6-3-2016

Data-Driven Synthesis and Evaluation of Syntactic Facial Expressions in American Sign Language Animation

Hernisa Kacorri
Graduate Center, City University of New York

Recommended Citation

Kacorri, Hernisa, "Data-Driven Synthesis and Evaluation of Syntactic Facial Expressions in American Sign Language Animation" (2016). *CUNY Academic Works.*
http://academicworks.cuny.edu/gc_etds/1375

DATA-DRIVEN SYNTHESIS AND EVALUATION OF SYNTACTIC FACIAL EXPRESSIONS IN AMERICAN SIGN LANGUAGE ANIMATION

by

HERNISA KACORRI

A dissertation submitted to the Graduate Faculty in Computer Science in partial fulfillment of the requirements for the degree of Doctor of Philosophy

The City University of New York

2016

This manuscript has been read and accepted for the Graduate Faculty in Computer Science in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Matt Huenerfauth

____________________________ ____

Date Chair of Examining Committee

Robert Haralick

____________________________ ____

Date Executive Officer

Vicki Hanson, Rochester Institute of Technology

Raquel Benbunan-Fich, Baruch College, CUNY

Andrew Rosenberg, Queens College, CUNY

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

DATA-DRIVEN SYNTHESIS AND EVALUATION OF SYNTACTIC FACIAL EXPRESSIONS IN AMERICAN SIGN LANGUAGE ANIMATION

by

HERNISA KACORRI

Advisor: Professor Matt Huenerfauth

Technology to automatically synthesize linguistically accurate and natural-looking animations of American Sign Language (ASL) would make it easier to add ASL content to websites and media, thereby increasing information accessibility for many people who are deaf and have low English literacy skills. State-of-art sign language animation tools focus mostly on accuracy of manual signs rather than on the facial expressions. We are investigating the synthesis of syntactic ASL facial expressions, which are grammatically required and essential to the meaning of sentences. In this thesis, we propose to: (1) explore the methodological aspects of evaluating sign language animations with facial expressions, and (2) examine data-driven modeling of facial expressions from multiple recordings of ASL signers.

In Part I of this thesis, we propose to conduct rigorous methodological research on how experiment design affects study outcomes when evaluating sign language animations with facial expressions. Our research questions involve: (i) stimuli design, (ii) effect of videos as upper baseline and for presenting comprehension questions, and (iii) eye-tracking as an alternative to recording question- responses from participants. In Part II of this thesis, we propose to use

generative models to automatically uncover the underlying trace of ASL syntactic facial expressions from multiple recordings of ASL signers, and apply these facial expressions to manual signs in novel animated sentences.

We hypothesize that an annotated sign language corpus, including both the manual and non-manual signs, can be used to model and generate linguistically meaningful facial expressions, if it is combined with facial feature extraction techniques, statistical machine learning, and an animation platform with detailed facial parameterization. To further improve sign language animation technology, we will assess the quality of the animation generated by our approach with ASL signers through the rigorous evaluation methodologies described in Part I.

Acknowledgments

I have been lucky to have Matt Huenerfauth as my advisor whose endless support, guidance, and understanding made this possible. He introduced me to the Deaf Culture, invited me to advance computer science for positive social impact, and taught me the value of interdisciplinary research. He is an exceptional scientist and inspires me to become one too. He is my role model.

I am grateful to my dissertation committee – Vicki Hanson, Raquel Benbunan-Fich, and Andrew Rosenberg – for generously devoting time and effort to review this dissertation and providing invaluable feedback. I am additionally grateful to Andrew for accepting me as a member of his Speech Lab, expanding my horizons into speech processing, and receiving valuable career advice.

As a member of Matt's Linguistic and Assistive Technologies Laboratory, both at CUNY and RIT, I experienced wonderful collaborations. I am grateful to my lab mates, Pengfei Lu and Allen Harper, for their expertise and knowledge. I enjoyed our conversations and benefited from the exchange of great ideas in our work together. Many thanks to all the people who helped with recruiting participants, preparing ASL animation stimuli, and collecting data: Wesley Clarke, Sarah Ebling, Dhananjai Hariharan, Jonathan Lamberton, Jennifer Marfino, Kellie Menzies, Sandra Sidler-Miserez, Fatimah Mohammed, Giovanni Moriarty, Miriam Morrow, Kasmira Patel, Evans Seraphin, Christine Singh, Mackenzie Willard, and Fang Zhou Yang.

Part of my work was also made possible through a joint effort between our lab and labs at Boston University, Rutgers University, and the German Research Centre for Artificial Intelligence (DFKI). Specifically, I would like to thank Carol Neidle, Dimitris Metaxas, and

Alexis Heloir for providing data and software, as well as lending their expertise in ASL linguistics, computer vision, and computer animation.

I am fortunate to have received funding from the National Science Foundation (award 1065009) and The Graduate Center, CUNY (Science Fellowship, Doctoral Student Research Grants, Research Presentation Support, and Mina Rees Dissertation Fellowship in the Sciences).

I would also like to acknowledge the support of early mentors and friends who shaped my path to the Graduate Center and this dissertation. I am indebted to Georgios Kouroupetroglou, my undergraduate and masters advisor, who gave me the guidance and opportunity to conduct research. I am grateful to old friends who helped and encouraged me in my efforts to pursue a graduate degree: Harald Kllapi, Anna Nika, Dimitis Paparas, Vivi Riga, and Eva Sitaridi.

I am also grateful to new friends who made New York feel like a second home: Antonis Achilleos, Thomas Flynn, Effie Korkou, Agis Mesologitis, Valia Mitsou, Antonios Papaioannou, and Jordan Salvit.

More than anything, I am the product of the love and support of my wonderful family. My parents, Doloreza and Spiro Kacorri, taught me the importance of education and sacrificed their dreams to realize mine. My brothers, Armando and Nikolin, always believed in me and inspired me to keep going. My partner, Ali Raza Syed, gave me New York, patience, love, and hope for a wonderful future to come.

Contents

Abstract .... **iv**

Acknowledgments .... **vi**

Contents .... **viii**

List of Tables .... **xv**

List of Illustrations .... **xvii**

Chapter 1 Introduction **1**

1.1 Focus of This Thesis 3

1.2 Overview of This Thesis 4

Chapter 2 Background on Sign Language Facial Expressions and Animation Generation6

2.1 Syntactic Facial Expressions 6

2.2 Non Syntactic Facial Expressions 9

2.2.1 Lexical *9*

2.2.2 Modifier or Adverbial *10*

2.2.3 Paralinguistic *11*

2.3 Animation Technologies in Sign Language 11

Part I **14**

Evaluation of Facial Expressions in ASL Animations **14**

Prologue to Part I **15**

Chapter 3 Related Work on Evaluation of Sign Language Animations **18**

3.1 Engineering Stimuli and Comprehension Questions 19

3.2 Modality of Upper Baselines Used for Comparison 21

3.2.1 No Upper Baseline *22*

3.2.2 Video Upper Baseline *22*

3.2.3 Animation Upper Baseline *23*

3.3 Modality of Instructions and Evaluation Questions 25

3.4 User Studies with Deaf Participants and Eye-tracking 26

3.4.1 Eye-Tracking Participants who are Deaf *28*

3.4.2 Eye-Tracking with Sign-Language Video *30*

3.5 Reported Participant's Demographics and Technology Experience 32

3.5.1 Demographics in Prior Studies *33*

3.5.2 Acceptance of Multiple Signing Avatars *36*

Chapter 4 Stimulus and Comprehension Question Design **40**

4.1 Pilot Testing of Experimental Stimuli 42

4.1.1 English-to-ASL versus ASL-Originated Stimuli Design *42*

4.1.2 Results and Comparison of the Studies *47*

4.2 Release of Experimental Stimuli and Questions 50

4.2.1 Overview of the Collection *51*

4.2.2 Design of Stimuli and Comprehension Questions *53*

4.2.3 Likert-scale Questions *57*

4.2.4 Availability of the Collection 58

Chapter 5 Effect of Videos as Upper Baseline and Questions 59

5.1 Research Methodology and Hypotheses 62

5.1.1 Three Phases of This Research 65

5.2 Phase 1: Animation vs. Video Upper Baselines 67

5.2.1 Design of Evaluation Studies for Phase 1 67

5.2.2 Recording the Human Video Upper Baseline for Phase 1 70

5.2.3 Data Collection and Results for Phase 1 73

5.3 Phase 2: Animation vs. Video Baselines with Facial Expression 81

5.3.1 Design of Evaluation Studies for Phase 2 82

5.3.2 Creation of Human Video Upper Baseline for Phase 2 86

5.3.3 Data Collection and Results of Phase 2 88

5.4 Phase 3: Animation vs. Video Comprehension Questions 93

5.4.1 Design of Evaluation Studies for Phase 3 93

5.4.2 Results of Phase 3 95

5.5 Conclusions and Recommendations for Future Researchers 99

5.5.1 Recommendations for Future Researchers 102

Chapter 6 Evaluation of Facial Expressions with Eye-Tracking 109

6.1 Eye-gaze Metrics and Hypotheses 111

6.2 User Study 114

6.3 Initial Analysis and Results 117

6.3.1 Areas of Interest in the Stimuli 118

6.3.2 Results and Comparisons 119

6.4 Extended Analysis and Results 122

6.4.1 Areas of Interest in Stimuli 123

6.4.2 Results and Comparisons 124

6.5 Conclusions and Recommendations for Future Researchers 127

Chapter 7 Demographics and Other Factors Influencing Acceptance of ASL Animation 132

