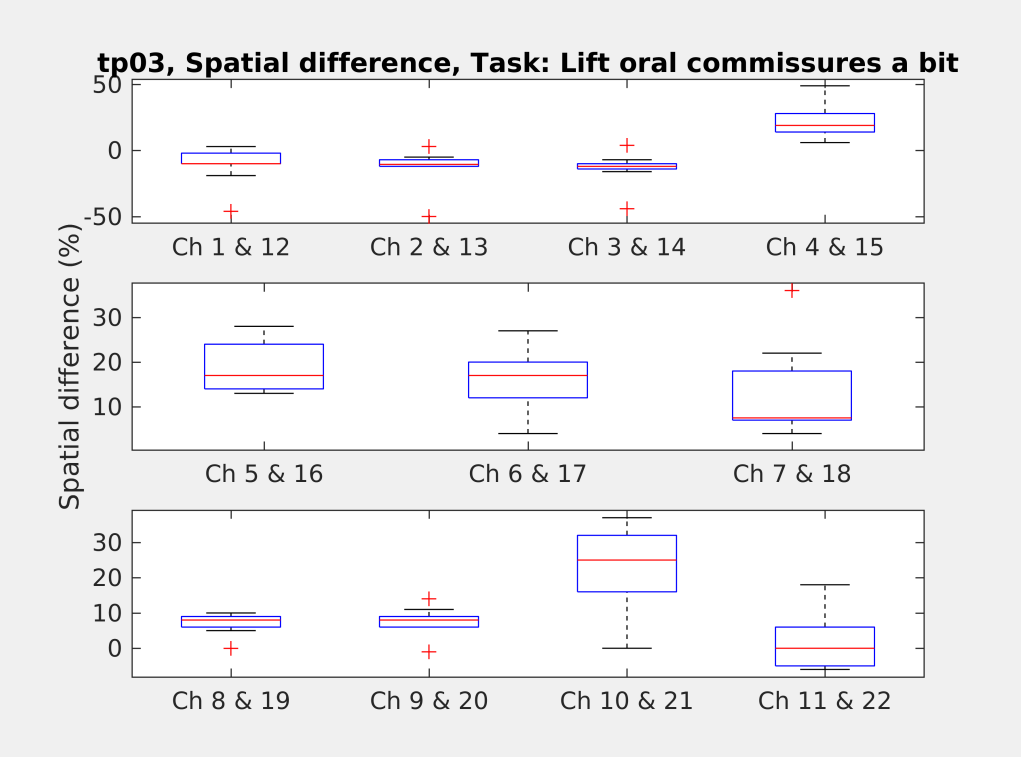

Figure C.13 *Contralateral spatial difference for repetitions of symmetrical smile at medium intensity, test person 3.*

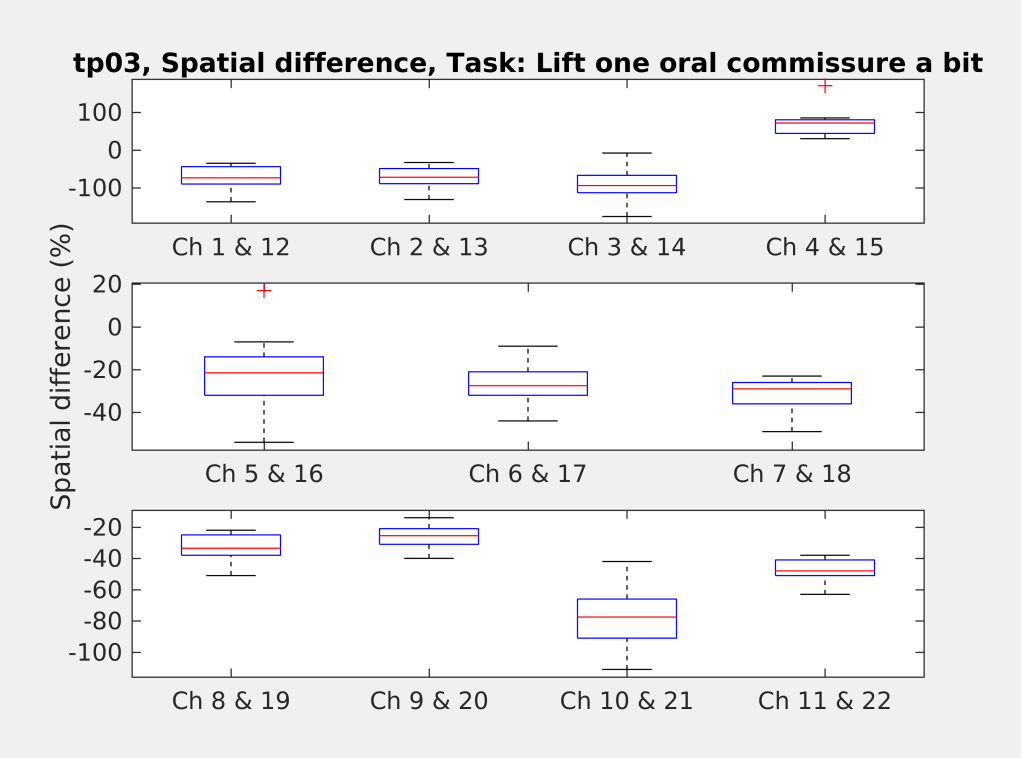

Figure C.14 *Contralateral spatial difference for repetitions of asymmetrical smile at medium intensity, test person 3.*

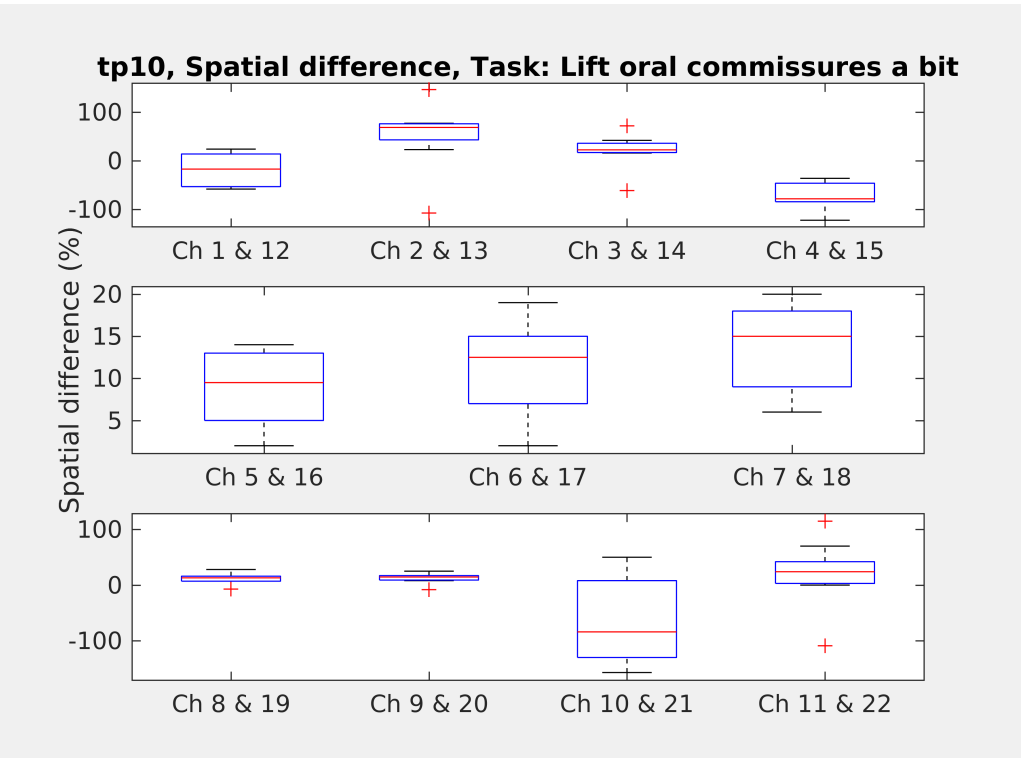

Figure C.15 *Contralateral spatial difference for repetitions of symmetrical smile at medium intensity, test person 10.*

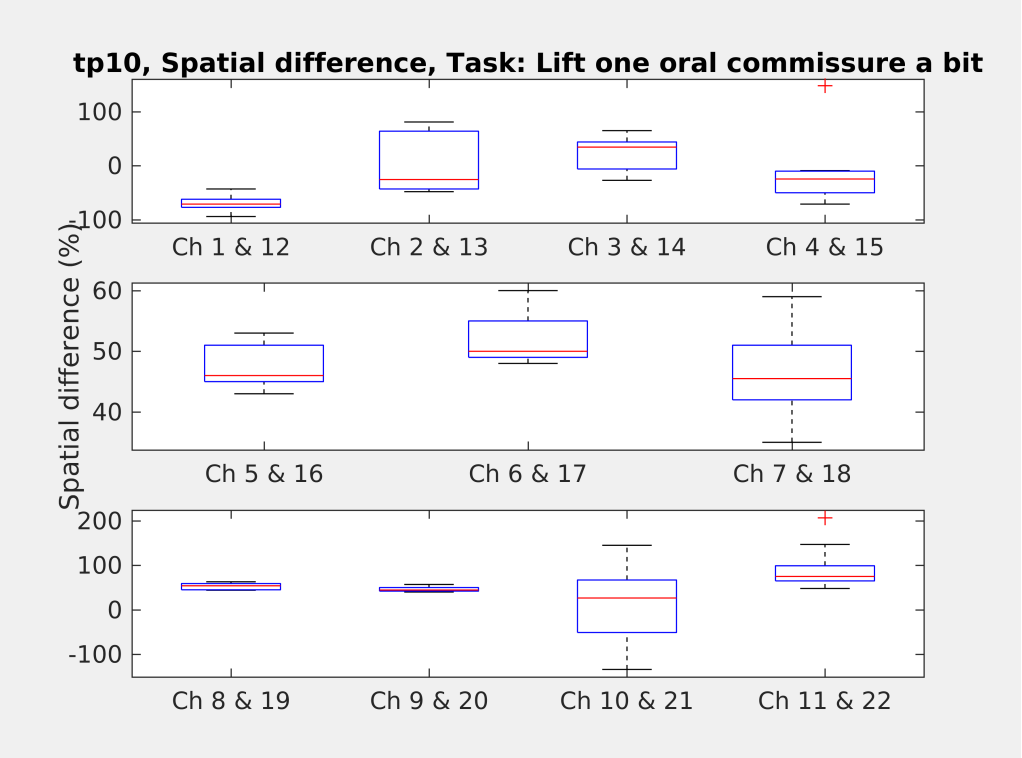

Figure C.16 *Contralateral spatial difference for repetitions of asymmetrical smile at medium intensity, test person 10.*

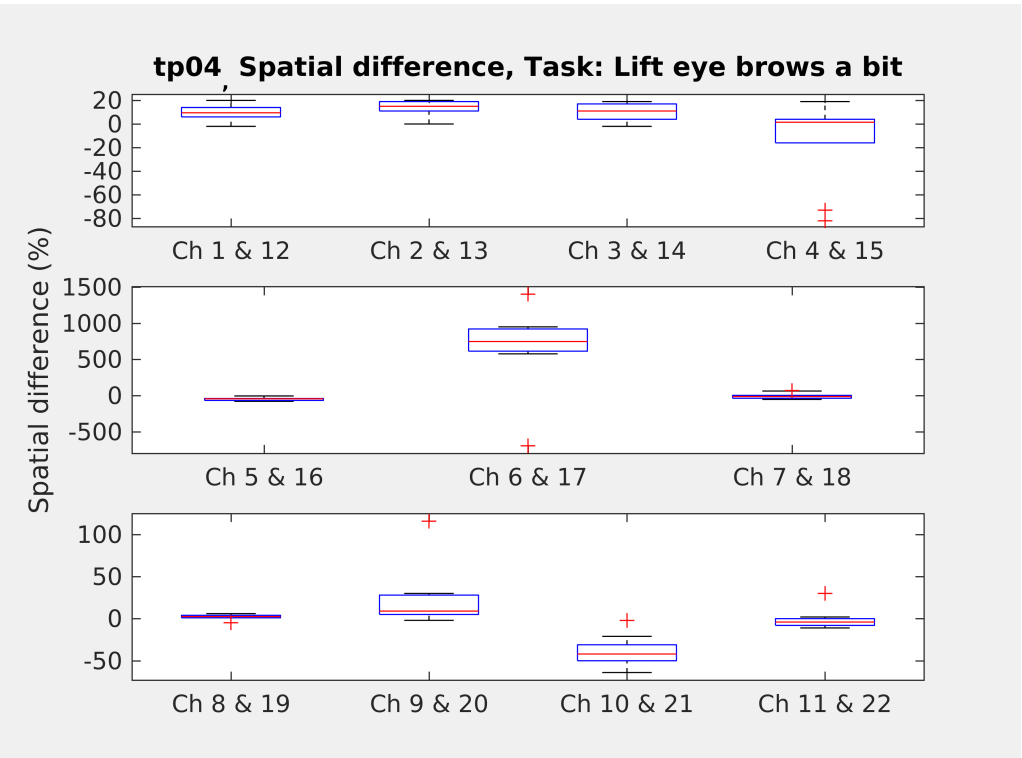

Figure C.17 *Contralateral spatial difference for repetitions of symmetrical eyebrow lift at medium intensity, test person 4.*

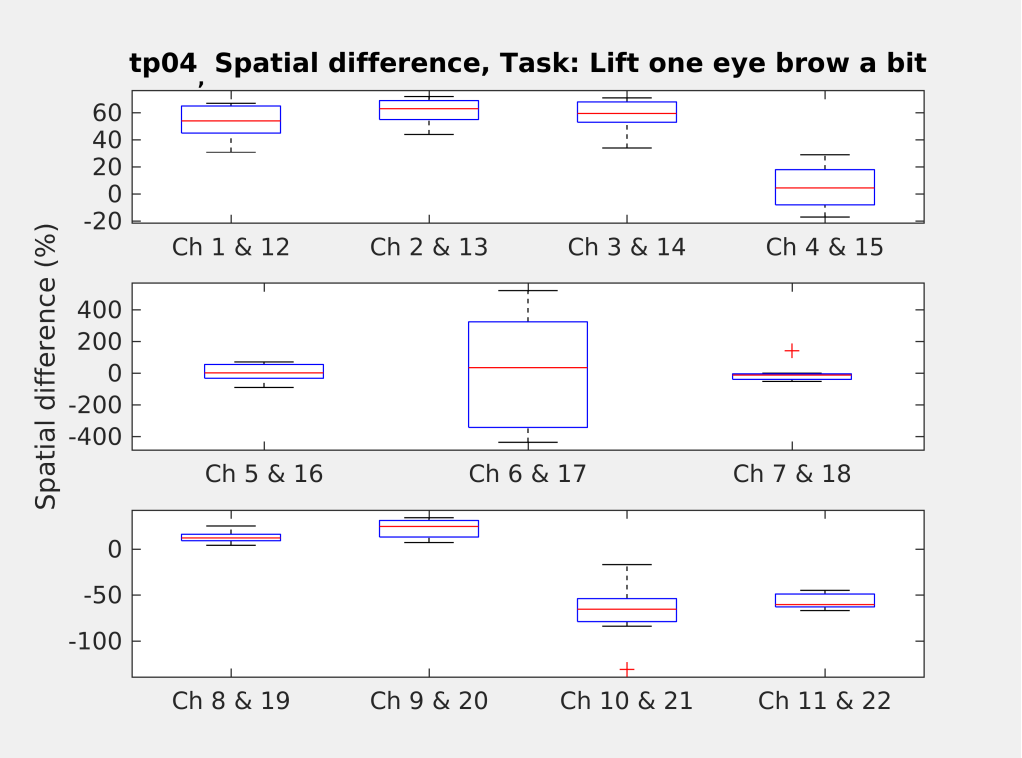

Figure C.18 *Contralateral spatial difference for repetitions of successful asymmetrical eyebrow lift at medium intensity, test person 4.*