7.1 Collecting Independent Variables 133

7.1.1 Demographic Questions 134

7.1.2 Technology Experience and Attitudes 135

7.2 Collecting Dependent Variables 137

7.2.1 Recruiting and Data Collection 141

7.2.2 Analysis and Results 142

7.3 Discussion of the Relative Importance of the Factors 146

7.4 Conclusions and Recommendations for Future Researchers 150

Epilogue to Part I 152

Part II: Data-driven Models for Syntactic Facial Expression Synthesis 155

Prologue to Part II 156

Chapter 8 Background on Synthesis of Facial Expression Animations 158

8.1 Facial Parameterization 159

8.1.1 Facial Action Coding System (FACS) 160

8.1.2 MPEG-4 Facial Animation (MPEG-4 FAPS) *160*

8.1.3 Proprietary Facial Parameterization *162*

8.2 Facial Feature Extraction 162

Chapter 9 Literature Survey of Facial Expression in Sign Language Animation 164

9.1 HamNoSys-based 167

9.1.1 Approach *167*

9.1.2 Data Resources *170*

9.1.3 Evaluation *171*

9.2 VCOM3D 172

9.2.1 Approach *172*

9.2.2 Data Resources *174*

9.2.3 Evaluation *175*

9.3 DePaul 175

9.3.1 Approach *176*

9.3.2 Data Resources *177*

9.3.3 Evaluation *178*

9.4 SignCom 180

9.4.1 Approach *180*

9.4.2 Data Resources *182*

9.4.3 Evaluation *182*

9.5 ClustLexical 183

9.5.1 Approach *184*

9.5.2 Data Resources ... 185

9.5.3 Evaluation ... 187

9.6 Conclusions ... 188

Chapter 10 Extracted Facial-Feature Data for ASL Animation Synthesis ... 191

10.1 Extracted MPEG4 Head/Face Features ... 191

10.2 Initial Pilot Data ... 193

10.3 Linguistically Diverse Data ... 196

Chapter 11 Virtual Human Platform ... 201

11.1 Animation Platform Extended with ASL Handshapes ... 201

11.2 Animation Platform Extended with MPEG4 Facial Controls ... 203

11.2.1 Supported MPEG4 Parameters ... 205

11.2.2 MPEG4 Features-to-Animation Pipeline ... 207

11.3 Evaluation of the Extended Animation Platform ... 209

11.4 Comparison of the Extended EMBR Animation Platform to Our Prior ASL Animation Platform ... 213

11.4.1 Comparison of New versus Old Avatar Platform ... 214

11.4.2 Experiment Setup and Results ... 216

Chapter 12 ASL Facial Expression Synthesis with Continuous Profile Models ... 222

12.1 Continuous Profile Model ... 223

12.2 Obtaining the CPM Latent Trace on ASL Facial Expressions ... 226

12.3 Comparison of CPM Latent Trace and Training Data with DTW ... 230

12.3.1 Dynamic Time Warping (DTW) ... 231

12.3.2 DTW Distance of the Latent Trace from the Training Data 232

Chapter 13 Evaluation of the Data-Driven ASL Facial Expression Synthesis with Continuous Profile Models 236

13.1 Metric Evaluation 237

13.1.1 Results 239

13.2 User Evaluation 239

13.2.1 Animation Stimuli 242

13.2.2 Results 243

Epilogue to Part II 247

Chapter 14 Conclusions and Contributions 249

14.1 Contributions 253

14.2 Future Work 256

14.3 Summary 258

APPENDIX A ASL Facial Expression Stimuli Collection 260

APPENDIX B Best Practices For Conducting Evaluations of Sign Language Animations 269

APPENDIX C CPM Training Set Per ASL Facial Expression Subcategory 275

Bibliography 282

List of Tables

Table 1: Similarity of the upper baseline animation to the animation being evaluated. 25

Table 2: Demographic and Technology Experience Characteristics Reported in Example User Studies. 35

Table 3: Demographic Profile of Participants in Prior Studies 35

Table 4: Collection Overview. 53

Table 5: Summary of Three Phases of Experiment...65

Table 6: Correlations between Eye Metrics and Responses 122

Table 7: Multiple Regression Model – Comprehension 143

Table 8: Multiple Regression Model – Subjective 145

Table 9: Comparison of state-of-the art approaches on sign language animation with facial expressions..165

Table 10: Grouping of syntactic ASL facial expressions into subcategories...........................198

Table 11: Supported MPEG4 controls in the enhanced EMBR platform. 206

Table 12: Training data and the obtained latent traces for each of the CPM models on ASL facial expression subcategories. The length of time series N corresponds to the duration in video frames of the longest example in the data set. Both the obtained latent trace and each time series in the training set has 14 dimensions corresponding to the extracted features from Visage: 'Head x', 'Head y', 'Head z', 'Head pitch', 'Head yaw', 'Head roll', 'raise_l_i_eyebrow', 'raise_r_i_eyebrow', 'raise_l_m_eyebrow', 'raise_r_m_eyebrow', 'raise_l_o_eyebrow', 'raise_r_o_eyebrow', ‘squeeze_l_eyebrow’, ‘squeeze_r_eyebrow’. The latent trace has a time axis of length M, which is approximately double the temporal resolution of the original training examples. 226

Table 13: CPM latent trace, centroid, and gold standard for each ASL facial expression subcategory. 237

Table 14: ASL passages in our stimuli collection including codenames, ASL glosses, and English translation of two possible interpretations based on the facial expression being performed. 260

Table 15: Comprehension Questions and Their Suggested Answers based on the stimuli meaning when the ASL facial expression is correctly perceived. (In this table the questions are phrased in English.) 264

Table 16: Video recordings (9) in the CPM training set for YesNo_A. 275

Table 17: Video recordings (10) in the CPM training set for YesNo_B. 276

Table 18: Video recordings (4) in the CPM training set for WhQuestion_A. 276

Table 19: Video recordings (8) in the CPM training set for WhQuestion_B. 276

Table 20: Video recordings (2) in the CPM training set for Rhetorical_A. 277

Table 21: Video recordings (8) in the CPM training set for Rhetorical_B. 277

Table 22: Video recordings (29) in the CPM training set for Topic_A. 278

Table 23: Video recordings (15) in the CPM training set for Topic_B. 279

Table 24: Video recordings (16) in the CPM training set for Negative_A. 280

Table 25: Video recordings (25) in the CPM training set for Negative_B. 281

List of Illustrations

Figure 1: Still images taken from videos included in the stimuli collection described in this thesis, with each image illustrating a moment when a particular facial expressions is occurring: (a) Y/N-Question, (b) Negative, (c) WH-Question, (d) Topic, and (e) RH-Question. 8

Figure 2: Differentiating (a) an ASL statement from: (b) an ASL question, (c) an ASL question with topic the doctor, (d) an ASL question with Negative based on the associated facial expression. 9

Figure 3: Example of ASL facial expression in lexical level (a) NOT-YET and (b) LATE signs. (source: www.handspeak.com)........10

Figure 4: Example of adverbial ASL facial expressions: (a) 'OO' the book has a small number of pages, (b) 'MM' the book has an average number of pages, and (c) 'CHA' the book is thick in pages. (source: www.handspeak.com) 10

Figure 5: Still images with each image illustrating a moment when a particular facial expressions is occurring: (a) angry, (b) sad, and (c) ironic. 11

Figure 6: Example of ASL animations produced by (a) a 3D animation artist using motion capture data over a few months (source: Hurdich, 2008) and (b) an ASL signer using a "scripting" software in a few minutes. 12

Figure 7: Upper baseline categories adopted by researchers when evaluating animations. 22

Figure 8: ASL character and some of the available facial expressions........43

Figure 9: English-to-ASL Stimuli Set examples: (1) yes/no-question and (2) emphasis. The name

"Charlie" refers to the animated character or human signer performing the passage. 45

Figure 10: Five categories of facial expression in the English-to-ASL Stimuli Set. 45

Figure 11: Example of an y/n-question (1) and a topic (2) stimulus in the ASL-originated set. 47

Figure 12: Six categories of facial expression in the ASL-originated Stimuli Set. 47

Figure 13: Comprehension question scores for both types of stimuli, showing results for animations with- and without-facial-expressions, with results for emotion and non-emotion categories. 49

Figure 14: Still images taken from videos included in the stimuli collection, with each image illustrating a moment when a particular facial expressions is occurring: (a) YN-Question, (b) WH-Question, (c) RH-Question, (d) Topic, (e) Negative, and (f) Emotional Affect (an example of anger). 52

Figure 15: Example script for a story shown in the study...68

Figure 16: Screenshots from the side-by-side comparison, as seen by participants in (a) 2010 or (b) 2012..69

Figure 17: Diagram of an overhead view of recording studio. 73

Figure 18: Results of comprehension scores in Phase 1. 78

Figure 19: Results of grammaticality, understandability, and naturalness scores in Phase 1. 79

Figure 21: Example script (originally shown in Fig. 12, repeated here for convenience) and corresponding comprehension questions for a story shown in the study. 84

Figure 22: Screenshots of the side-by-side comparison portion of the studies as shown to participants in (a) animator-upper-baseline study and (b) video-upper-baseline study.85

Figure 23: Results of comprehension scores in Phase 2. 90

Figure 24: Results of grammaticality, understandability, and naturalness scores in Phase 2. 91

Figure 25: Results of Side-by-side comparison scores in Phase 2. 92

Figure 26: Screenshots of two forms of comprehension questions presented in 2010 study (a) and 2013 study (b). 95

Figure 27: Results of comprehension scores in Phase 3. 97

Figure 28: Results of grammaticality, understandability, and naturalness scores in Phase 3. 98

Fig. 29: Screenshots from the three types of stimuli: i) video of human signer, ii) animation with facial expressions, and iii) animation without facial expressions............................115

Fig. 30: An example of a stimulus used in the study: story, subjective and notice questions, and comprehension questions...116

Figure 31: Screen regions for the face and hands AOIs of the animated character and the human signer. 119

Figure 32: Proportional Fixation Time on the Face 120

Figure 33: Transition Frequency Between the Face and Hands..121

Figure 34: Screen regions for the upper face, lower face, and hands AOIs. 124

Figure 35: Transitions per second between: (a) the hands-body AOI and the upper-face AOI ("TransUFH") and (b) the hands-body AOI and the lower-face AOI ("TransLFH").125

Figure 36: Proportional fixation time on: (a) the upper-face AOI (labeled as "UFacePFT") and (b) the lower-face AOI (labeled as "LFacePFT")...126

Figure 37: Correlations between participants responses (rows) and eye metrics (columns), including proportional fixation time and transition frequency for upper-face and lower-face...127

Figure 38: Fixation trail length for each type of stimulus (a) and correlations to responses (b).127

Figure 39: Screenshots from the three avatars shown in the study: (a) EMBR, (b) JASigning, (c) VCom3D. 139

Figure 40: Regression model comparison summary. (Significance codes: 0 '***' 0.001 '**' 0.01) 146

Figure 41: Relative importance (normalized to sum to 100%) of factors in Comprehension Model 2 and in Subjective Model 2, with 95% bootstrap confidence intervals. 147

Figure 42: A FACS action unit example: (a) AU12 Lip Corner Puller (source: ISHIZAKI Lab) mapped to (b) zygomaticus major facial muscle (source: Wikipedia). 160

Figure 43: MPEG-4 facial animation (a) feature points and (b) scale factors. 161

Figure 45: An example of the HAMBURG sign transcribed in HamNoSys. (source: Hanke, 2010) 168

Figure 46: Face movement options in JASigning organized by alphanumeric tags for: (a) eyebrows and nose, and (b) mouth gestures. (source: San-Segundo Hernández, 2010) 169

Figure 48: VCom3D (a) deformable mesh model of the avatar's head (source: DeWitt et al., 2003), (b) a 'one key-frame pose' topic facial expression, and (c) a 'few key-frames in a looping mode' negative facial expression..173

Figure 49: Timing diagram example that shows how a '3 key-frame' facial expression is applied to the hand movements (interpolation, looping, transitions, and holds). (source: DeWitt et al., 2003) 174

Figure 50: The ASL sentence "How many books do you want?" (source: Wolfe et al., 2011)177

Figure 51: Eyebrow height on wh-questions, angry, and quizzical expressions. (source: Grossman and Kegl, 2006) 178

Figure 52: Co-occurrence of wh-question with emotions (source: Schnepp et al., 2012). 179

Figure 53: (a) Motion-capture sensors on a signer's face and (b) blended faces of the avatar driven by the values of the facial markers. (source: Gibet et al., 2011) 181

Figure 54: An example of ID-glosses alignment to the spoken language words and variants extraction. (source: slides of Schmidt et al., 2013)........185

Figure 55: High-level feature extraction with Active Appearance Models. (source: Schmidt et al., 2013) 186

Figure 56: Fitted 3D mask in Visage. (Source: visagetechnologies.com) 192

Figure 57: ASL signer's a) fitted face shape mask and b) tracking screenshot in Visage software. 194

Figure 58: Extracted features and gloss annotation for an ASL passage in the pilot data that includes facial expressions such as topic and negative. The features displayed in each graph include: (a) pitch, yaw, roll, (b) head x, y, z, (c) raise_l_i_eyebrow, raise_r_i_eyebrow, raise_l_m_eyebrow, raise_r_m_eyebrow, raise_l_o_eyebrow, raise_r_o_eyebrow, (d) squeeze_l_eyebrow, squeeze_r_eyebrow, and (e) close_t_l_eyelid, close_t_r_eyelid, close_b_l_eyelid, close_b_r_eyelid. 196

Figure 59: Neutral facial poses to be used for the signer's profile in Visage. 197

Figure 60: Extracted features and gloss annotation for an ASL passage in the linguistically diverse data that includes facial expressions such as topic and negative. The features displayed in each graph include: (a) pitch, yaw, roll, (b) head x, y, z, (c) raise_l_i_eyebrow, raise_r_i_eyebrow, raise_l_m_eyebrow, raise_r_m_eyebrow, raise_l_o_eyebrow, raise_r_o_eyebrow, (d) squeeze_l_eyebrow, squeeze_r_eyebrow, and (e) close_t_l_eyelid, close_t_r_eyelid, close_b_l_eyelid, close_b_r_eyelid. 200

Figure 61: Naming convention for the new ASL handshapes added to EMBR. (Source: ASLLVD) 202

Figure 62: A skeleton view of the avatar in Blender performing the ASL handshape “W”. 203

Figure 63: (a) Some feature points and (b) FAPU scale factors in MPEG-4 standard. 204

Figure 64: Wireframe Max in (a) the initial EMBR platform and (b) in the enhanced platform (MPEG4 feature points are illustrated with black crosses). 205

Fig 65: (a) forehead with eyebrows raised before the addition of MPEG-4 controls, facial mesh with wrinkling, and lighting enhancements, (b) eyebrows raised in our current system. 207

Figure 66: Screenshots from a human-recording-driven and neutral Yes/No-question stimulus in the study. 211

Figure 67: Notice and Comprehension scores for animations with Driven and Neutral facial expressions..212

Figure 68: This graphic depicts the same timeline of an ASL sentence consisting of four signs (shown in the “Glosses” row) with co-occurring facial expressions in (a) VCOM3D and (b) extended EMBR. 214

Figure 69: Notice Scores for OLD and NEW Animation Platform. 219

Figure 70: Comprehension Scores for OLD and NEW Animation Platform. 220

Figure 72: Graphical model depiction of a CPM for a particular series xk. Top nodes, are the hidden state variables underlying each observation, xik. The table illustrates the state-space: time-state/scale-state pairs mapped to the hidden variables, where time states belong to the integer set (1… M) and scale states belong to an ordered set with 7 evenly spaced scales in logarithmic space (as in prior CPM experiments in Listgarten et al., 2004). 224

Figure 73: Rhetorical_B (a) training examples before CPM (lines represent values of the 14 face features over time), (b) training examples after CPM (time alignment and rescaling), and (c) the obtained latent trace (based upon all eight examples). Note that the y-axis limits differ across the subplots. 229

Figure 74: Rhetorical_B per head/face-feature time series in the a) training examples before CPM, and b) training examples after CPM, and c) obtained latent trace. 230

Figure 75: *Example of DTW alignment between the "raise_l_i_eyebrow" values detected in human recordings of two ASL stories containing a Negative facial expression. The duration of the facial expression in Story 1 and Story 2 is 1414 and 924 frames, respectively and their calculated normalized distance was found to be 8.76.* 232

Figure 76: Visualization of DTW distances and the centroid in the comparison set for each of the ASL facial expression subcategories. The centroid is represented as the node with the smallest sum distance; hence the lightest color in the graph. For all graphs, the latent trace was found to be the centroid corresponding to the following node numbers: YesNo_A – 9, YesNo_B – 10, WhQuestion_A – 4, WhQuestion_B – 8, Rhetorical_A – 2, Rhetorical_B – 8, Topic_A – 29, Topic_B – 15, Negative_A – 16, and Negative_B – 25. 235

Figure 77: DTW distances on the FAP36 feature, squeeze_l_eyebrow, during a Negative_A ASL facial expression: (left) between the CPM latent trace and the gold standard and (right) between the centroid and the gold standard. The timeline for both graphs is given in milliseconds. 238

Figure 78: Grouped barplots of overall normalized DTW distances for latent trace and centroid (left) and per each subcategory of ASL facial expression (right). 239

Figure 79: Screenshots of YesNo_A stimuli of three types: a) neutral, b) centroid, and c) latent trace. 240

Figure 80: Before and after pose sampling of the FAP37 feature, squeeze_r_eyebrow, during a Negative_A ASL facial expression driven by the latent trace. 243

Figure 81: Comprehension scores for animations with facial expressions driven by centroid, latent trace, and with a neutral head/face pose. 245

Figure 82: Subjective scores for animations with facial expressions driven by centroid, latent trace, and with a neutral head/face pose. 246

Figure 83: Screenshots of (a) the comprehension questions and (b) the subjective/“notice” questions HTML forms. 268

Chapter 1 Introduction

Deaf[1] adults who use sign language as a primary means of communication may have low literacy skills in written languages due to limited spoken language exposure and other educational factors. For example, in the U.S., standardized testing has revealed that a majority of deaf high school graduates (students aged 18 and older) perform at or below fourth-grade English reading level (typically age 10) (Traxler, 2000). If the reading level of the text on websites, television captioning, or other media is too complex, these adults may not comprehend the conveyed message despite having read the text. While more than 500,000 people in the U.S. use American Sign Language (ASL) as a primary means of communication (Mitchell et al., 2006), the number of people using sign language as their first language worldwide rises to 70 million (World Federation of the Deaf, 2014). Fluency in sign language does not entail fluency in the written language; sign languages are distinct natural languages with a different word order, syntax, and lexicon from spoken and written languages. Therefore, technology to automatically synthesize grammatically correct and natural looking sign language animations from written text can benefit this population.[2]

While videos of human signers are often incorporated in media for presenting information to deaf users, there are several reasons to prefer animated sign language avatars. It is often prohibitively expensive to re-film a human performing sign language for information

[1] This dissertation adopts the following conventions: *Deaf* (capitalized) describing members of the linguistic community of sign language users, and *deaf* (un-capitalized) describing the audiological state of a hearing loss. When referring to prior publications (e.g. in related work sections) we maintain the notation originally used by the author(s). The questionnaires used in our user studies include the terms deaf/Deaf, hard-of-hearing, and hearing for self-identification of the participants.