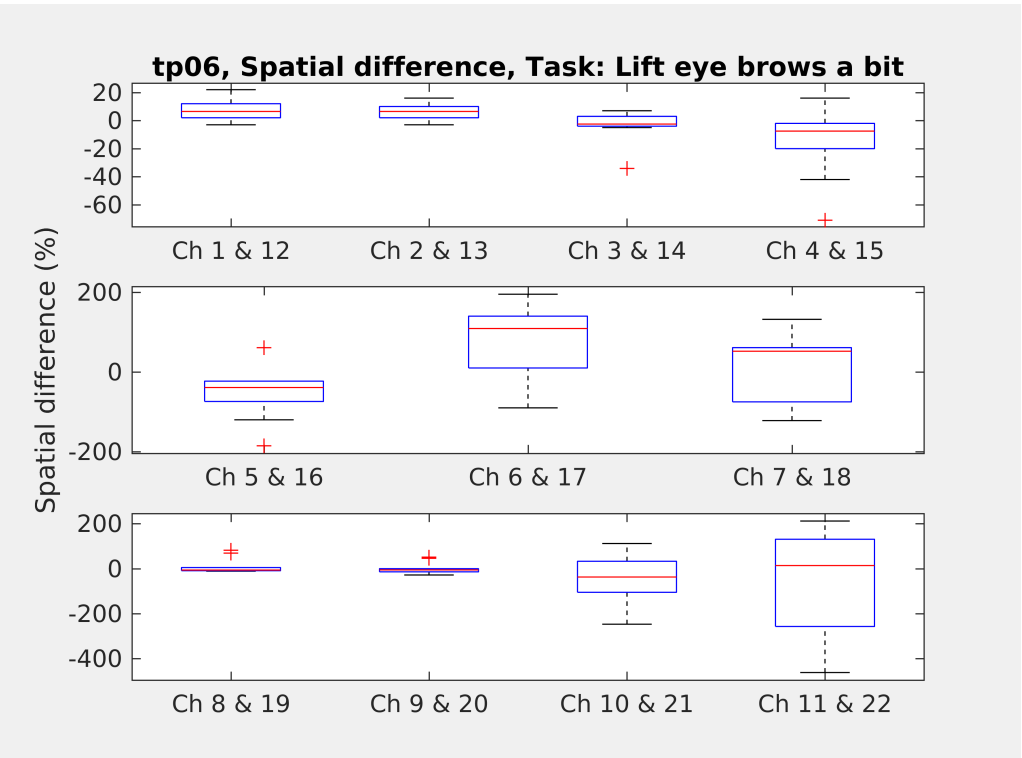

Figure C.19 *Contralateral spatial difference for repetitions of symmetrical eyebrow lift at medium intensity, test person 6.*

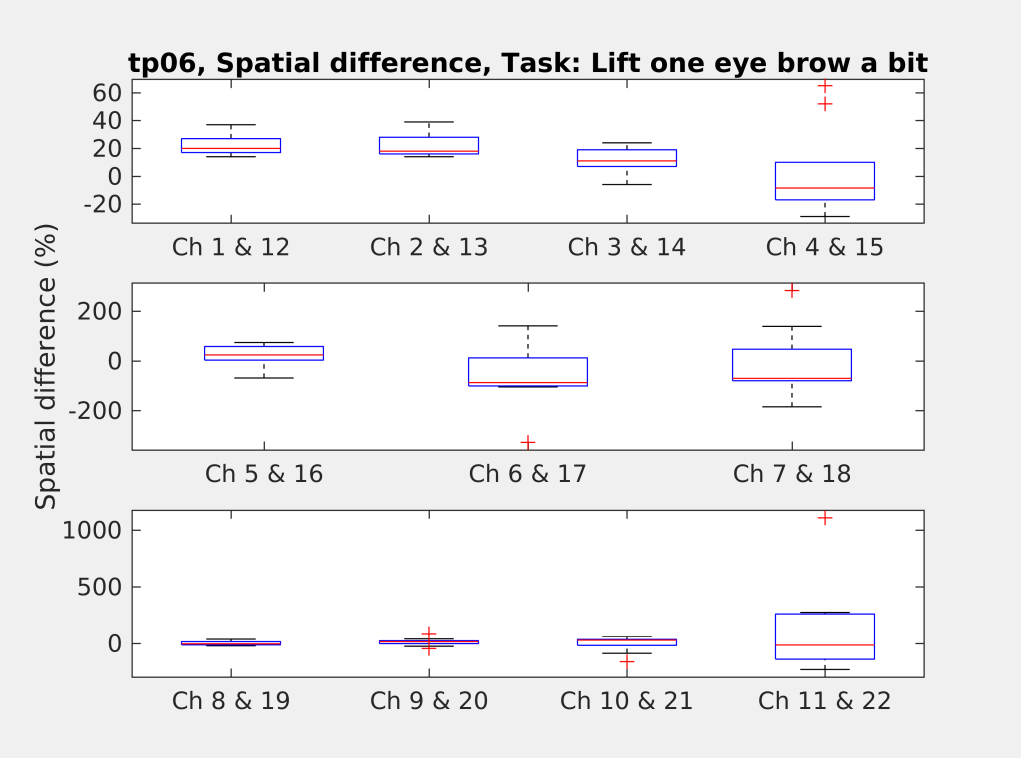

Figure C.20 *Contralateral spatial difference for repetitions of unsuccessful asymmetrical eyebrow lift at medium intensity, test person 6.*

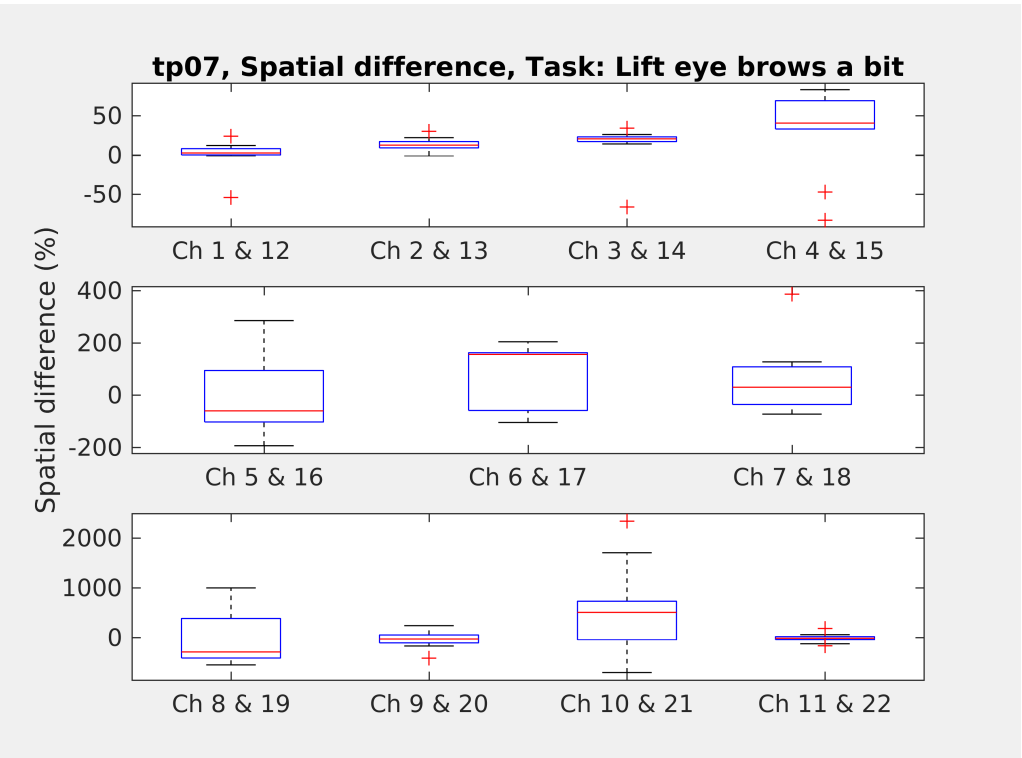

Figure C.21 *Contralateral spatial difference for repetitions of symmetrical eyebrow lift at medium intensity, test person 7.*

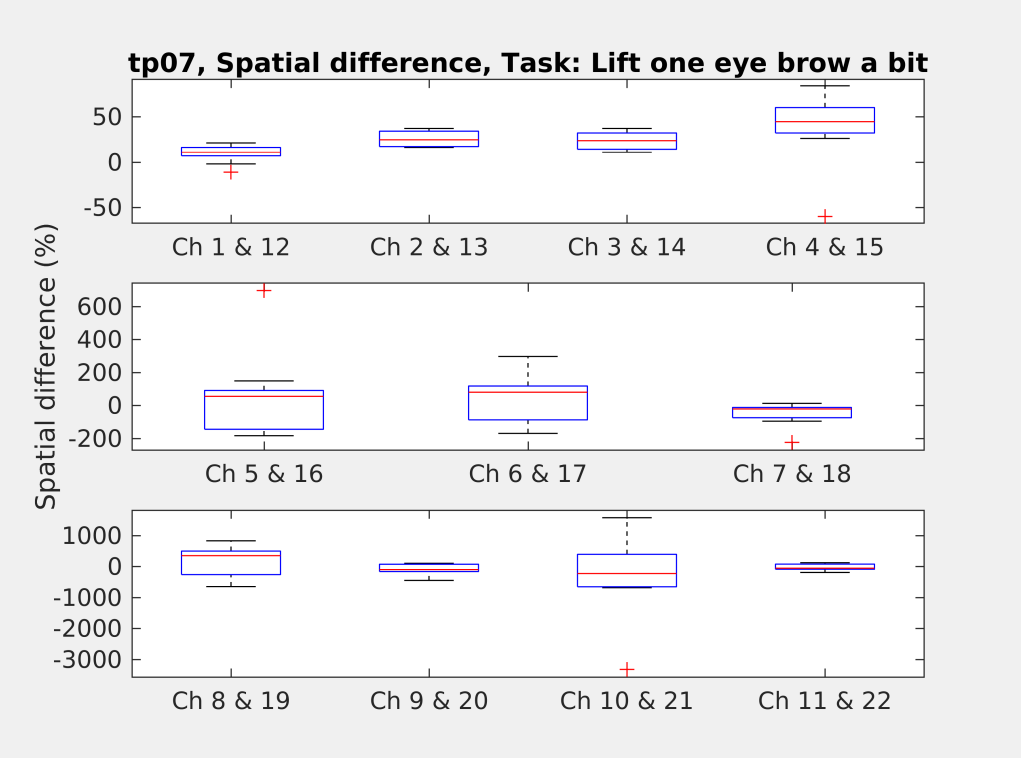

Figure C.22 *Contralateral spatial difference for repetitions of unsuccessful asymmetrical eyebrow lift at medium intensity, test person 7.*

APPENDIX D: PATIENT STUDY'S PARTICIPANTS

Table D.1 lays out the few notes written during the patient measurements. The fitting of the prototype is commented in that table. If a participant does not have a note present, then there was nothing particular in his/her prototype fitting. In Subsection 4.6.1, general overviews of the data collection step and participants are given.

Table D.1 *Notes on prototype fitting on patient test participants 21-37.*

No.	**Notes**
tp30	Left side, the lowermost extension, the third sensor reading from the middle of the face touching (sensor 21)
tp33	Very good fitting, quite much space between the sensors and face though
tp34	Did not reach very far to the center of the face
tp36	Sensors placed very near to the face

The following Tables D.2 and D.3 contain the patient study's test participants' details.

Table D.2 *Necessary details and anonymization of the patient test participants 21-29.*

No.	Gender	Age	Year	Aetiology	Operations	Weight
tp21	Male	60	1998	Choleastoma operation	Temporalis transposition, Brow lift x 2	-
tp22	Female	82	2014	Parotis carcinoma operation	Operated 3 weeks ago: swelling present, Botox: eye and cheek	Pt
tp23	Male	67	1973	Idiopathic	Operated 7 weeks ago, Botox: eye, cheek, mouth	Au
tp24	Female	18	1998	Congenital	-	-
tp25	Female	59	1997	Acoustic neurinoma operation	Operated 2 months ago: Pt weight, Botox: eye, cheek, mouth, Several nerve transfers incl. crossover	Pt
tp26	Female	44	2015	Ear infection	-	-
tp27	Male	57	1996	Post choleastoma operation	Nerve transfer, Temporalis transposition, Gracilis transfer	Pt
tp28	Male	26	2006	Post choleastoma operation	Reanimation, Orbicularis plication	-
tp29	Female	90	2007	Acoustic neurinoma operation	Canthopexia, Blepharoblastia, Botox: eye	Au

Table D.3 *Necessary details and anonymization of the patient test participants 30-37.*

No.	Gender	Age	Year	Aetiology	Operations	Weight
tp30	Female	54	1992, 2014	Parotis carcinoma operation I, Parotis carcinoma operation II	Nerve transfer: crossover	-
tp31	Female	73	1976	Post skull base hemangioma operation	Orbicularis plication, new weight, Botox: eye	Au
tp32	Female	61	1978/01-2016	Post brain-stem hemangioma operation, renewed hemangioma	Lateral tarsal strip (eye), Botox: eye	-
tp33	Male	64	2014	Salivary gland tumor	Nerve reconstruction	Au
tp34	Male	69	2015	Acoustic neurinoma operation	Lateral tarsal strip, Botox: eye	-
tp35	Male	60	2015	Acoustic neurinoma operation	-	-
tp36	Female	46	2015	Acoustic neurinoma operation	-	-
tp37	Female	66	2008	Meningeoma	Botox: eye, Muscle membrane graft (forehead)	-

APPENDIX E: SOFTWARE SPECIFICATIONS

This appendix contains software specifications to complement methodology Subsection 4.6.2. At first, the functionality of the application concentrating on the measurement-mode is described. Next, the main demands for the application's fault tolerance are covered. This appendix does not intend to be a specification document but to provide general overview to the central requirements. The purpose of this appendix is to address the objective of this thesis.

The software described is to replace the existing application and the Matlab demo. The idea is to get a program that is specific purpose application and does both measuring and analysis of the data. Thus, few Face Hugger features will be ignored. There **main features** of the application are listed next:

1. **Selecting the mode**. The user should be able to choose between two main functionalities: to conduct measurements, or to analyze existing data.
2. **Configuration step**. When the **user selects to conduct measurements**, the user should be able to select values for all the parameters from a single view. The user should be able to define values for:
    - Which side is paralyzed, left or right?
    - How many repetitions of each movement is measured. However, a minimum of three repetitions is set.
    - The language for instructions to a patient: Finnish, Swedish or English.
    - Which channels are measuring? Default settings are placed and a warning should appear if the user attempts to change the recording channels.
3. **Offset measuring- and test step**. When the user has selected to conduct measurements and has set the values for parameters, the user should be able to calibrate and test the system from a single view.
    - The calibration should be optional. If no new calibration data is measured, an already existing default calibration data will be used. If the user attempts to conduct calibration data measurement, a warning should appear.
    - The channel test is mandatory and should be visualized well and thus a simple matter to perform. For example, a facial contours at the background and channels at approximately correct locations. When a single channel is tested, the user waves their finger in front of the channel and checks from the visual that for example a bar increases at the specific channel.

4. **Measurements step**. When the user has selected to conduct measurements and has set values for parameters and calibrated and tested the channels, the measurements begin. During the measurements, a single view for the patient and a single view for the user should be shown.

    - The patient should see only the instruction text either to perform a task or to relax.
    - The user should see:
        - The instruction that is currently shown to the patient.
        - How long the instruction has been on the screen in seconds.
        - How long the overall measurement has been on.
        - How long the overall measurement still has time left.
        - How many smile-movements have been done and how many are left.
        - How many eye-brow-movements have been done and how many are left.
    - The user should also have an option to neglect past or present movements in case the patient missperforms (for example starts a wrong movement), the prototype clearly moves on the head or something else surprising and data compromising happens. In case of neglecting a measurement, the program should give one extra instruction for the patient for each neglected measurement to guarantee enough data to be collected.
    - The user should be able to exit the measurement. The user should be asked if the data is saved or not.
    - The user should be able to pause the measurement and to continue the measurement after the pause. If the pause-function is used, the current instruction to the patient is carried on.
        - If the current instruction was to relax, the pause occurs after the relax-period.
        - If the current instruction was to perform a movement, the pause occurs after the relax instruction has been shown.
        - In other words, there is always a full relax-period before the pause occurs. Thus, it should be signaled to the user somehow that the pause is about to happen after the button is pushed.

- When the restart occurs, the first instruction is relax. In this way, the pause is surrounded by relax-commands on both sides of the pause, before and after, and filtering the data is successfull.

There is no limit to the pause time. When the pause is happening, the pause-button should be marked as a restart button.

5. **Results step**. After the measurements are ready, the results are visualized in their own single window. The results are presented separately for the different movements - now smile and lifting eyebrows. Results based on each parameter are represented. The data is also saved in a form that is easy to observe even with naked eye.

If the **user selects to analyze existing data** in step 1, the steps 2-4 are replaced with analysis specific steps. To put it briefly, at first the user chooses if they want to analyze clinical correlation (in other words, study multiple patients simultaneously and especially the inter-patient trends) or a single patient's development (in other words, the intra-patient status over time). The next step is to import the data into the app. Then, the results are represented (step 5 from measurements mode).

The faults that should be tolerated are listed next. This brief list is not perfect but instead concentrates on the situations that are not covered by the existing solution. The **faults that should be tolerated**:

1. **Bluetooth connection lost**: The data measured so far is not lost. It should be possible to continue the measurement where it is left. If a movement is interrupted, it should be asked again and the partial data ignored. Otherwise, the program should respond just as the pause button was pushed.
2. **Not enough hard disk space**: This should be checked before starting analysis or measurements and a warning given to the user if needed.
3. **A channel/channels not recording**: This should be checked during the test step of the measurement mode's flow (feature 3 in this current section.) Give warning to the user if needed.

APPENDIX F: USER INTERFACE

This appendix contains notes and figures related to the GUI and usability to complement Subsection 4.6.2. Section F.1 lays out factors that should be addressed in usability. Section F.2 contains figures of the wireframe demo used as a basis for usability testing and application development. Finally, Section F.3 describes concisely the main points considered in usability testing. The purpose of this appendix is to provide necessary information to address the objective of this thesis. This appendix does not aim to be be a complete plan or documentation of usability, but provide the necessary details within the scope of this thesis.

F.1 Usability

This section contains a brief listing of usability factors and general requirements that are requested to be addressed. The end-users are engineers and medical professionals such as surgeons and nurses.

Usability factors that should be considered:

1. **Preselected default settings**. Default settings should be possible to restore easily. A warning is shown if the user attempts to change critical settings such as which channels record, or to measure new offset removal data.
2. **Intuitive usability**. No extensive training should be needed to use the application. This requirement can be further divided into several factors such as:
    - **Instructions** about functionalities behind a symbol such as question mark that is located in close proximity of the functionality. For example, hovering over a question mark close to the mode selection buttons would show a tooltip describing the modes.
    - **Colors**. Using intuitive colors.
    - **Simple navigation**. In the current software the preparations for the measurements are done in several tabs that need to be found. This is something that should be avoided; this is a specific purpose application and thus irrelevant features can be excluded. For example, own window for each step (steps listed in Section E) in a logical order is an option. In that case, the moving between the steps should be enabled and filled information and chosen settings should be preserved if revisting previous steps.

3. **GUI language** should be Finnish. The doctors' working language is Finnish. However, the language demand may change in future.

F.2 Graphic User Interface Wireframe Mock-Up

This section represents GUI wireframe images. There is an image for each step described in Section E. The figures are shown in chronological order. As explained in Section F.1, the GUI should be in Finnish.

Figure F.1 visualizes the **selection of the mode**.

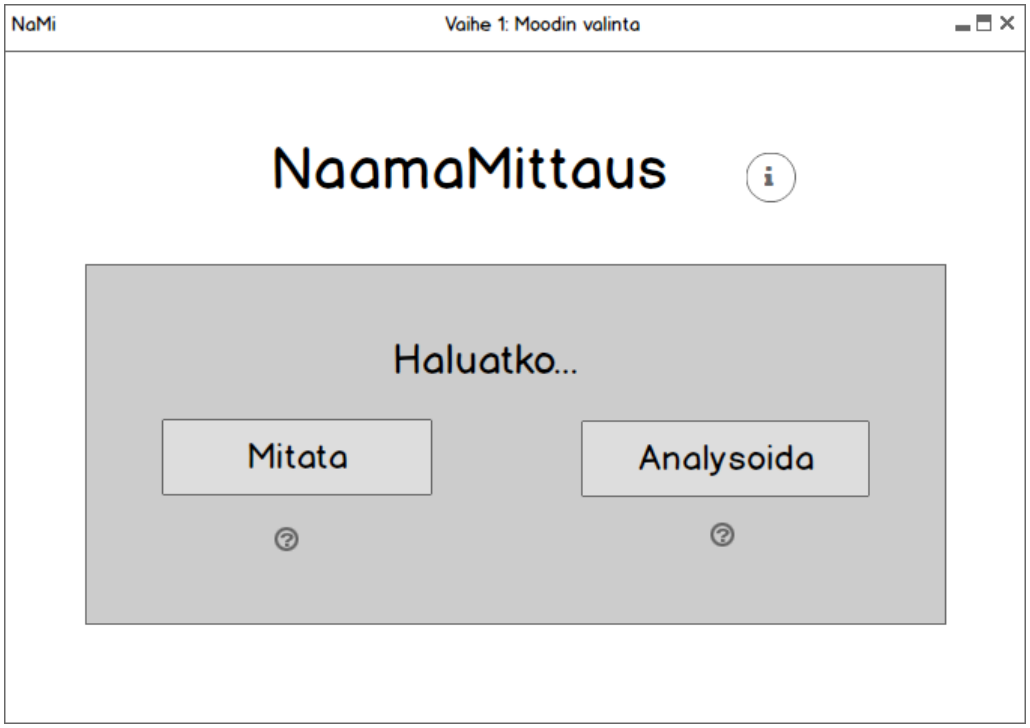

Figure F.1 *GUI wireframe image: mode selection.*

Figures F.2-F.4 show the **measurement-mode steps**.

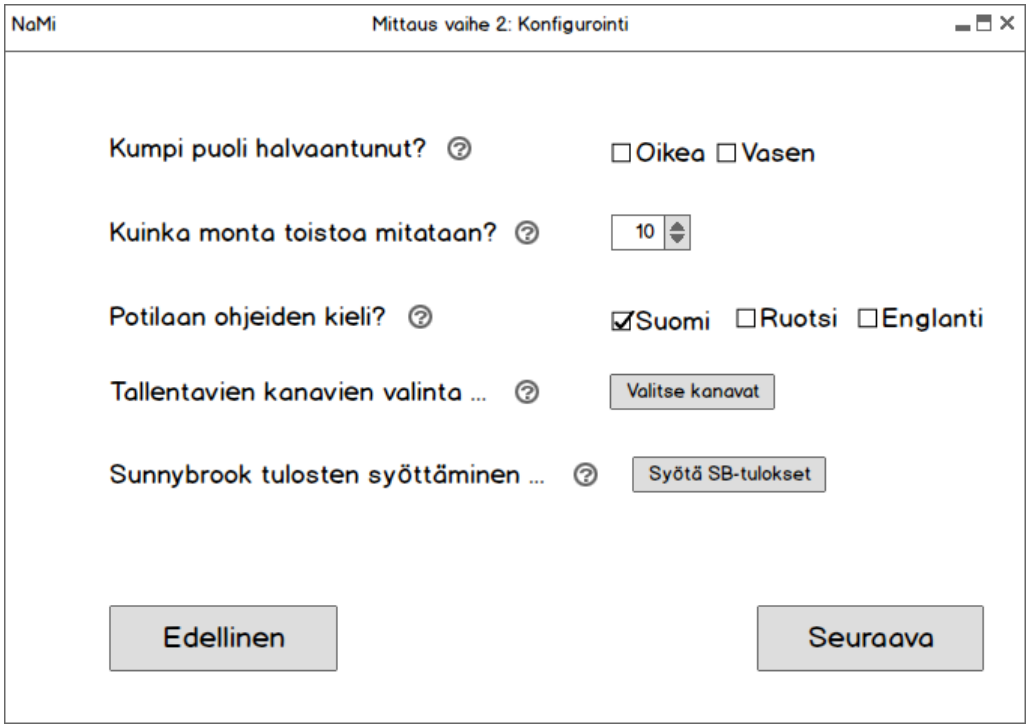

Figure F.2 *GUI wireframe image: configuring measurement settings. Channel choosing and Sunnybrook result filling open their own pop-up windows.*

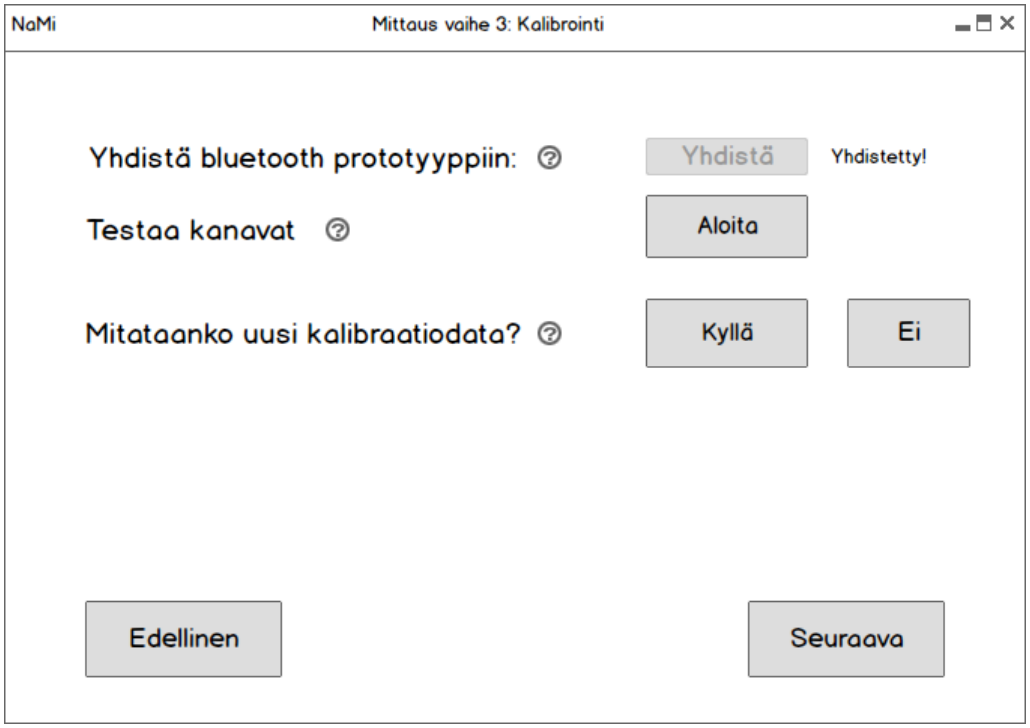

Figure F.3 *GUI wireframe image: connecting to the prototype (step 1), testing channel functioning (step 2), and measuring offset data (optional step 3). The buttons are enabled once the previous ones are pressed and successfully completed.*

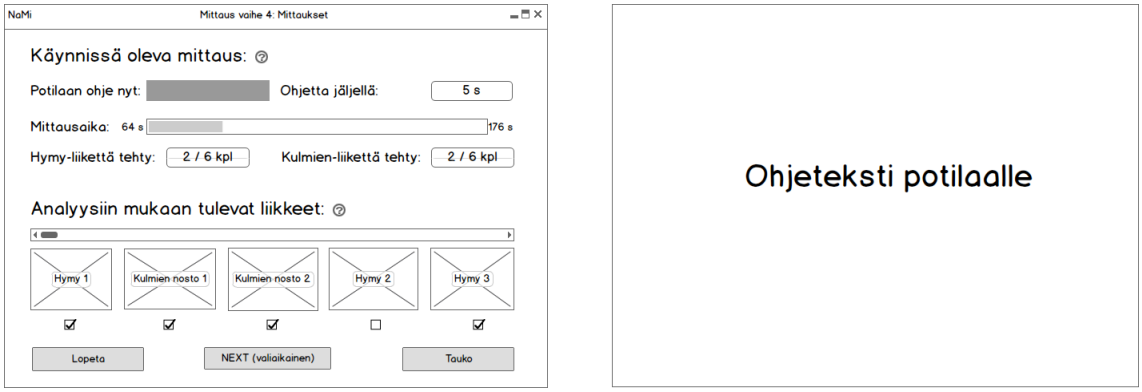

Figure F.4 *GUI wireframe image: measuring. There are two views; one for the person conducting the measurement, and one for the patient.*

Figure F.5 visualizes **the final step for both measurement- and analysis mode**.

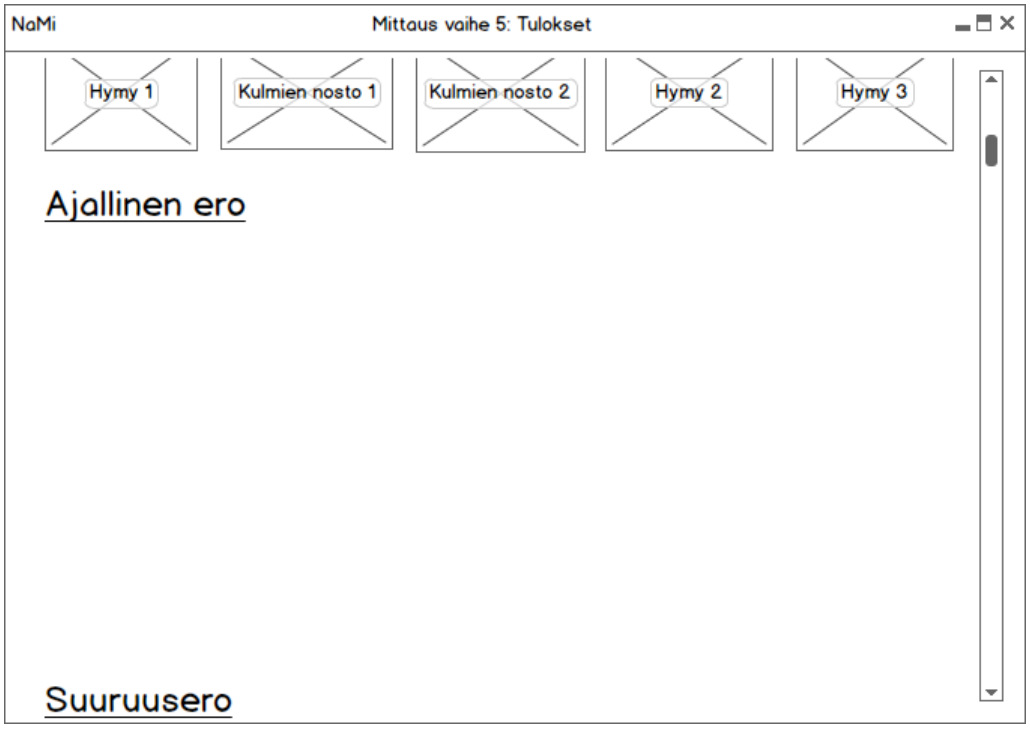

Figure F.5 *GUI wireframe image: results. The details are not illustrated.*

Figures F.6-F.8 show **the steps specific for the analysis-mode**.

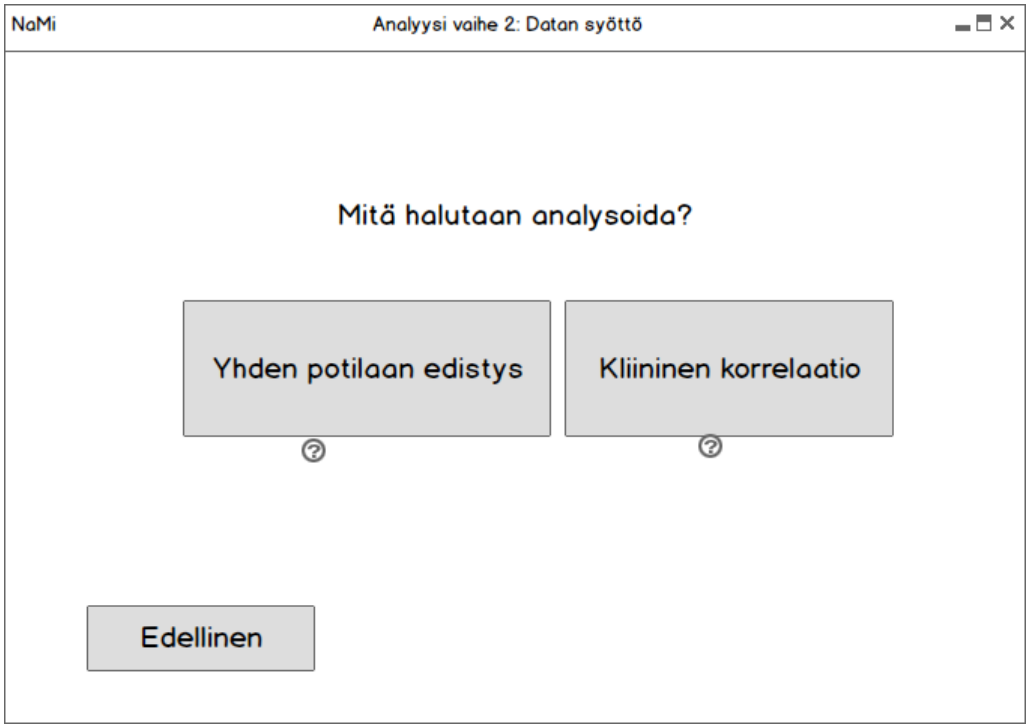

Figure F.6 *GUI wireframe image: choosing whether to analyze clinical correlation or a single patient's development.*