[2] The target users, of the technology developed in this dissertation, both: (1) use ASL and (2) have lower English literacy skills.

that is frequently updated, thus leading to out-of-date information. Automatically synthesized animations allow for frequent updating as well as presenting content that is dynamically generated from a database query. Assembling video clips of individual signs together into sentences does not produce high-quality results. Animation offers additional flexibility through customized signing speed/style, avatar appearance, and view perspective. It preserves author anonymity through the use of avatars and enables the collaboration between multiple authors for scripting a message in sign language. For these reasons, many accessibility researchers favor sign language animations for presenting information to deaf users. For example, Adamo-Villani and Wilbur (2010) investigated digital lessons annotated with ASL animations to improve the mathematical abilities of deaf pupils, and many researchers used signing avatars for the output of their speech/text-to-signing translation system (e.g. Krnoul et al., 2008; San-Segundo et al., 2012). However, it is still challenging for modern sign language animation software to support accurate and understandable signing via virtual human characters.

State of art sign language animation tools focus mostly on accuracy of manual signs rather than facial expressions (Elliott et al., 2008; Filhol et al., 2010; Fotinea et al., 2008). However, the production of grammatical facial expressions and head movements coordinated with specific manual signs is crucial for the interpretation of signed sentences. For example, in ASL there is a significant difference in deaf users' comprehension of animations when linguistically and emotionally meaningful facial expressions are supported (Huenerfauth et al., 2011).

1.1 Focus of This Thesis

Our research is focused on data-driven modeling, generation, and evaluation of facial expressions in ASL animations. Facial expressions are an essential part of the fluent performance of ASL. They can convey emotional information, subtle variations in the meaning of words, and other information, but this thesis focuses on a specific use of facial expressions: to convey grammatical information during entire syntactic phrases in an ASL sentence.

To produce an animation with good facial expressions an animation artist could carefully edit the facial mesh of an animated character, but this is very time-consuming. We want to support automatic synthesis of sign language and scripting of sign language animations. We are studying how to model and generate ASL animations that include facial expressions to convey grammatical syntax information, such as negative; topic; and yes/no, wh-word, and rhetorical questions. Our objective is to determine when signers use these facial expressions, how they perform each, how the timing of these facial expressions occurs in relation to the manual signs, and how the sequential occurrence of facial expressions affect one another. Thus, we are investigating technologies for automatically planning aspects and timing of face movement. In addition to planning algorithms, we also need a succinct representation of ASL (that can encode a good-quality performance with as few parameters as possible). This makes it practical for a generation system to plan the animation, and it makes it possible for a human using a scripting tool to produce an animation with facial expressions in an efficient manner. We must test both our planning algorithms and our ASL script representation to ensure that they encode sufficient detail for ASL facial expressions that are understandable (and deemed natural) by signers.

While user studies are necessary to advance research in the field of sign language animation, the evaluation of synthesized facial expressions is still challenging due to the subtle and complex manner in which facial expressions affect the meaning of sentences (Huenerfauth et al., 2011). Evaluation methodologies and stimuli-design approaches that address these challenges would be of value for animation researchers improving generation of facial expressions in sign language.

Driven by the above limitations in the field, the research focus of this thesis is dual. First (in Part I) it investigates methodological aspects when evaluating sign language animations with facial expressions, and second (in Part II), it examines data-driven modeling of ASL facial expressions from multiple recordings of a human signer.

1.2 Overview of This Thesis

To provide the reader with essential background knowledge, Chapter 2 will provide a basic introduction to sign language animation synthesis, discuss the different types of facial expressions in sign language, and explain their importance for sign language animation with a primary focus on syntactic ASL facial expressions.

In Part I, we will begin by discussing related work on evaluation of sign language animations in Chapter 3. In the subsequent chapters, we will describe rigorous methodological research on how experiment design affects study outcomes when evaluating sign language animations with facial expressions. Specifically, Chapter 4 will focus on stimuli design, Chapter 5 will discuss the effect of videos as upper baseline and for presenting comprehension questions, Chapter 6 will investigate eye-tracking as an alternative to recording question-responses from participants, and last Chapter 7 will examine the use of participants' demographics and technology-experiences as predictors of their evaluation scores to the animation quality.

Part II of this thesis begins with Chapter 8, which will briefly discuss the key aspects of synthesis of facial expressions, such as parameterization and feature extraction. In Chapter 9 we will survey and critique the literature on state-of-art facial expression synthesis for sign language animations. The two key components necessary for data-driven facial expression animations, data and animation platform, will be discussed in detail in Chapter 10 and Chapter 11, respectively. Chapter 12 will then discuss how we can build a data-driven model for each of the supported facial expressions by using multiple recordings of ASL signers followed by their evaluation in Chapter 13.

Finally, this thesis will conclude in Chapter 14 with an outline of the research activities, expected contributions, and future work.

Chapter 2 Background on Sign Language Facial Expressions and Animation Generation

This chapter provides an overview of facial expression categories in sign language and briefly discusses animation technologies for sign language. We can group the non-manual behaviors (facial expressions and head movements) of the signer into four classes based on their linguistic level (Reilly and Anderson, 2002); lexical, adverbial, syntactic, and paralinguistic. For each of the facial expressions types, we provide the reader with an example from American Sign Language, to demonstrate their dynamics and their temporal coordination to the manual signs. For synthesizing an animation of ASL, the time synchronization between facial expressions and the manual signs in a sentence is a key challenge.

2.1 Syntactic Facial Expressions

Syntactic facial expressions convey grammatical information during entire syntactic phrases in a sign language sentence and are thus constrained by the timing and scope of the manual signs in a phrase (Baker-Shenk, 1983). A sequence of signs performed on the hands can have different meanings, depending on the syntactic facial expression that co-occurs with a portion of the sentence. For instance, a declarative sentence (ASL: “JOHN LIKE PIZZA” / English: “John likes pizza.”) can be turned into a yes/no question (English: “Does John like pizza?”), with the addition of a Yes/No-Question facial expression during the sentence. Similarly, the addition of a Negative facial expression during the verb phrase “LIKE PIZZA” can change the meaning of the sentence to “John doesn’t like pizza.” Optionally, the signer may include the word NOT during the sentence. For interrogative questions (typically with a “WH” word in English such as what, who, where, when, how, which, etc.), it is necessary for there to be a WH-word Question facial expression during the ASL sentence, e.g., “JOHN LIKE WHO.” Still images of an ASL signer performing these three ASL facial expressions, Yes/No-Question, Negative, and WH-Question are illustrated in Figure 1a, 1b, and 1c, respectively. While we use the term “facial expressions,” these phenomena also include movements of the head, which we also propose to model in this thesis. There are five types of ASL facial expressions (illustrated in Fig. 1) that we are focusing on this thesis. ASL linguistics references contain more detail about each, e.g., (Neidle et al., 2000), but they are described briefly below:

Yes/No-Questions: The signer raises his eyebrows while tilting the head forward during a sentence to indicate that it should be interpreted as a question.

Negative: The signer shakes his head left and right during the verb phrase, which should be interpreted with a negated meaning, often with the sign NOT.

WH-word Questions: The signer furrows his eyebrows and tilts his head forward during a sentence that should be interpreted as information-seeking, typically with a “WH” word such as what, who, where, when, how, which, etc.

Topic: The signer raises his eyebrows and tilts his head backward during a phrase at the beginning of a phrase that should be interpreted as a topic.

RH-Questions: The signer raises his eyebrows and tilts his head backward and to the side to indicate a question that should be interpreted rhetorically.

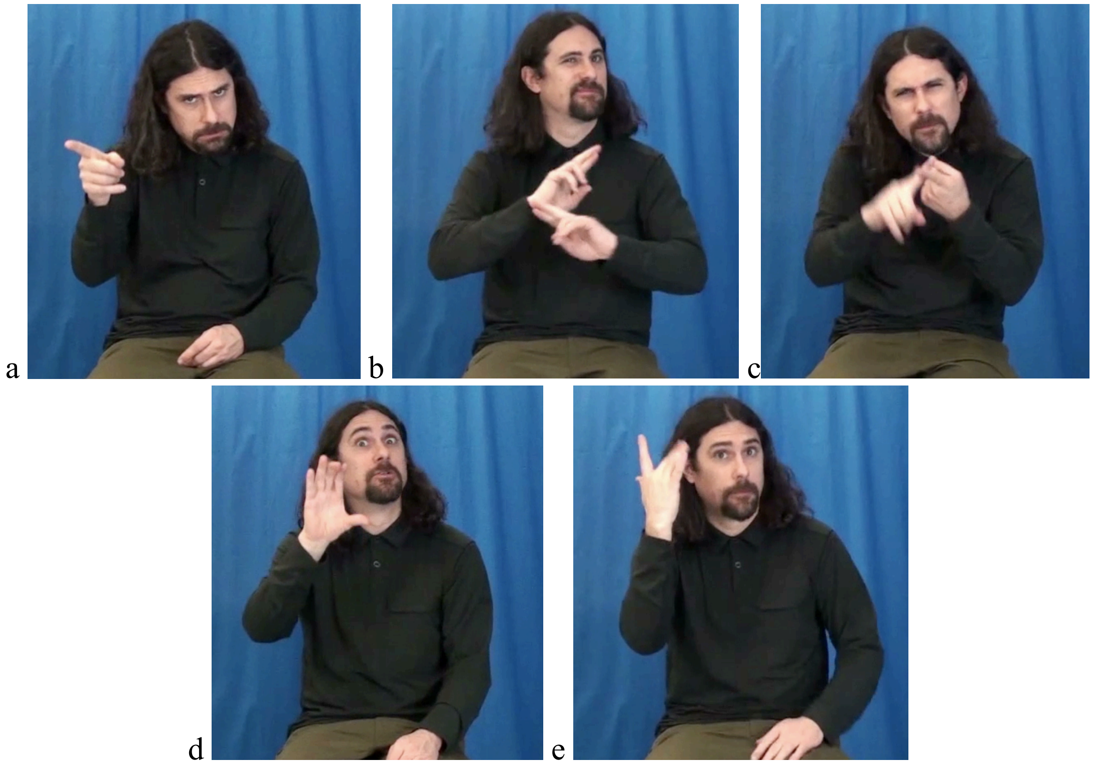


Figure 1: Still images taken from videos included in the stimuli collection described in this thesis, with each image illustrating a moment when a particular facial expressions is occurring: (a) Y/N-Question, (b) Negative, (c) WH-Question, (d) Topic, and (e) RH-Question.

There is variation in how syntactic facial expressions are performed during a given sentence, based on the length of the phrase when the facial expression occurs, the location of particular words during the phrase (e.g., NOT or WHO), the facial expressions that precede or follow, the overall speed of signing, and other factors. Figure 2 shows an example where ASL sentences with identical sequence of manual signs are interpreted differently based on the accompanying sequential or co-occurring facial expressions. Thus, for an animation synthesis system, it is insufficient to simply play a single pre-recorded version of this facial expression whenever it is needed. For this reason, we are researching how to model the performance of facial expressions in various contexts.

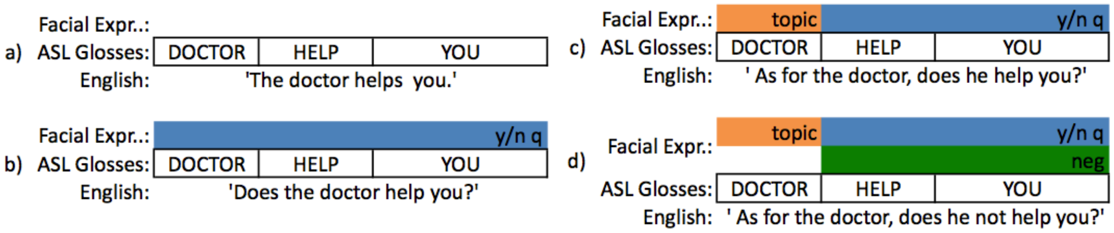


Figure 2: Differentiating (a) an ASL statement from: (b) an ASL question, (c) an ASL question with topic the doctor, (d) an ASL question with Negative based on the associated facial expression.

2.2 Non Syntactic Facial Expressions

We have grouped all the other types of sign language facial expressions under the category "non-syntactic" facial expressions. Although these types of facial expressions are not a primary focus for this thesis, they are briefly described here for the following reason: In the prior work surveyed in Chapter 9, sign language animation researchers have investigated the synthesis of some of these facial expressions for their animation platforms. To understand and benefit from their prior work, it is necessary that we understand the dynamics of these facial expressions and their coordination to the manual signs.

2.2.1 Lexical

A lexical facial expression usually involves mouthing, mouth patterns derived from the spoken language. The co-occurrence of such an expression distinguishes the meaning of two or more identical single lexical signs. E.g. the only difference between the ASL signs NOT-YET and LATE is the use of the 'TH' facial expression where the tongue is slightly protruding between the teeth (Reilly and Anderson, 2002). Figure 3 illustrates both signs as performed in the HANDSPEAK.COM dictionary.

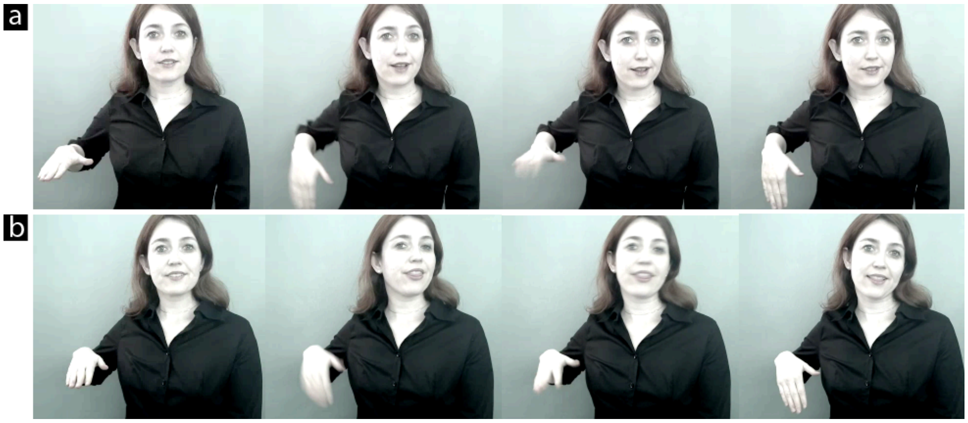


Figure 3: Example of ASL facial expression in lexical level (a) NOT-YET and (b) LATE signs. (source: www.handspeak.com)

2.2.2 Modifier or Adverbial

This facial expression often co-occurs with a predicate to semantically modify the meaning of the manual predicate. E.g. in ASL, the mouth morpheme 'OO', 'MM', and 'CHA' can describe the size small, average, and huge, respectively. Figure 4 illustrates an example of these modifiers in the phrase while describing the size of a book.

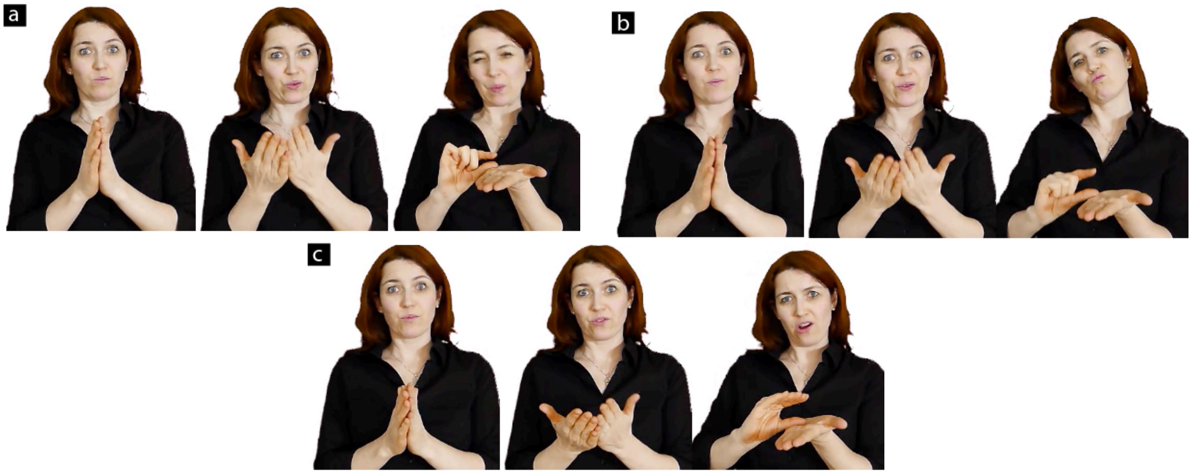


Figure 4: Example of adverbial ASL facial expressions: (a) 'OO' the book has a small number of pages, (b) 'MM' the book has an average number of pages, and (c) 'CHA' the book is thick in pages. (source: www.handspeak.com)

2.2.3 Paralinguistic

This group of facial expressions includes affective behaviors (e.g. emotions, illustrated in Figure 5) and prosodic behaviors such as emphasis and presentation of new information and old information. These facial expressions are not linguistically constrained in time by the manual signs and their scope can vary over the signed passages.