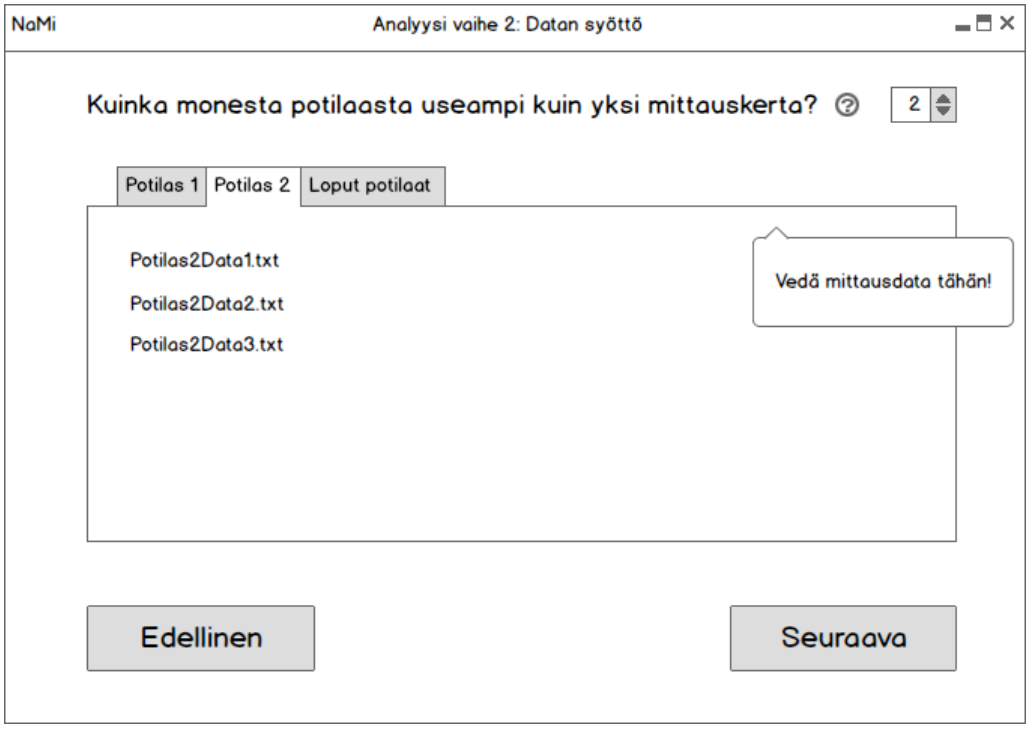

Figure F.7 *GUI wireframe image: importing the data into the app for clinical correlation analysis.*

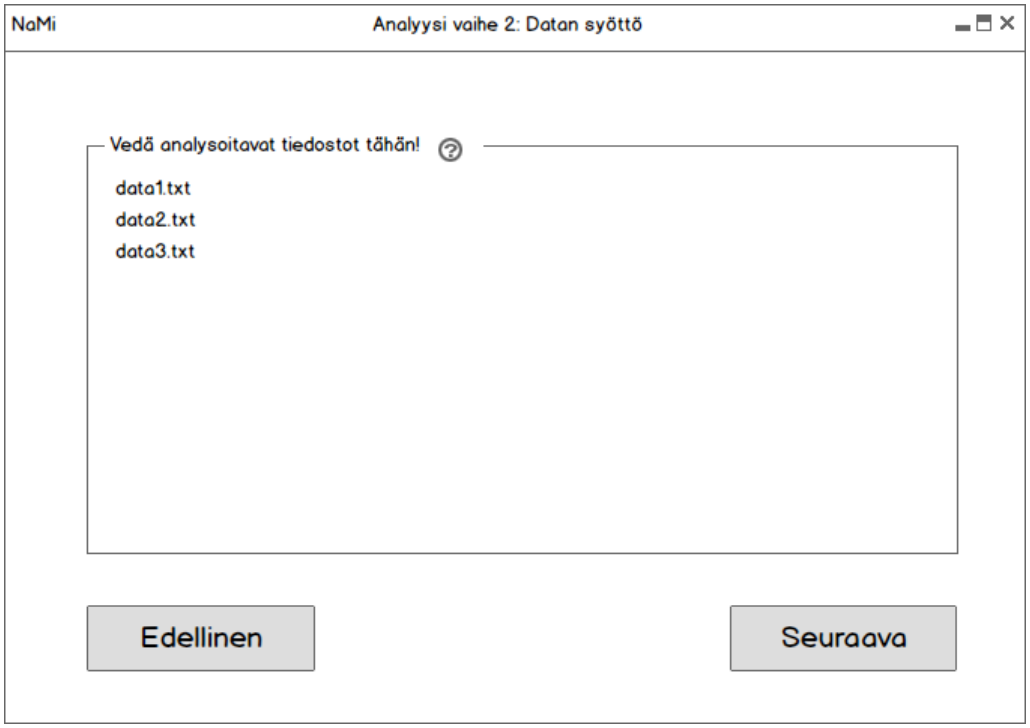

Figure F.8 *GUI wireframe image: importing the data into the app for single patient's development analysis.*

Combining these wireframe images into a clickable demo gives us a mock-up to be used in usability testing (see the next Section F.3) and development.

F.3 Usability Testing

This section details the usability testing. As this thesis is a part of a larger project that includes both medical doctors and engineers, proper **User Experience** (UX) should be verified for both professions. As the project workers represent the end-user groups of the software, they are employed in the user research as test participants. One participant from both user groups is involved.

To answer the objective of this thesis within the scope suitable for a thesis requires delimiting the coverage depth of different steps. For example, a usability research could be its own thesis topic and thus is completed in a very brief and shallow manner here. However, the phase names of UX steps are used here for communication purposes even though they are not even intended to be completed as such.

Due to the preliminary work such as literature research, measuring healthy volunteers and having discussions with medical doctors and experienced engineers, the

discovery and strategy parts of a typical UX project get completed as a side product. The so called **mental model** [116] and **UX objectives** [117] are clarified as a result. Thus, the GUI design and implementation steps are left to be conducted [118].

To measure the **success of the design**, usability testing in **scavenger hunt** form [119] is chosen. The scavenger hunt usability testing is based on giving pre-determined tasks to a user and observing that at which stages the design is helping, slowing down or frustrating the user [119]. By noting the difficult spots the design can be further improved [119]. The typical downside of scavenger hunt usability testing is that the tester should be certain that what does the user need and do with the product as those determine the scavenger tasks [119]. Due to the nature of this project those tasks (measure and analyze) have been definitive from the very beginning, thus the test type suits this project well.

In more detail, the mock-up demo visualized in Section F.2 is used for the usability testing. The scavenger usability test takes place in a calm room (to mimic a situation where a patient is being measured by a doctor).The user is informed briefly:

- That the feedback is for improving the GUI design.
- The purpose of the software is to enable further studies about facial palsy patients' degree of paralysis and the ultimate goal would be to replace the subjective rating scales currently in use.
- The user is told that no education is needed to use the software, it should be intuitive.
- The test user is also encouraged to give ideas and tell any feelings they have when using the software.
- The user is given a possibilility for questions.

Then, the software is given to the user, and the user is asked:

1. To conduct a new measurement.
2. To analyze existing data.

The test user is asked one question at time.

Once the design is approved by the test users to certain extent, the implementation with QML begins.

APPENDIX G: CLASS DIAGRAM

This appendix provides a class diagram to complement Subsection 4.6.3. The class diagram visualizes the main architecture of the developed application. The class diagram is a draft; this is the state that the application is at the end of the second BML-feedback loop.

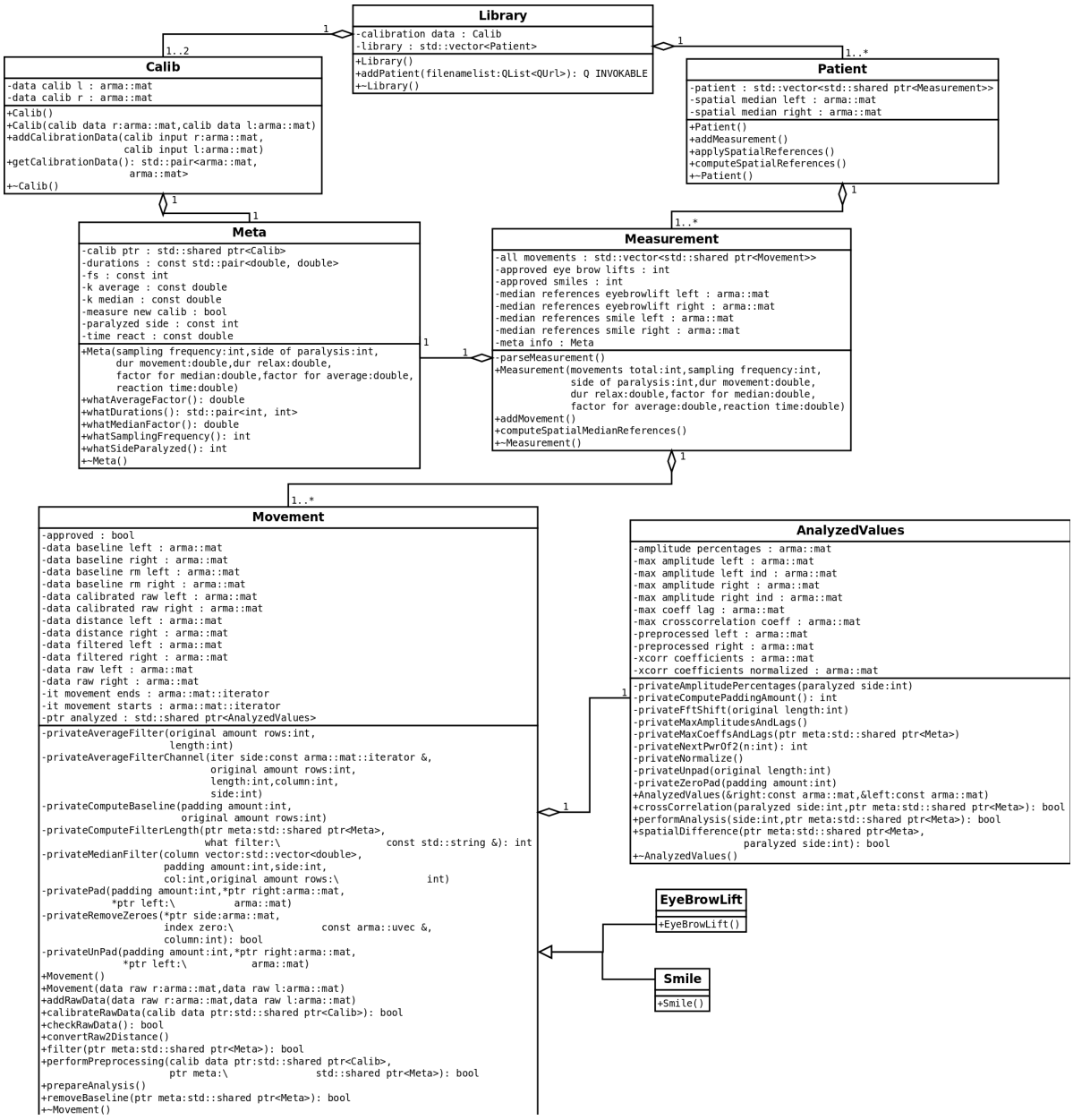

Figure G.1 *The class diagram of the developed application.*

In Figure G.1, the **Library-class** aggregates two classes of different types. Firstly, the Library-class has **Calib-class** that contains the data for the base capacitance removal (see Subsection 4.4.2) which is a preprocessing step. There is at least one Calib-class instance; the default data set. There may be a second additional set

if the user chooses to measure their own. Further on, the Calib-class aggregates Meta-class. **Meta-class** contains the meta information related to a data set. Secondly, Library-class contains the patient data. There are as many **Patient-class** instances created as there are patients. Each Patient-class then stores as many **Measurement-classe**s as there are measurement sessions per the patient. For each measurement session there is meta data stored, hence the Meta-class aggregation into Measurement-class. Also, a measurement session and thus Measurement-class aggregates movements; there are as many instances of **Movement-class** as there are completed movements. The Movement-class is a virtual base class, and the two **inherited classes EyeBrowLift and Smile** reveal the type of the movement. Finally, Movement-class aggregates **AnalyzedValues-class**; for each movement there is one instance of analyzed movement.

In other words, the class diagram shows a **large nested structure**: library consists of patients. Those patients have measurements sessions that consist of meta data and movements. There are two types of movements: eyebrow lifts and smiles. Movements are then analyzed. The library also stores the calibration data set and its meta data.

The class diagram shows the **class variables** and **methods**. There are both **private and public methods**. A class variable and a private method can be distinguished from the "-" symbol, public interface methods have a '+' in front of the function signature. There are private methods, or methods that cannot be called outside the class, to maintain the **single responsibility principle** (SRP). In other words, the public methods implement a complex task that is further chopped into smaller steps. These steps do not need to be visible outside the class. A representational example is Movement-class' filter-function. Once the filter-method is called to a Measurement-object, the function further calls the computation for the filter length (privateComputeFilterLength), necessary padding (privatePad), then the actual filtering function (privateMedianFilter), and finally unpadding (privateUnPad) to cancel the effects of the padding.

It is worthy to emphasize that the class diagram does not show a ready application. Instead, it visualizes the **application state at the pivot point** (see Subsection 4.6.4). At that point, the data preprocessing and analysis excluding the statistics are implemented and tested. Most parts of the GUI are not yet connected to this back-end functionality.

APPENDIX H: ADDITIONAL RESULTS

This appendix provides additional figures to complement Chapter 5. Section H.1 lays out images showing smile repetitions and Section H.2 visualizes eyebrow lift repetitions. These images are provided to illustrate how does the capacitive facial signals of facial palsy patients appear; to the best of knowledge facial palsy patients' movements have not been measured capacitively before. Secondly, these figures picture the data that is used to gain the waveform results (Section 5.2) and the temporal results (Section 5.3). Thus, these figures give context to the computed results.

All figures in this appendix show the preprocessed ready-to-be-analyzed signals. These signals are chopped to contain 8.5 seconds before the movement, the 5 second period when the movement instruction was shown to the patients, and 10 seconds after the movement. The time periods before and after the movement are the relaxation periods. There is a figure for each patient. All ten repetitions are plotted into the same figure. The background color indicates which facial side is paralyzed for each patient; green background is used for the healthy side, and red for the paralyzed side. The amount of channels shown is limited to channels 4-6 for the smiles, and channels 1 and 2 for the eyebrow lifts.

H.1 Smile

This section visualizes all the smile repetitions for each patient. There is one image for each patient and the **figures are ordered according to increasing total Sunnybrook value**. In other words, the first Figure H.1 contains the data of the most severe palsy patient, and the last Figure H.17 the repetitions of the healthiest patient.