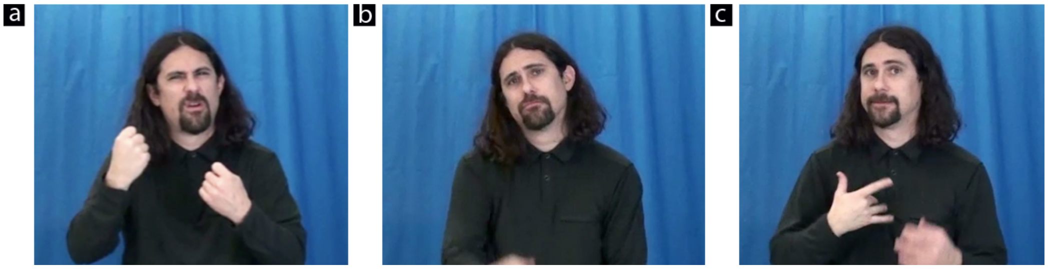


Figure 5: Still images with each image illustrating a moment when a particular facial expressions is occurring: (a) angry, (b) sad, and (c) ironic.

2.3 Animation Technologies in Sign Language

Animations of sign language refer to coordinated simultaneous 3D movements of an animated human-like character's body such as hands, shoulders, head, and face while conveying a message in sign language. One way to produce sign language animations would be for a skilled animator (fluent in that language) to create a virtual human using general-purpose 3D animation software based on motion capture data of a signer, as illustrated by the example in Figure 6a. Since this is time-consuming and largely dependent on the 3D animator's skill, researchers study automated techniques. The most automated approach is to develop "generation" software to automatically plan the words of a sign language sentence based on some input. For instance, in an automatic translation system, the input could be a written text or the output of a speech recognition system, which must be translated into sign language. While some researchers have investigated sign language generation and translation technologies (e.g. Lu, 2013), the state-of-the-art is still rather limited due to the linguistic challenges inherent in planning sign language sentences (Huenerfauth and Hanson, 2009). A less automated approach for producing sign language animation is to develop "scripting" software allowing a human to efficiently "word process" a set of sign language sentences, placing individual signs from a dictionary onto a timeline to be performed by an animated character. Such tools, e.g., VCom3D (DeWitt et al., 2003), EMBR (Heloir and Kipp, 2009), and JASigning (Jennings et al., 2010), can make use of pre-built dictionaries of sign animations. They incorporate software for automating selection of the transition movements between signs, and other detailed, and time-consuming to specify, aspects of the animation. Figure 6b illustrates an example of an animation as a result of one such approach from VCom3D.

Figure 6: Example of ASL animations produced by (a) a 3D animation artist using motion capture data over a few months (source: Hurdich, 2008) and (b) an ASL signer using a "scripting" software in a few minutes.

The linguistic complexity of sign languages and their use of the 3D space makes developing animation technologies challenging e.g., classifiers (e.g., Huenerfauth, 2006), inflective verbs (e.g., Lu, 2013), role shifting (e.g., McDonald et al., 2014), and facial expressions that communicate essential information during sentences, which are the topic of this thesis. There has been recent work by several groups to improve the state-of-art of facial expressions and non-manual signals for sign language animation. For example, Wolfe et al. (2011) used linguistic findings to drive eyebrow movement in animations of interrogative (WH-word) questions with or without co-occurrence of affect. Schmidt et al. (2013) used clustering techniques to obtain lexical facial expressions. Gibet et al. (2011) used machine-learning methods to map facial motion-capture data to animation blend-shapes. The contributions and limitations of these projects and others will be discussed in our literature survey in Chapter 9.

Part I

Evaluation of Facial Expressions in ASL Animations

Prologue to Part I

The overall objective of research on facial expression generation for sign language animations is to increase the naturalness and understandability of those animations, ultimately leading to better accessibility of information for people who are deaf with low English literacy. Therefore, a common agreement among researchers in sign language animation is the importance of involving signers in the evaluation process (e.g., Gibet et al., 2011; Wolfe et al. 2011; Kipp et al., 2011b).

However, sign language synthesis is a relatively new field; therefore few researchers have explicitly discussed methodological aspects when evaluating their work and even fewer have examined the challenges tied specifically to the facial expressions in their animations. For example, when assessing the quality of animated facial expressions, signers may not consciously notice a facial expression during a sign language passage that serves as stimuli in the user study (Huenerfauth et al., 2011). While participants may understand the meaning of the sentence as conveyed by both the manual signs and the face, they may not be overtly conscious of having seen the particular categories of facial expressions involved.

In Part I of this thesis we propose to conduct rigorous methodological research on how experiment design affects study outcomes when evaluating facial expressions in sign language animations; with ASL syntactic and emotional facial expressions as the case study. To provide the reader with background information and prior methodological choices of researchers evaluating their animations, we start this part of the thesis with discussion of related work in Chapter 3. Next, in Chapter 4 we will investigate the effect of ASL signers' involvement in design of stimuli and comprehension questions. The set of the stimuli as a result of this process will be released to the research community and details on their engineering steps will be described. We will also discuss the effect of videos as upper baseline and for presenting comprehension questions (Chapter 5), and eye-tracking as an alternative to recording question-responses from participants who are assessing sign language animations (Chapter 6). Last, we will identify relationships between (a) demographic and technology experience/attitude characteristics of participants and (b) the subjective and objective scores collected from them during the evaluation of sign language animation systems (Chapter 7).

Specifically, Part I of this thesis will explore each of these following five research questions:

RQ1*: Can our stimuli and comprehension questions that contain linguistic facial expressions measure whether participants understand the indented facial expression effectively?* (We will examine ***RQ1*** in Chapter 4.)

RQ2*: How does the modality (video of a human vs. a human-produced high-quality animation) of an upper baseline, presented for comparison purposes, affect the comprehension and subjective scores for the animation being evaluated?* (We will examine ***RQ2*** in Chapter 5.)

RQ3*: Does the modality (video of a human vs. a human-produced high-quality animation) of instructions/comprehension questions in a study affect the comprehension and subjective scores for the animation being evaluated?* (We will examine ***RQ3*** in Chapter 5.)

RQ4: *Could eye-tracking be effectively used as a complementary or an alternative unobtrusive way of evaluating sign language animations with facial expressions?* (We will examine **RQ4** in Chapter 6.)

RQ5: *Which are the eye-tracking metrics that correlate with evaluation judgments from participants during the evaluation of sign language animations with facial expressions?* (We will examine **RQ5** in Chapter 6.)

RQ6: *What demographic and technology-experience variables are predictive of participants' judgments during evaluation of sign language animations?* (We will examine **RQ6** in Chapter 7.)

Chapter 3 Related Work on Evaluation of Sign Language Animations[3]

While focusing on the evaluation challenges tied specifically to the evaluation of facial expressions, in this chapter we discuss the state-of-art in sign language animation evaluation. The methodological choices of previous researchers are organized in such a way that they follow the research questions, and thus the chapters in Part I of this thesis.

To evaluate their animation synthesis approaches, sign language researchers typically ask signers to view stimuli animations and then answer subjective Likert-scale questions (e.g., Wolfe et al., 2011), respond to comprehension questions (e.g., Huenerfauth, 2008), provide comments (e.g., Ebling and Glauert, 2013), write down the perceived message (e.g., Cox et al., 2002), or re-perform the perceived sign language passage (e.g., Kipp et al., 2011a). A side-by-side comparison between animations under different conditions (e.g. Huenerfauth and Lu, 2010) or between videos and animations (e.g. Krnoul et al., 2008) is often adopted.

An important issue in conducting any type of evaluation of sign language stimuli is recruiting and screening participants for the study. This task can be more challenging than one might originally expect. Researchers cannot simply ask potential participants if they are good signers or whether sign language is their first language: potential participants may answer those questions based on their feelings of connection to the Deaf Community, not based upon their

[3] The information presented in this chapter first appeared in manuscripts submitted as joint work with Professor Matt Huenerfauth, and graduate students working at LATLAb: Pengfei Lu, Allen Harper, Sarah Ebling, Kasmira Patel, Mackenzie Willard (Lu and Kacorri, 2012; Kacorri et al., 2013a; Kacorri et al., 2013b; Kacorri et al., 2013c; Kacorri et al., 2014; Kacorri et al., 2015).

sign language skills (Huenerfauth et al., 2008). Instead, researchers must elicit participants' signing skills through detailed questions about the age when they first started signing and other sign language exposure factors such as their educational experiences, family members who use sign language, and social groups that use sign language (e.g. Huenerfauth et al., 2008).

3.1 Engineering Stimuli and Comprehension Questions

Researchers studying facial expression of non-signing virtual humans, often evaluate only static faces, e.g., participants must identify the category of the facial expression or assign scores for intensity or sincerity from looking at a static screenshot of a virtual human (e.g. Wallraven et al., 2008). Because sign language facial expressions convey grammatical information and are governed by linguistic rules, additional care is needed to design useful stimuli and questions for evaluations. There is additional complexity in this case because users must evaluate the degree of intensity of facial expressions and the timing of the various phases of the facial movements over the signed message. Static images of faces are insufficient: animations are necessary in order to evaluate the timing dynamics of the facial movements during sentences. However, few researchers have explicitly discussed methodological aspects of animation stimuli design for facial expressions in sign language animation user-studies.