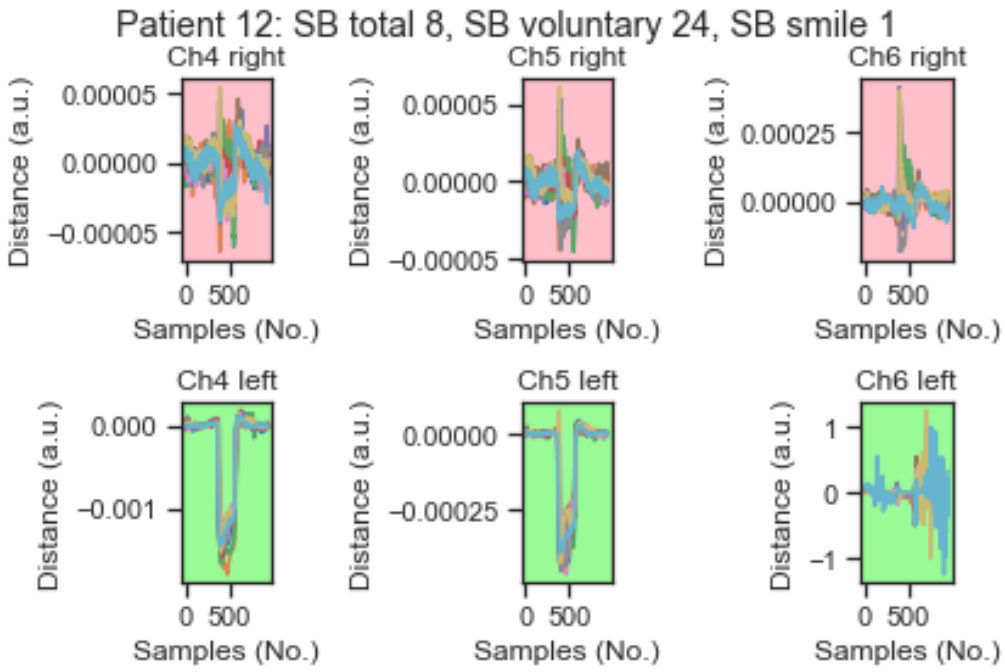

Figure H.1 *All smile repetitions of patient number 12.*

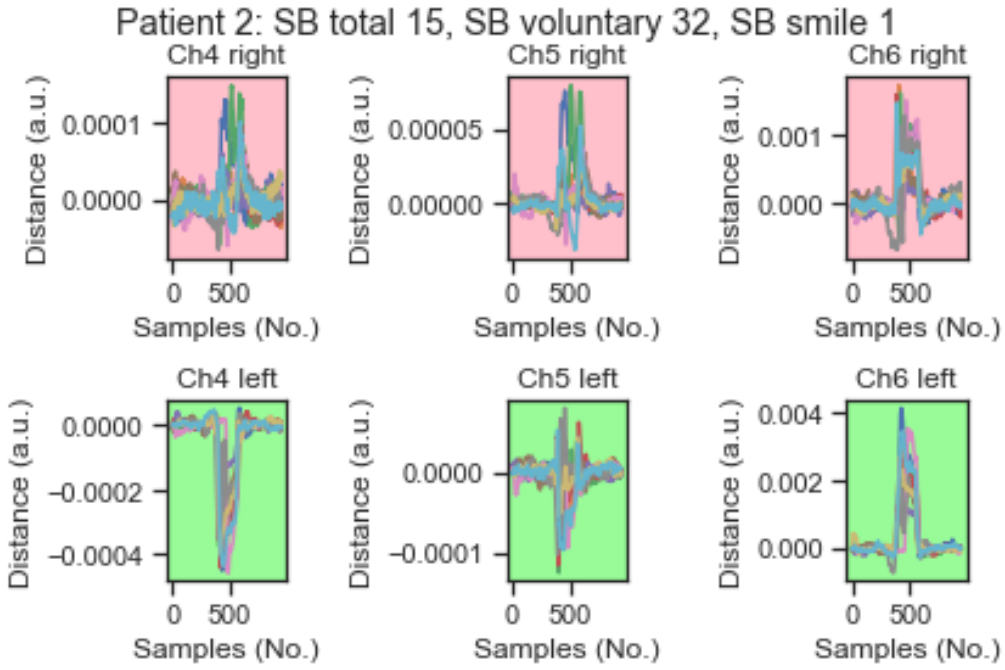

Figure H.2 *All smile repetitions of patient number 2.*

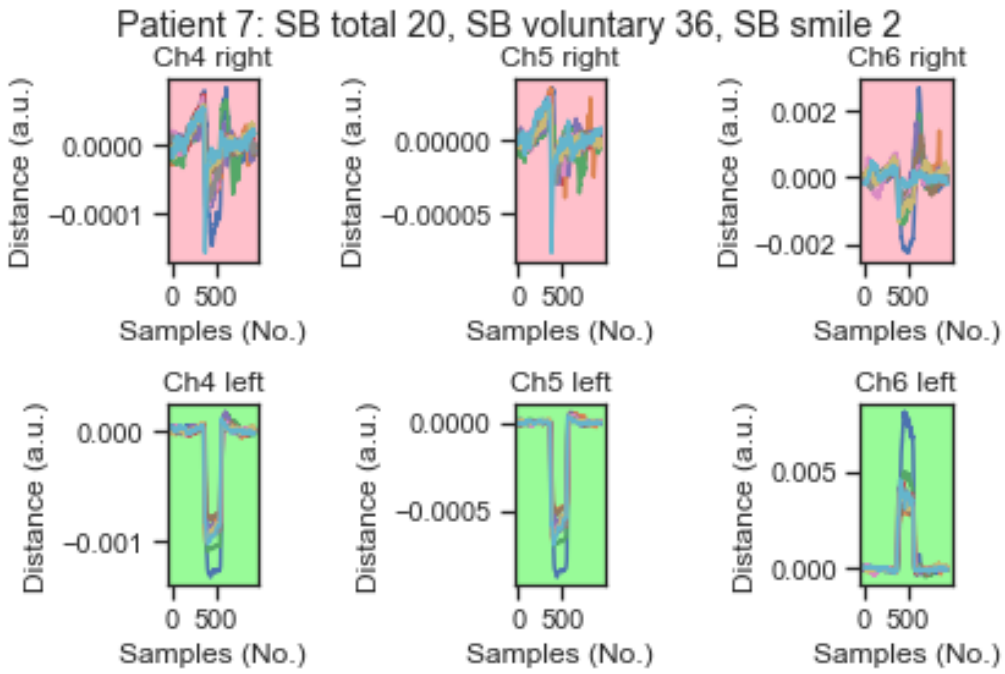

Figure H.3 *All smile repetitions of patient number 7.*

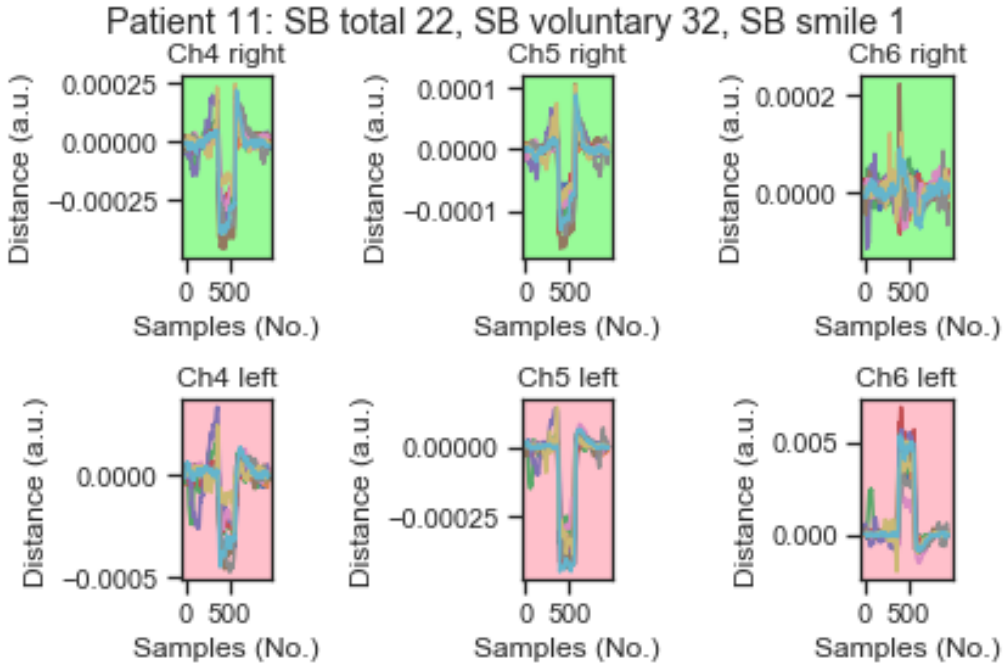

Figure H.4 *All smile repetitions of patient number 11.*

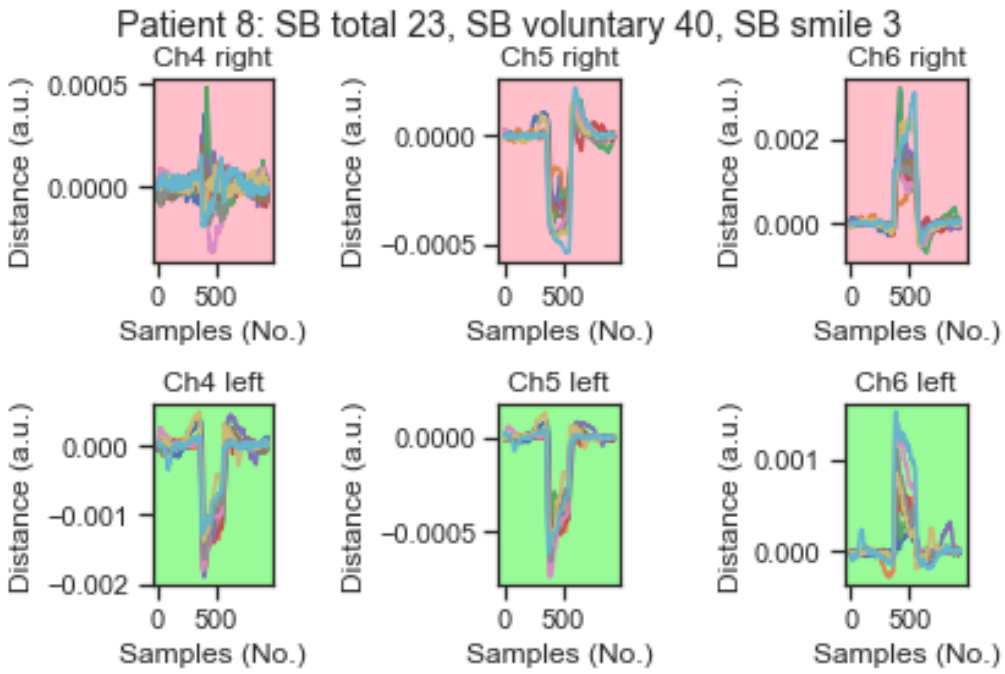

Figure H.5 *All smile repetitions of patient number 8.*

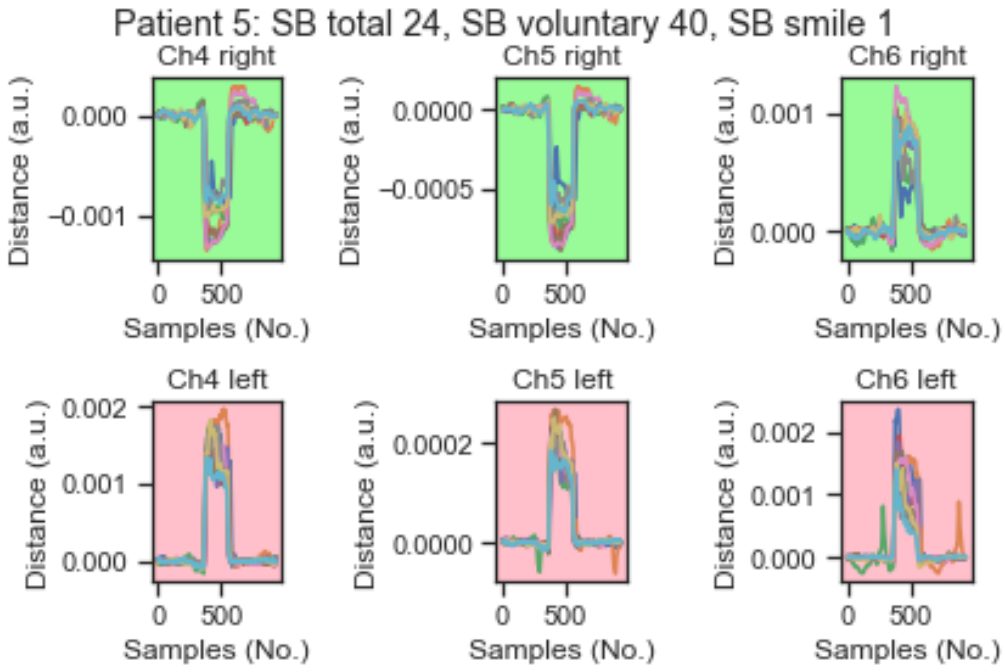

Figure H.6 *All smile repetitions of patient number 5.*

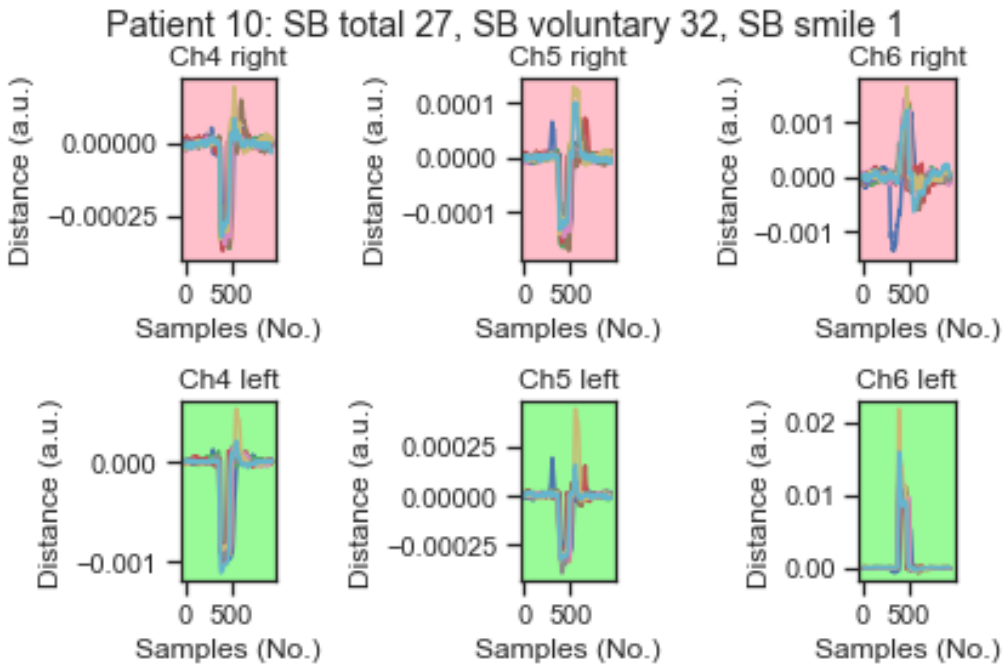

Figure H.7 *All smile repetitions of patient number 10.*

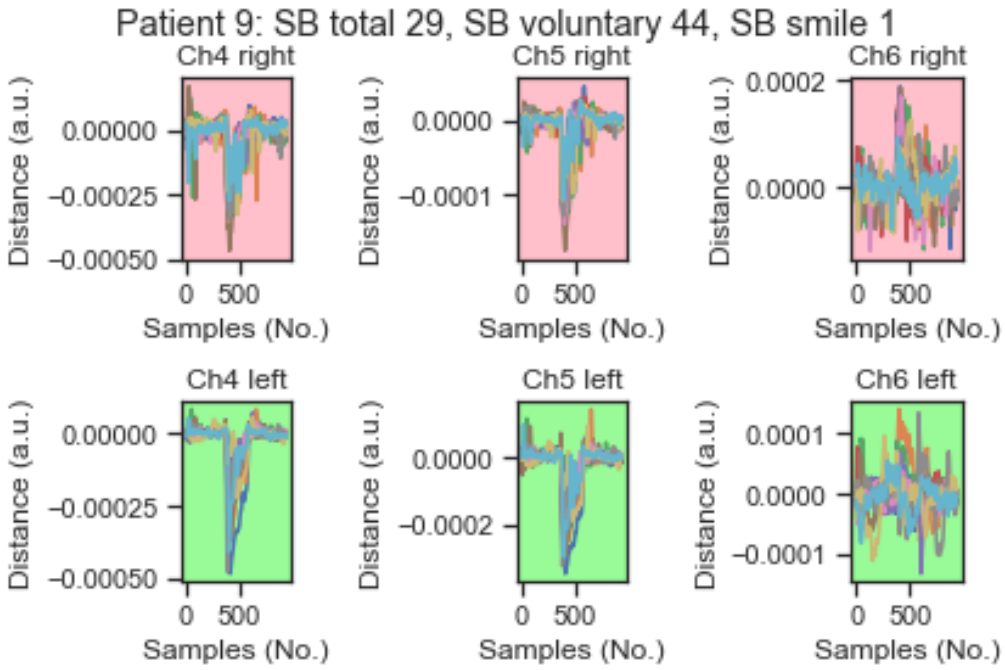

Figure H.8 *All smile repetitions of patient number 9.*

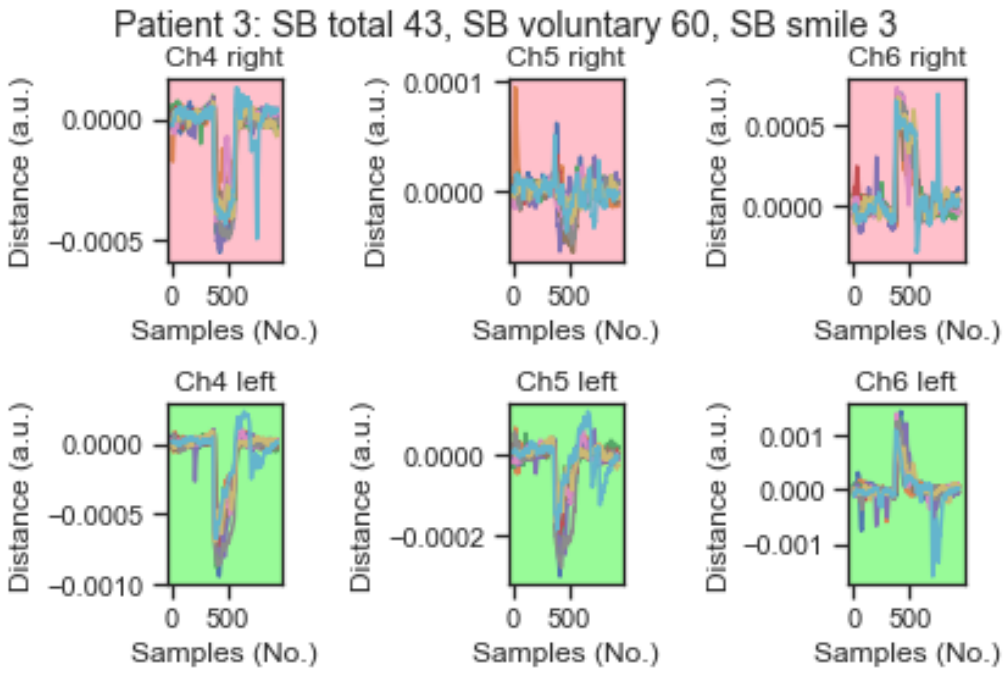

Figure H.9 *All smile repetitions of patient number 3.*

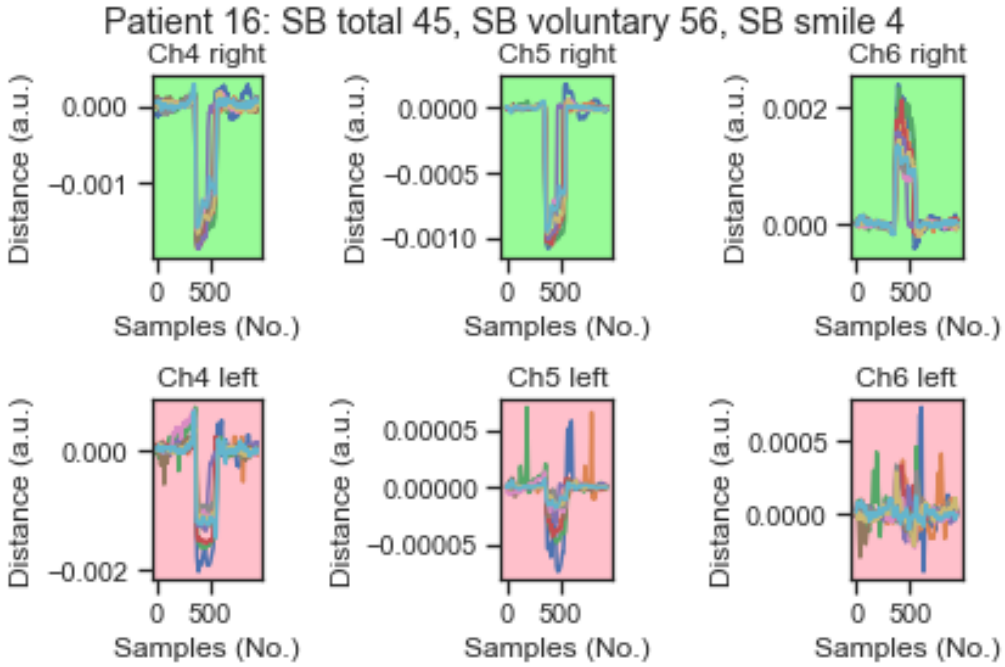

Figure H.10 *All smile repetitions of patient number 16.*

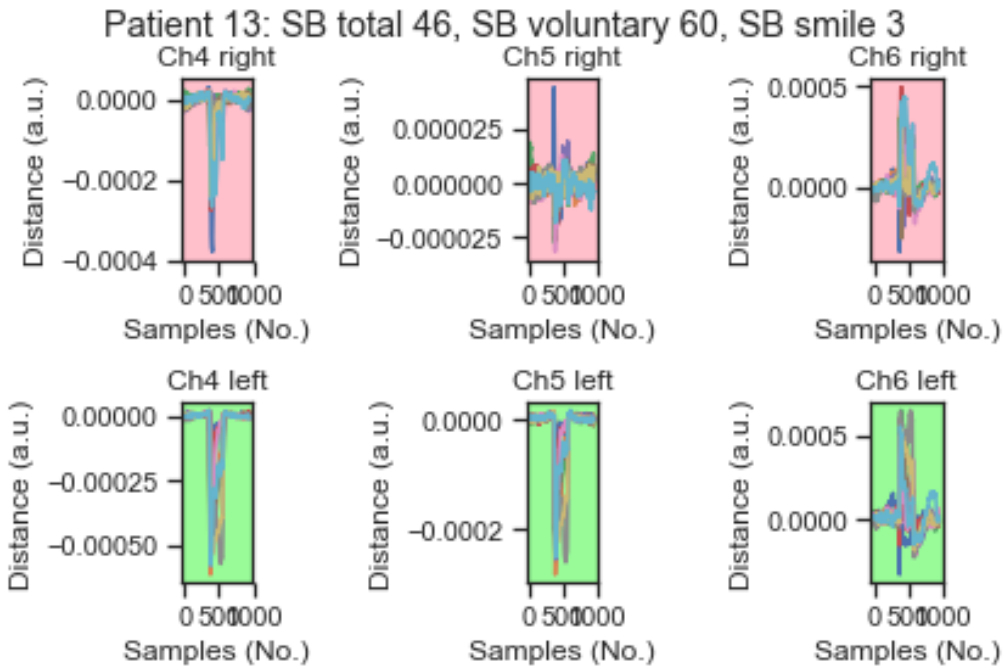

Figure H.11 *All smile repetitions of patient number 13.*

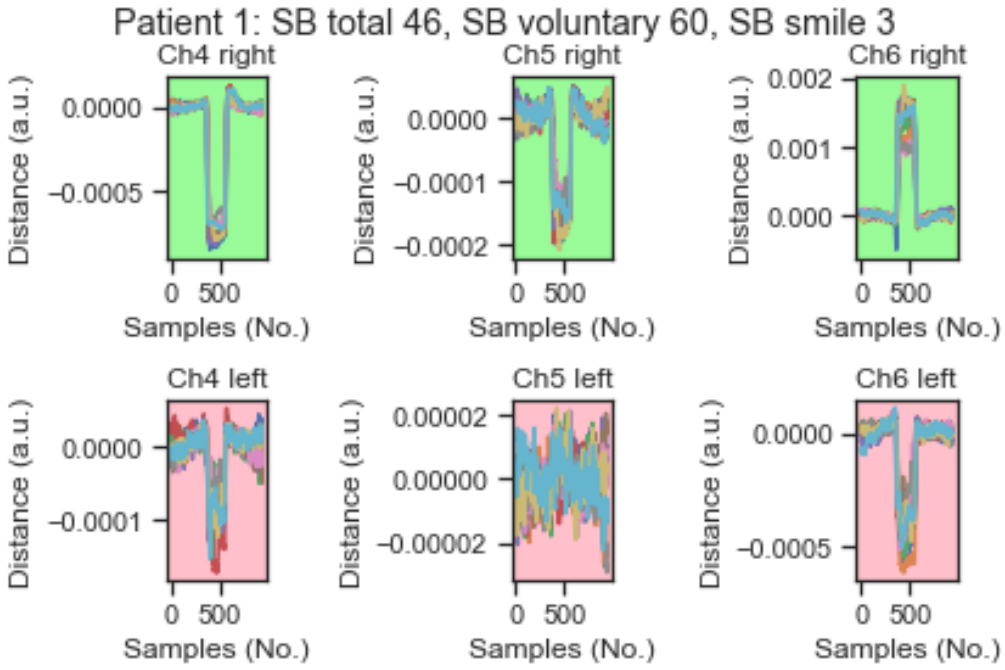

Figure H.12 *All smile repetitions of patient number 1.*

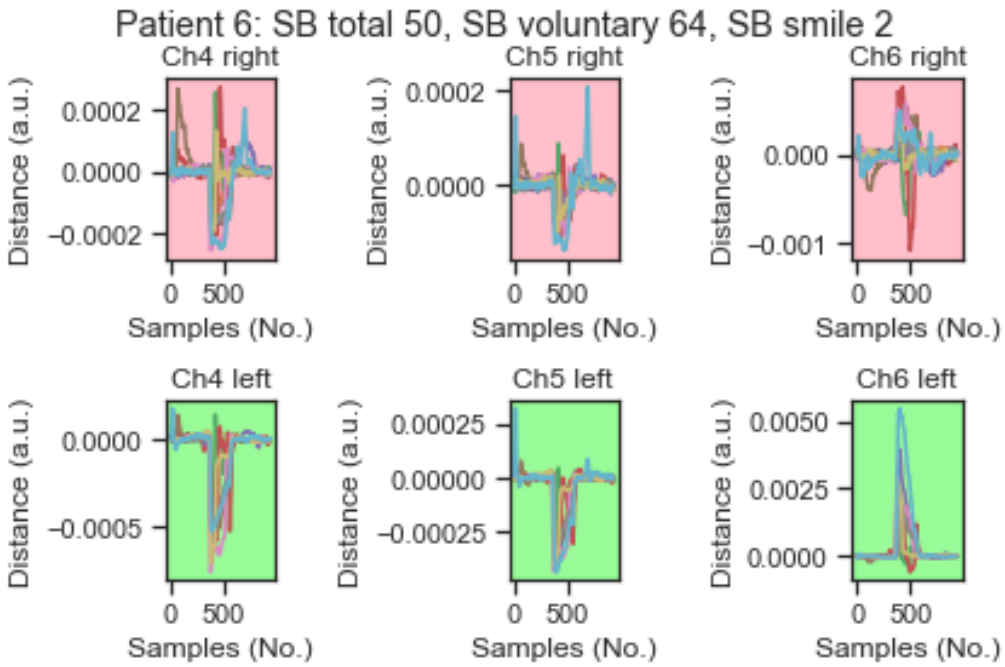

Figure H.13 *All smile repetitions of patient number 6.*

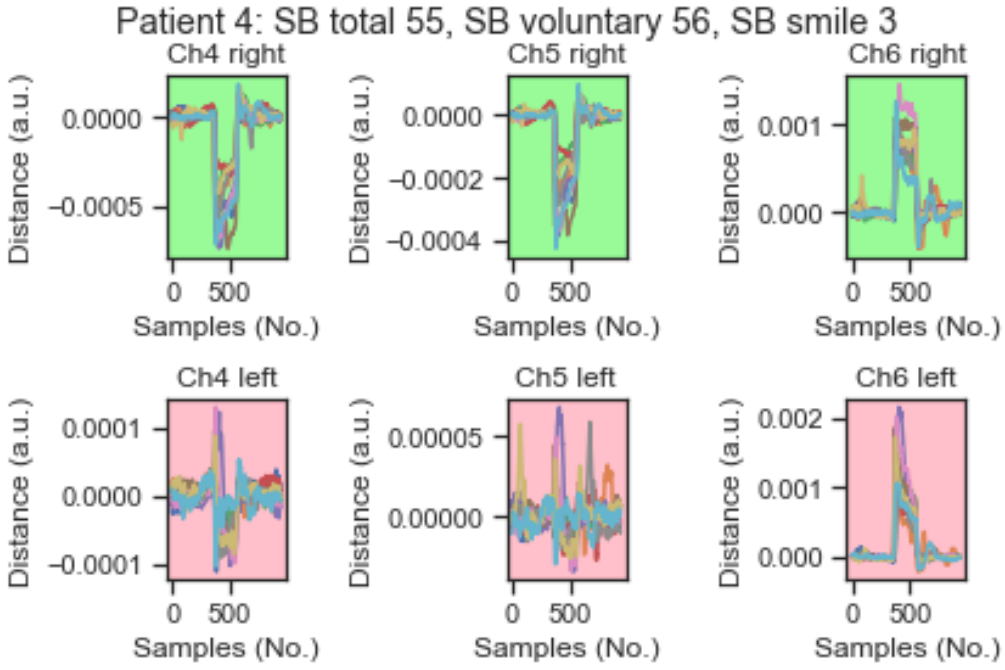

Figure H.14 *All smile repetitions of patient number 4.*

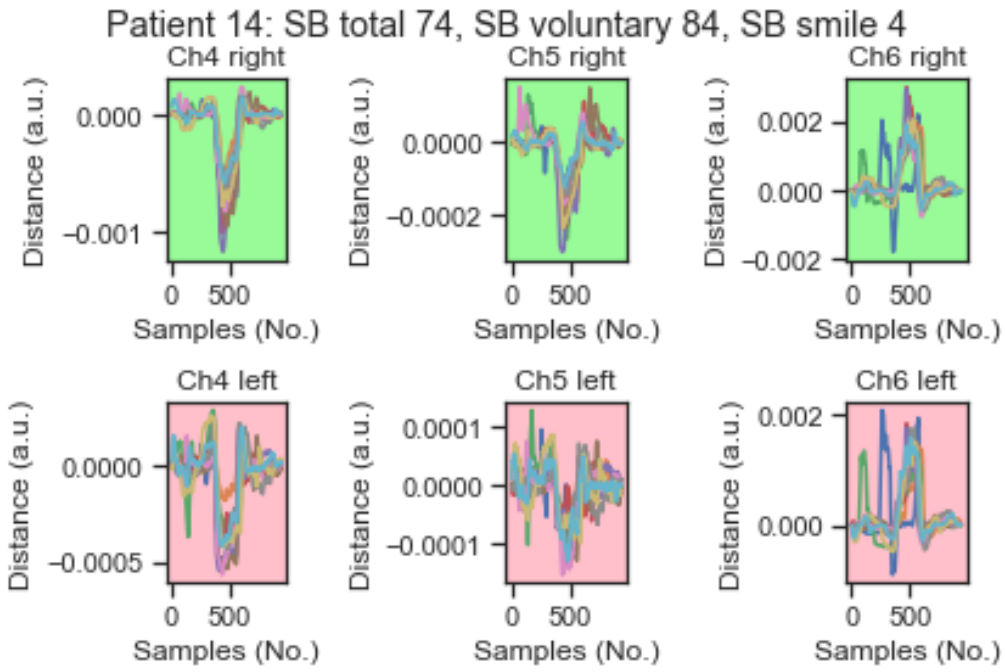

Figure H.15 *All smile repetitions of patient number 14.*

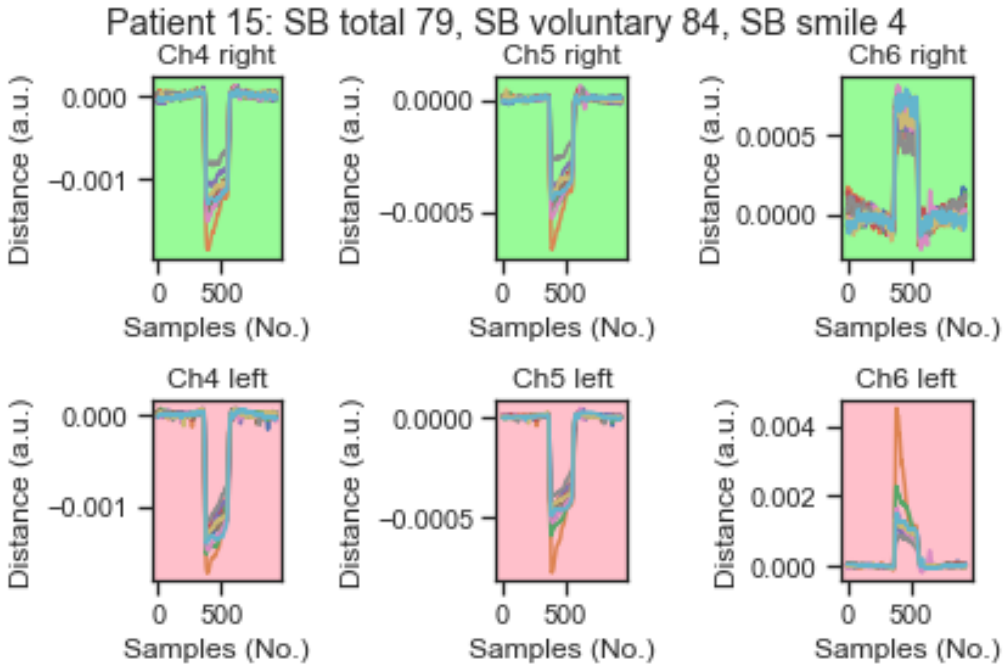

Figure H.16 *All smile repetitions of patient number 15.*

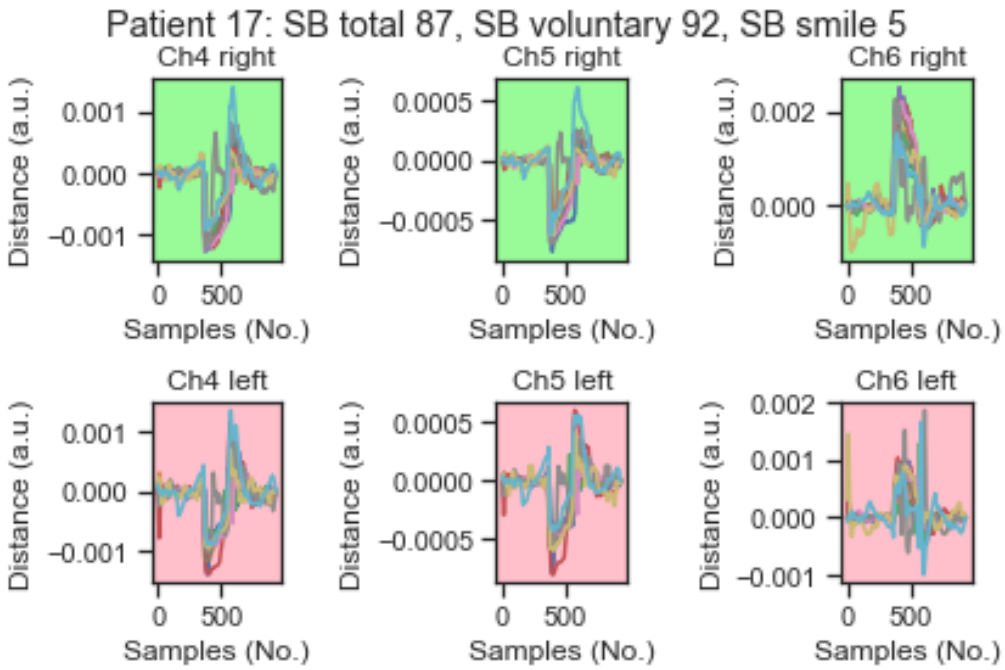

Figure H.17 *All smile repetitions of patient number 17.*

The data shown in Figures H.1-H.17 is used for analysis of waveform and temporal delay.