When considering the methodological issues in stimuli design where signers are involved, published research differs as to whether researchers invented their stimuli (1) originally as sign language sentences (Gibet et al., 2011; Schnepp et al., 2010) or (2) originally as written/spoken language sentences that were subsequently translated into sign language stimuli (Boulares and Jemni, 2012; San-Segundo et al., 2008). In Chapter 4, we will compare these two types of stimuli to further investigate the impact of signers' involvement early in the stimuli design process.

While it is relatively easy to ask a participant to rate subjectively whether they believe a particular animation stimulus was understandable, researchers have observed low correlation between a user's subjective impression of the understandability of a sign language animation and his/her actual success at answering comprehension questions about that animation (Huenerfauth et al., 2008). It is for this reason that researchers have made efforts to include an actual comprehension task (either a comprehension question about information content in the stimulus or a matching task that the user must perform based on this information). Given that the format of the comprehension questions and the answer choices must be accessible to the participants, these questions are often presented as sign language animations or video recordings of a human signer, e.g. (Schnepp and Shiver, 2011; Huenerfauth and Lu, 2010).

Previous researchers have structured the responses to be collected from participants in various forms. For instance, the participants may respond to questions by selecting a choice on a range from "Definitely Yes" to "Definitely No" (e.g., Huenerfauth and Lu, 2010), select an image that corresponds to their answer (thereby enabling participation by signers with limited English skills) (Huenerfauth and Lu, 2012), or provide open-ended answers performed in sign language (Schnepp and Shiver, 2011). In Chapter 4, we will discuss the response ranges for the subjective and comprehension questions we have adopted in our experimental studies.

3.2 Modality of Upper Baselines Used for Comparison

A challenge in using comprehension questions to evaluate animations of sign language is that the accuracy scores obtained from participants may depend on factors beyond the quality of the stimulus itself, e.g., the difficulty of the comprehension question or the memory skill of the participant. In order to obtain results whose meaning is more easily interpreted (independent of the difficulty of comprehension and subjective questions), researchers often compare their synthesized animations to a baseline for comparison, e.g. animations produced manually by signers or video recordings of human signers.

Many researchers have studied the usability of computer animations of virtual humans in various applications, including comparisons to videos of humans, e.g. (Ham et al., 2005; McDonnell et al., 2008; Russell et al., 2009). Obviously, a video of a human signer is an ideal of the visual fidelity and movement that animation researchers would like to achieve (and therefore it makes sense to use such a video as a basis for comparison in an experiment). Of course, the virtual human may never look visually identical to an actual human; so, one could argue that an animation produced manually by a skilled animator (who is a fluent signer) could be a more "fair" ideal for comparison.

In order to understand the diversity of experiment designs that previous researchers have employed, we searched the literature for examples of prior user studies (of both sign language animation and non-signing virtual human animation), and we noted the type of upper baseline used in those studies. We found that prior user studies can be organized into three categories (as illustrated in Figure 7).

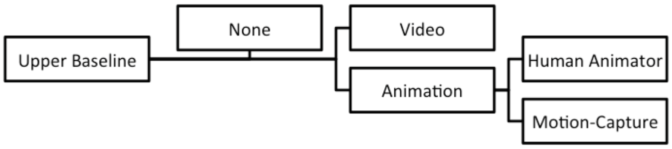


Figure 7: Upper baseline categories adopted by researchers when evaluating animations.

3.2.1 No Upper Baseline

The first category is research in which no upper baseline is mentioned. Although evaluation against a baseline usually results in more meaningful scores, many user-studies don't include any baselines. For example, researchers asked users to evaluate their deployed human animation without any baselines for comparison (Schnepp et al., 2010), improved their animations through iterative experiments (Davidson et al., 2000), defined the best parameters for their animation models through presentation of multiple versions (Davidson et al., 2000), compared the suitability of their available animated characters for a given task (Ow, 2009), or compared their results to previous similar studies (López-Colino et al., 2011). As discussed above, without an upper baseline for comparison in a study, it is more difficult to meaningfully interpret the results of an experiment. It is difficult to determine whether the evaluation scores obtained in the study were due to the quality of the stimuli themselves, or were due to the nature of the questions that were asked. For this reason, many researchers do choose to include upper baselines for comparison in their studies, as discussed in the following two sections.

3.2.2 Video Upper Baseline

The second category of studies that we have identified in the literature, are those in which video recordings of a human were used as an upper baseline for comparison to the animation under evaluation. For example, researchers have assessed comprehensibility while comparing their avatars to human signers (Kipp et al., 2001a), verified the visual quality of their animations by comparing them to video of human interpreters (Baldassarri et al., 2009; Baldassarri and Cerezo, 2012), or compared interpreter videos to animations in a medical domain (Morimoto et al., 2003; Morimoto et al., 2006). Researchers studying non-signing virtual characters have also sometimes used videos of humans as a baseline for comparison. For example, the expressiveness of a MPEG-4 face model (Ahlberg et al., 2002) and eye gazing in humanoid avatars in dyadic conversations (Garau et al., 2001) have been evaluated in comparison to human videos.

3.2.3 Animation Upper Baseline

The last category of studies that we have identified in the literature are those in which animations of a virtual character were used as an upper baseline in the evaluation; this seems to be the most popular approach for sign language animation (and non-signing virtual human or embodied agents) research. We noticed that the similarity of appearance between the virtual characters in the "upper baseline" animation and the character in the animation under evaluation varied across studies, and so, we decided to divide this research into two subcategories, according to the way in which the upper baseline animation was created and manipulated.

The first subcategory of prior research used upper baseline animations that were controlled by a human animator, without any motion-capture data. This is the approach used in prior studies at our lab, in which a human animator have carefully produced an animation of a virtual human, with the same appearance as the animation for evaluation, to serve as the upper baseline (Huenerfauth et al., 2011; Lu and Huenerfauth, 2011; Lu and Huenerfauth, 2012). Researchers studying the animation of non-signing virtual human characters have employed a similar methodology, e.g. (Bergmann, 2012) compared "average" models learned from the combined data of several speakers with individualized generated gestures based on empirically observed gestural behavior.

The second subcategory includes studies where the upper baseline was an animation produced, at least partially, from a motion-capture recording of a human. Sign language animation researchers have used this type of upper baseline in a variety of studies: to rate the understandability, naturalness of movement, and grammaticality of animations (Huenerfauth, 2006); to measure the comprehension of synthesized facial expressions (Gibet et al., 2011); to evaluate synthesized signs (Kennaway et al., 2007); or to elicit feedback on a variety of signing animations (Kipp et al., 2011b). These papers are listed in Table 1, which highlights the degree to which the upper baseline animation was similar to the other animation being evaluated in the study. When an "X" appears in a column, it means that the upper baseline shared a property with the animation being evaluated: the language of signing, the content of the signed message, the animation tool used to produce the animation, the appearance of the virtual human, and the background of the animation. In addition, researchers studying non-signing animations have also used virtual humans driven by motion-capture as upper baselines (Pražák et al., 2010).

Table 1: Similarity of the upper baseline animation to the animation being evaluated.

	Gibet et al., 2011	Huenerfauth, 2006	Kennaway et al., 2007	Kipp et al., 2011b
Language	x	x	x	-
Message Content	x	x	x	-
Animation Tool	x	x	-	-
Character Appearance	x	-	-	-
Animation Background	x	x	x	-

Thus, as discussed in the previous three sections of this chapter, previous researchers have made use of both virtual-human-animation stimuli and human-video stimuli as comparison baselines in evaluation studies; however, we have found that no prior researchers have investigated empirically how this choice of upper baseline may affect the overall results of a study. For instance, perhaps participants would judge animations more harshly when seen in comparison to actual videos of real human signers. Such an empirical investigation of these possible effects is necessary so that it would be possible to fairly compare the results of studies that made use of different upper baselines. Therefore, we will investigate this issue in Chapter 5.

3.3 Modality of Instructions and Evaluation Questions

Our discussion in Section 3.2 focused on the upper baseline shown for comparison in a study, with a focus on whether this baseline was in the form of human-video or animation. During our review of the literature, we also noted another experiment design parameter that varied across studies: specifically, researchers vary in how they chose to convey instructions or comprehension/evaluation questions to participants. Based on the modality of presentation, we identified four categories of prior studies (listed below).

- The first category includes studies in which a *human experimenter* signs: instructions (Schnepp et al., 2010; Kipp et al., 2011a; Kennaway et al., 2007), questions (Davidson et al., 2000), or guidance to a focus group (Kipp et al., 2011b).
- The second category includes studies where *video recordings* of a human are used to present instructions within the user-interface of the software displaying the animation stimuli (Schnepp and Shiver, 2011; Schnepp et al., 2011) or provided as explanations for the questions in an online study (Kipp et al., 2011b).
- The third category includes studies in which an *animated character* (similar in appearance to the virtual human in the animations being evaluated) performs comprehension questions (e.g., Huenerfauth and Lu, 2010).
- The last category includes studies in which *written text* is used to present study instructions (Baldassarri et al., 2009; López-Colino and Colás, 2011) or questionnaires (Gibet et al., 2011; Ow, 2009).