H.2 Eyebrow Lift

This section visualizes all the eyebrow lift repetitions for each patient. There is one image for each patient and the **figures are ordered according to increasing total Sunnybrook value**. In other words, the first Figure H.18 contains the data of the most severe palsy patient, and the last Figure H.34 the repetitions of the healthiest patient.

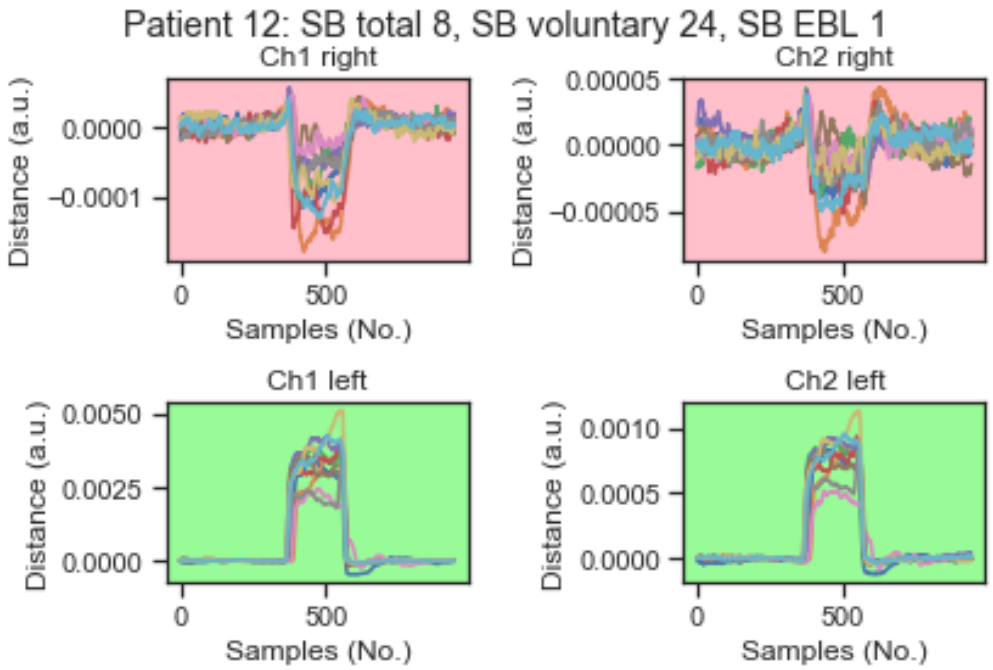

Figure H.18 *All eyebrow lift repetitions of patient number 12.*

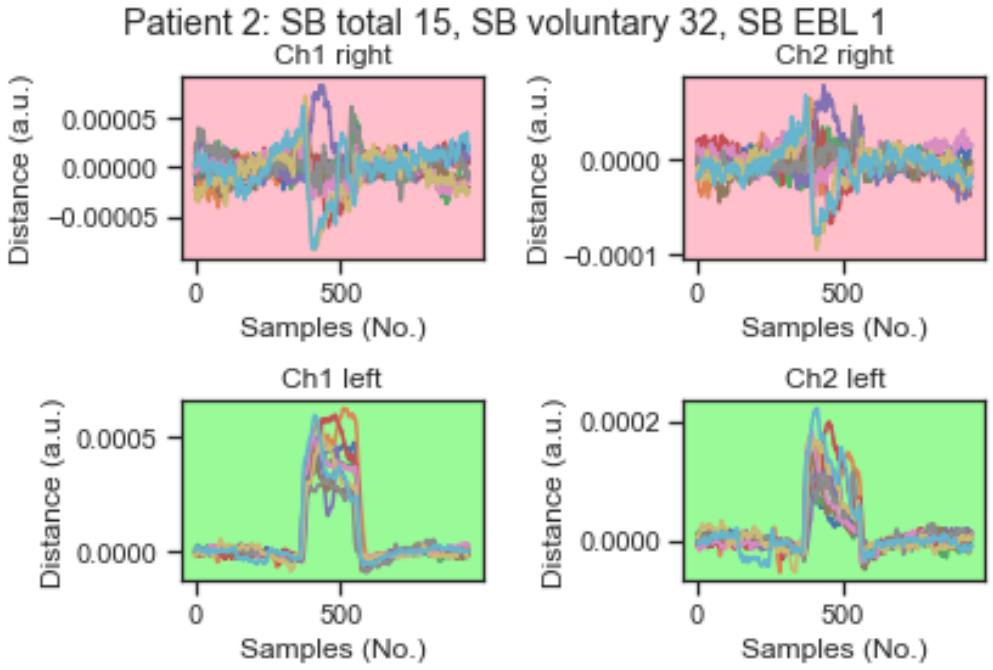

Figure H.19 *All eyebrow lift repetitions of patient number 2.*

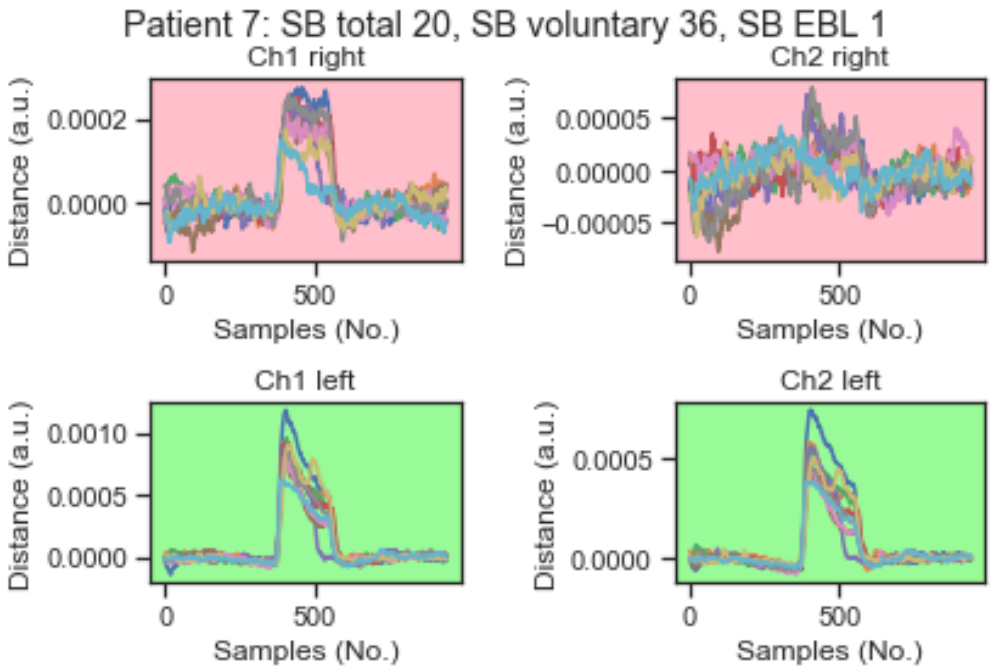

Figure H.20 *All eyebrow lift repetitions of patient number 7.*

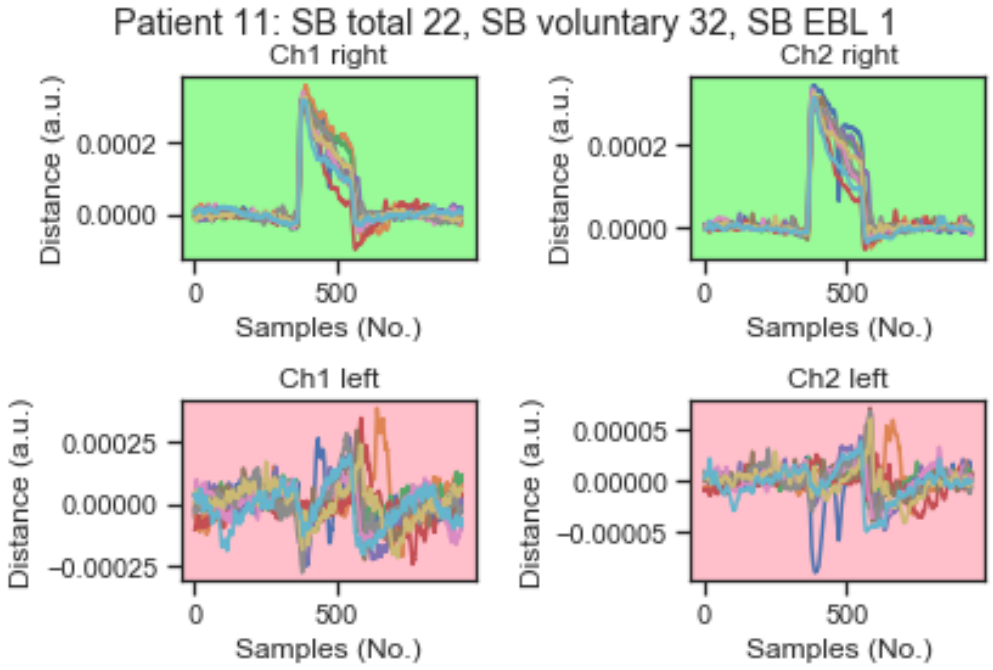

Figure H.21 *All eyebrow lift repetitions of patient number 11.*

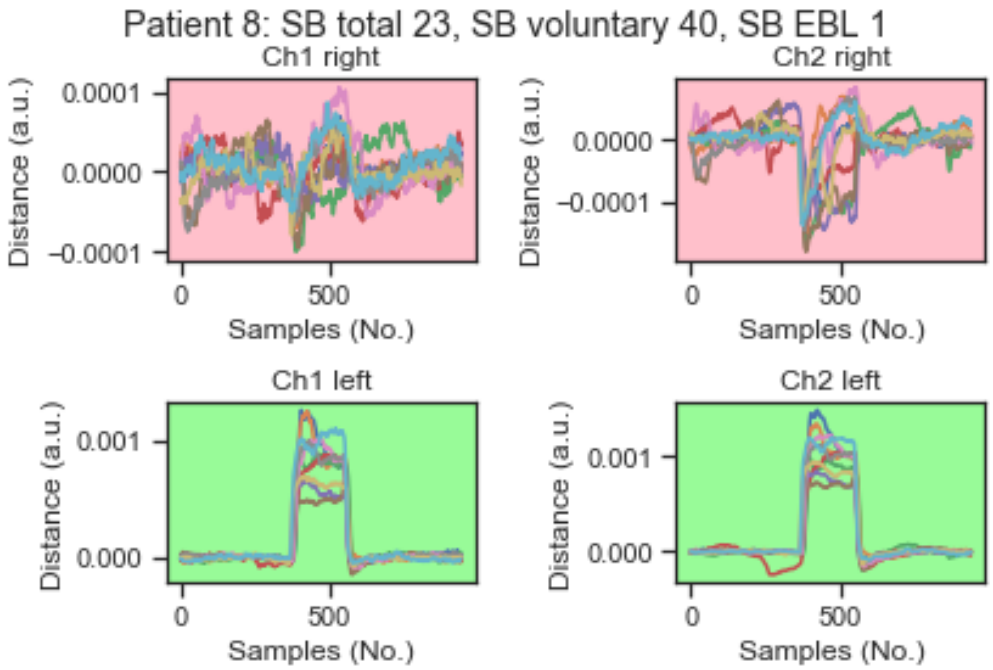

Figure H.22 *All eyebrow lift repetitions of patient number 8.*

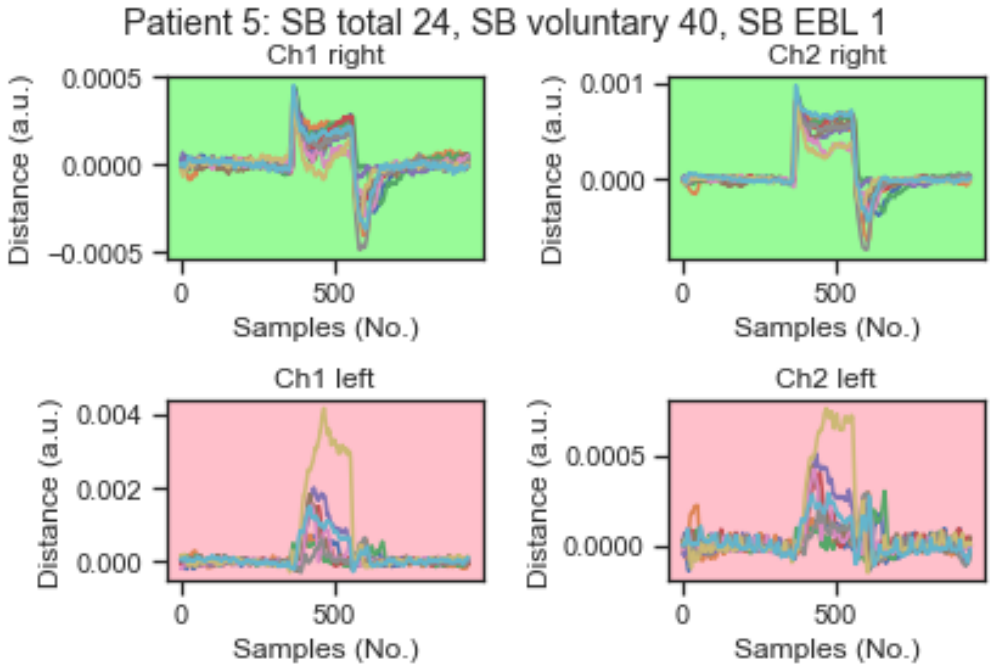

Figure H.23 *All eyebrow lift repetitions of patient number 5.*

Figure H.24 *All eyebrow lift repetitions of patient number 10.*

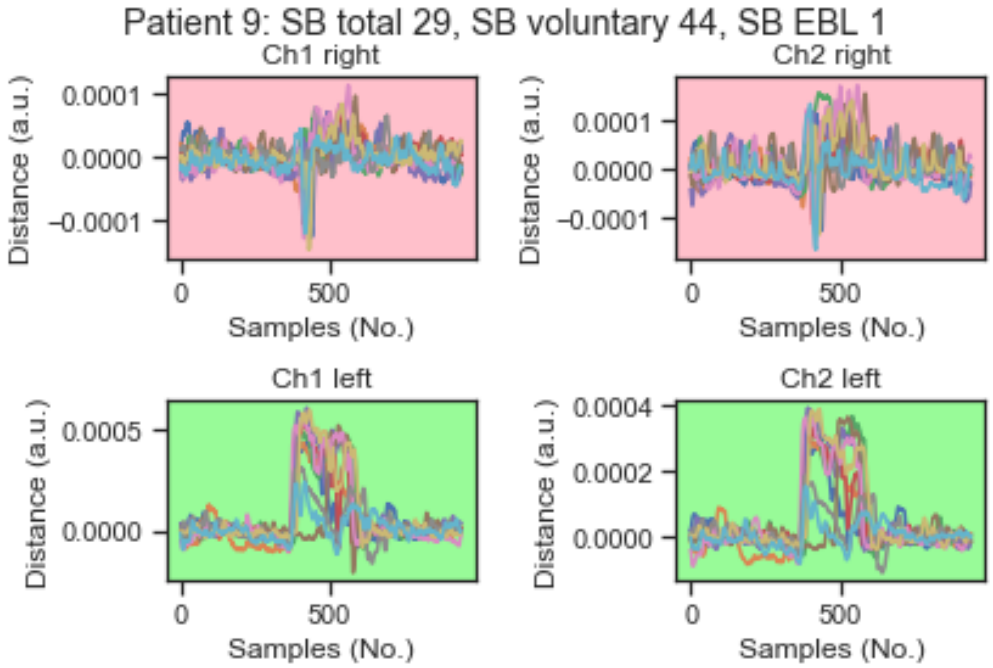

Figure H.25 *All eyebrow lift repetitions of patient number 9.*

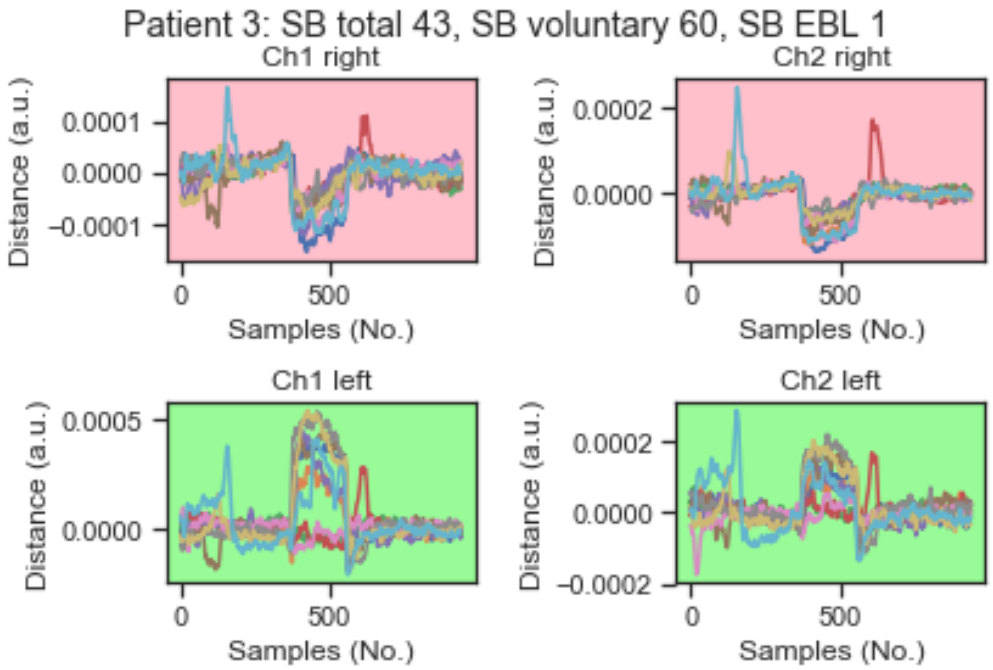

Figure H.26 *All eyebrow lift repetitions of patient number 3.*

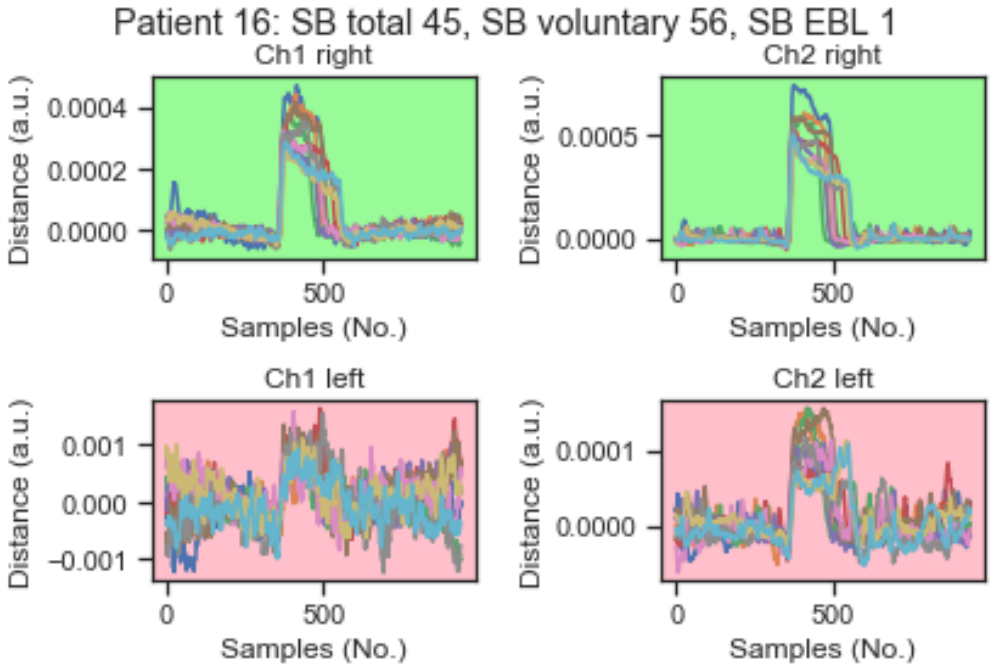

Figure H.27 *All eyebrow lift repetitions of patient number 16.*

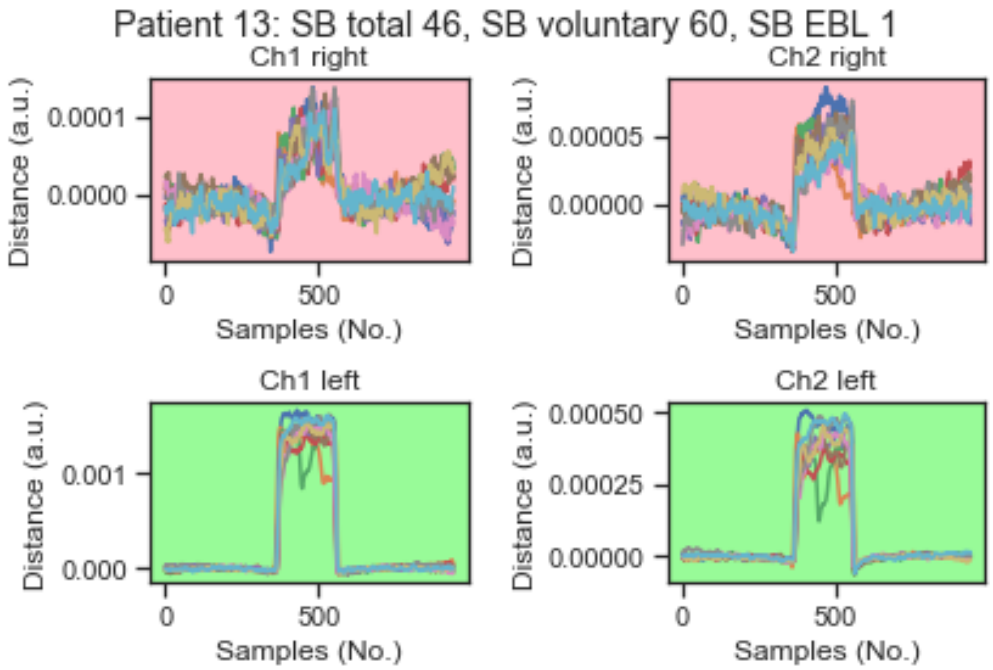

Figure H.28 *All eyebrow lift repetitions of patient number 13.*

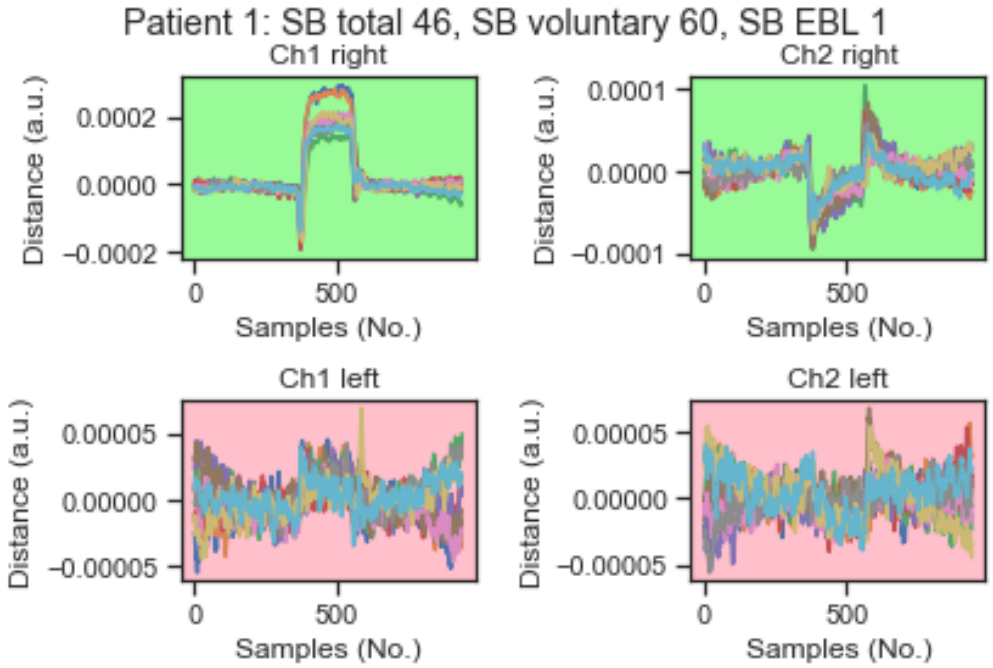

Figure H.29 *All eyebrow lift repetitions of patient number 1.*

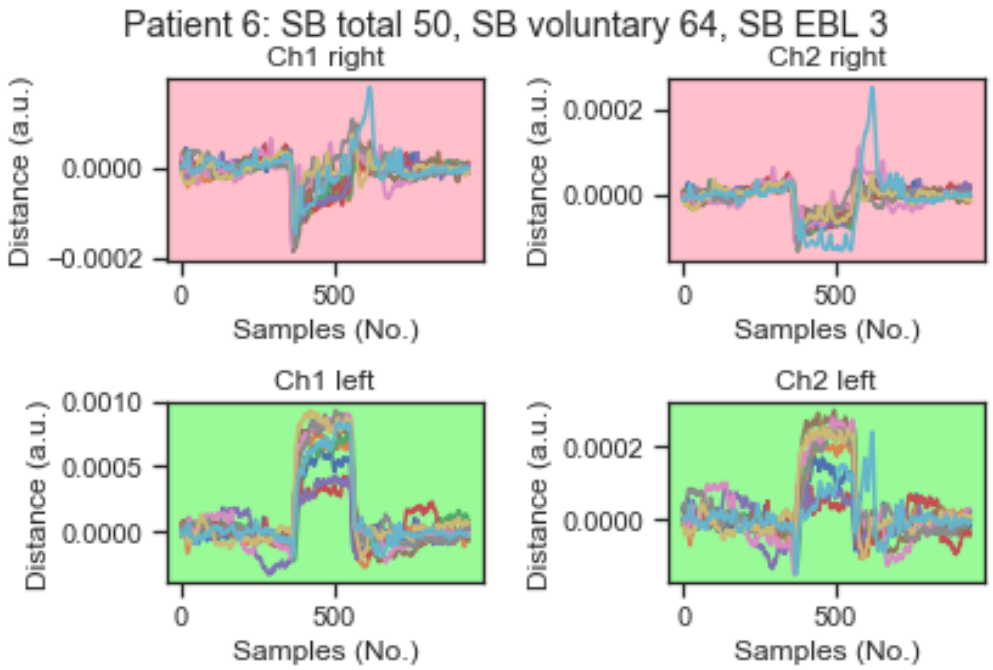

Figure H.30 *All eyebrow lift repetitions of patient number 6.*

Figure H.31 *All eyebrow lift repetitions of patient number 4.*

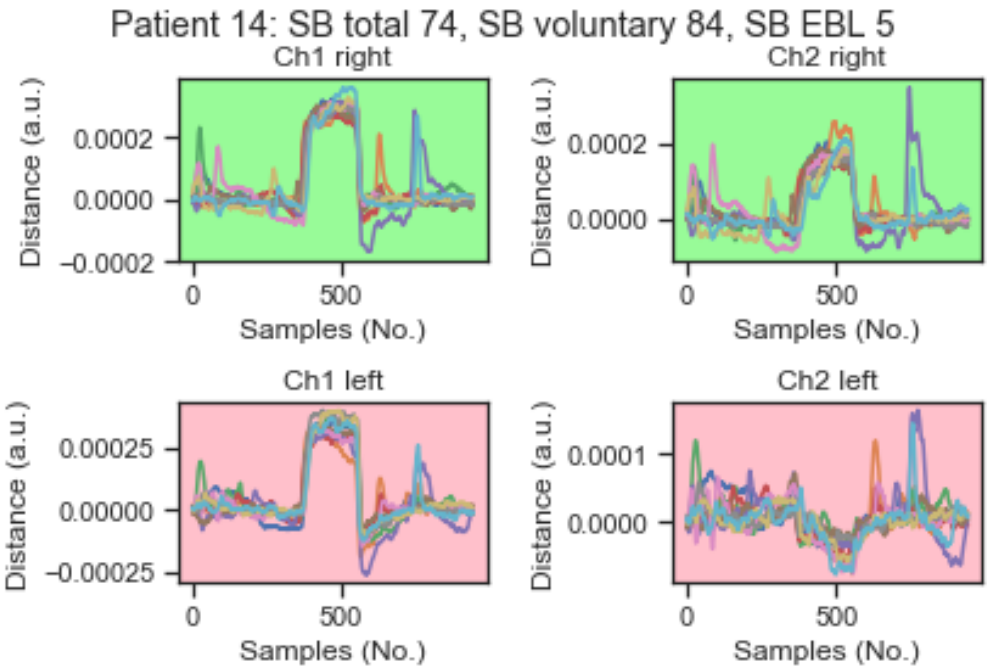

Figure H.32 *All eyebrow lift repetitions of patient number 14.*

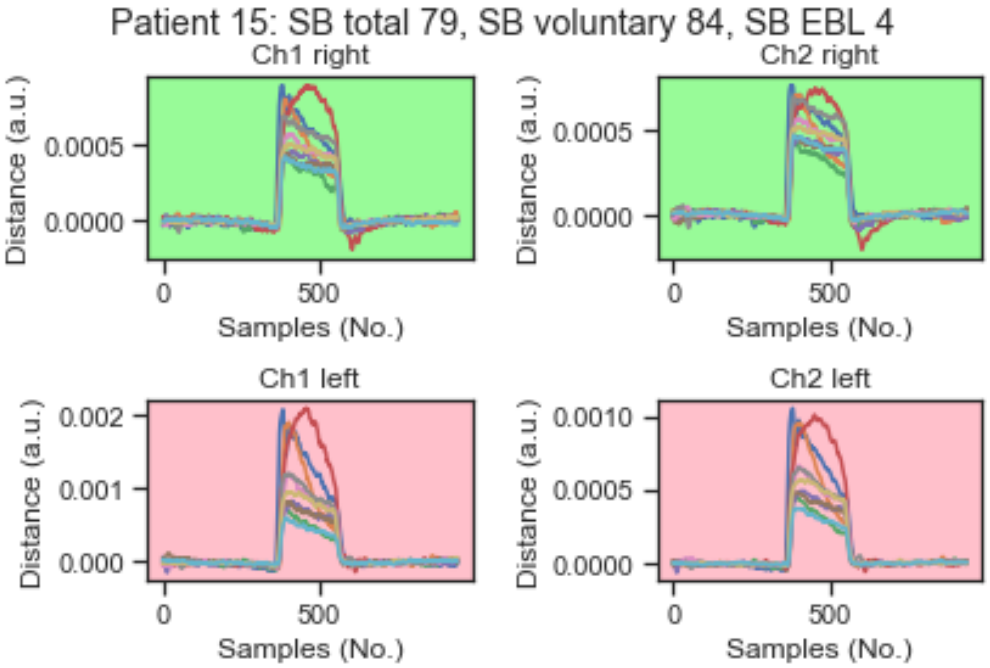

Figure H.33 *All eyebrow lift repetitions of patient number 15.*

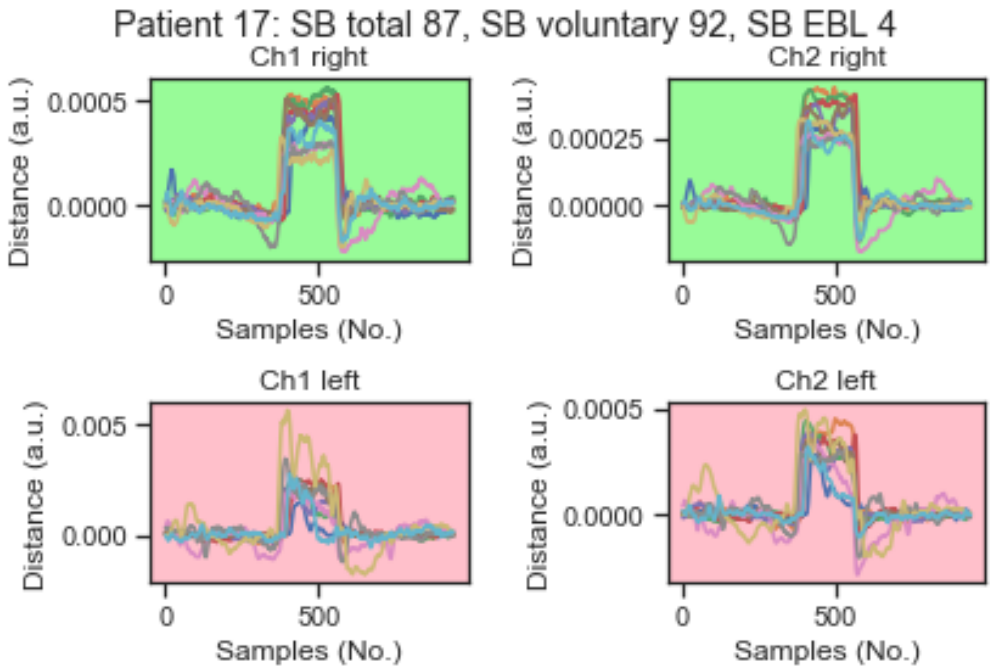

Figure H.34 *All eyebrow lift repetitions of patient number 17.*

The data shown in Figures H.18-H.34 is used for analysis of waveform and temporal delay.