Section 5.4 will investigate how the modality of presentation of study elements beyond the stories themselves (i.e., the comprehension questions) may affect the results of an experiment. For instance, perhaps users achieve higher scores in studies in which the instructions and comprehension questions were conveyed in the form of videos of a human signer, as opposed to animations of a virtual human.

3.4 User Studies with Deaf Participants and Eye-tracking

Another methodological issue that we propose to investigate in this dissertation research is how eye-tracking technology could be used to evaluate animations of sign language. This section provides essential information about this technology and discusses prior work in user studies that involve deaf participants and eye-tracking. While several prior authors have surveyed the use of eye tracking in the field of human computer interaction (Duchowski, 2002; Jacob and Karn, 2003; Rayner, 1998), some basic information about eye-tracking methodology is briefly summarized below.

The bright-pupil technique used in most of the state-of-art eye-tracking system and in our approach (to be discussed in Chapter 6) employs a near infrared light source, which illuminates the pupil (the "red-eye" effect) and creates a reflection on the cornea (first Purkinje image). Image processing software then identifies: (1) the center of the pupil and (2) the corneal reflection. By comparing the relationship between these two artifacts in the eye video, the point of gaze on the stimuli can be determined.

The eye-tracking system records the horizontal and vertical coordinates on the computer screen where the eye is aimed. Human eye gaze tends to move rapidly from one location to another, during movements called "saccades." Moments when the eye is relatively stationary are called "fixations." Thus, eye-tracking data is usually processed into a list of the fixations that occur during a study, each with a: start-time, end-time, horizontal and vertical screen coordinates, and other information.

To facilitate analysis, researchers typically perform one more step of processing on the fixation list. They define regions of the computer screen (during specific periods of time in the study) that are significant; such regions are called "Areas of Interest" or AOIs. For example, during a usability study, each button on a user-interface may be defined as an AOI, consisting of the shape and location of the button and the time duration when it was visible. Each fixation in the fixation list can thus be labeled as to whether it was within an AOI.

Eye tracking enables researchers to collect a detailed sequential record of how users visually interact with stimuli. While the link between visual attention and cognitive processes is not completely understood, there is a general consensus among eye-tracking researchers of the validity of the so-called "eye-mind" hypothesis, that: "eye movements and attention are assumed to serve useful purposes connected to the visual task" (Kowler, 2011). Modern video-based eye-tracking has been applied to diverse areas of research, including: the psychology of reading (e.g., Rayner, 2009), web search (e.g., Goldberg et al., 2002; Guan and Cutrell, 2007), user-interface usability (e.g., Goldberg and Kotval, 1999), cognitive workload estimation (e.g., Bartels and Marshall, 2006), and cognitive modeling (e.g., Halverson and Hornof, 2007; Salvucci and Anderson, 2001).

3.4.1 Eye-Tracking Participants who are Deaf

Researchers have conducted eye-tracking studies with participants who are deaf, to examine reading strategies, perception, or software usability; some of these studies involve comparisons between deaf and hearing participants. For instance, in (Watanabe et al., 2011), deaf and hearing subjects rated static face images for ten different emotional states while their eye movements were recorded with a desktop eye tracker. Eye metrics were calculated for proportional fixation time and average gaze duration. Interestingly, while no differences were found between the two groups in how they rated the images, there were measurable differences in eye movement patterns. Specifically, deaf subjects had greater proportional fixation times as well as mean gaze duration on the eyes AOI while hearing subjects had more fixation time as well as longer gaze durations on the nose AOI.

Other researchers have studied reading, e.g. (Jensen and Pedersen, 2011) compared reading strategies used by deaf and hearing participants. A desktop-mounted eye-tracker monitored eye-movement behaviors while participants read Dutch texts on websites. Chapdelaine et al. (2007) compared the eye movements of deaf and hearing subjects when watching captioned videos. They recorded proportional fixation time and gaze duration for several AOIs in their videos: faces in the videos, areas of motion in the videos, and caption regions. They found that deaf users spent less proportional fixation time on the captions than the hearing group, but the deaf users scored higher on recall tests of information from the videos.

Finally, several researchers have used eye-tracking technology to study how deaf students balance their attention across several sources of information during classroom lectures.

- Marschark et al. (2005) examined the eye-movements while college students were presented visual stimuli consisting of a lecturer, an interpreter, and a projection screen. Their participants included deaf students who were: skilled signers and less experienced signers. Conducted in a classroom setting, this experiment used a head mounted eye-tracker worn by the participants. Eye movement data was recorded for proportional fixation time, mean gaze duration (average length of time per gaze), and transitions between the three AOIs. They found that both groups of deaf students had similar eye metrics.

- Cavender et al. (2009) conducted a usability study of a multi-modal educational user interface for deaf students that had four regions (AOIs): the lecturer, the interpreter, slides, and captions. To assist students in noticing when a slide change occurred, notification schemes were implemented that altered the user interface component (e.g., color change) at that moment. A desktop eye tracking system was used to capture fixation data for the four AOIs. They found that the students spent more time looking at the interpreter and captions, as opposed to allocating their visual attention to the instructor or the slides.

3.4.2 Eye-Tracking with Sign-Language Video

While we are not aware of any prior studies that have used eye-tracking techniques to evaluate sign language *animations*, this section describes some examples of studies that have recorded participants viewing *videos* of sign language. For instance, Cavender et al. (2005) conducted a preliminary study to evaluate the understandability of videos of sign language displayed at different sizes (based on screen sizes of mobile phones) and video-compression rates. Four participants viewed videos while eye-tracked, and they answered evaluation questions about each video. The authors found most fixations were close to the signer's mouth in the videos. They also found that the path length traced by fixations was shorter for the medium-sized video in their study, which was the video that received the highest subjective scores from participants. Finally, the authors analyzed instances when participants' gaze transitioned away from the signer's face; this occurred during some fingerspelling, when hands moved to the bottom of the screen, when the signer looked away from the camera, or when the signer pointed to locations outside the video.

Muir and Richardson (2005) performed an eye tracking study to determine how British Sign Language (BSL) signers use their central (high-resolution) vision and peripheral vision when viewing BSL videos. Their earlier work had suggested that signers tend to use their central vision on the face of a signer, and they tend to use peripheral vision for hand movements, fingerspelling, and body movements. In (Muir and Richardson, 2005), BSL signers watched three videos that varied in how visually challenging they were to view: (1) close-up above-the-waist camera view of the signer with no fingerspelling or body movement, (2) distant above-the-knees view of the signer with use of some fingerspelling, (3) distant above-the-knees view of the signer with use of fingerspelling and body movements. Participants' eye movements were recorded and proportional fixation time was computed over five AOIs: upper face, lower face, hands, fingers, upper body, and lower body. (The researchers had to carefully view the recordings of their eye-tracking data to determine when the participant was looking at each of these moving portions of the signer's body.)

Detailed signs and fingerspelling did not accumulate large proportional fixation time, indicating that participants used their peripheral vision to observe these aspects of sign language video. For all three videos, the face AOIs received the most proportional fixation time: 88%, 82%, 60% respectively. Video 3 included upper body movement, and participants spent more time looking at the upper body of the signer. During video 1, participants looked at the upper face 72% and lower face 16%, but during video 2 (more distant view of the signer), they looked at the upper face 47% and lower face 35%. These results are of interest to our current research questions (RQ4 and RQ5) because they indicate that when participants view sign language videos that have lower clarity (because the signer is more distant from the camera), their attention may shift to different areas of the video image, perhaps in an effort to search for the AOI with the most useful and visible information. This suggests that studying proportional fixation time on the face might be a useful way to analyze eye-tracking data when participants are viewing sign language videos (or animations) of different quality.

Emmorey et al. (2009) conducted an eye tracking experiment to investigate differences in eye movement patterns between native and beginner ASL signers. The authors hypothesized that novice signers would have a smaller visual field from which to extract information from a signer. This in turn would lead to: less time fixating on the singer's face, more fixations on the lower mouth and upper body, and more transitions away from the face to the hands and lower body. Two stories were constructed which differed in the amount of fingerspelling and use of locative classifier constructions, signs that convey spatial information, investigated in (Huenerfauth, 2004), with the goal of inducing more transitions in novice signers due to a restricted perceptual span. Both native and novice signers had similar proportional fixation times (89%) on the face; however, novices spent significantly more time fixating on the signer's mouth than native signers, who spent more time fixating on the signer's eyes. Also, neither novices nor native signers made transitions to the hands during fingerspelling, but did make transitions towards classifier constructions.

3.5 Reported Participant's Demographics and Technology Experience

Our last methodological issue that we propose to examine in this dissertation research is how demographic characteristics or technology experience of participants in a study may affect the results collected. We therefore examine how prior sign language animation researchers have considered these demographic or attitude variables when conducting their studies. We focus on studies that have been conducted with deaf participants in the context of evaluating sign language animations (Section 3.5.1) or to determine general acceptance of such technology (Section 3.5.2). While some researchers have conducted studies of how various demographic or health factors affect technology use and acceptance, e.g., (CREATE, 2015; Crabb and Hanson, 2014; Rosen et al., 2013), this section focuses on studies with deaf participants evaluating sign language animations.

3.5.1 Demographics in Prior Studies

We surveyed prior sign language animation studies to identify the types of participant characteristics or technology experience/attitudes that researchers reported. Our goal was to understand the diversity of participants in prior studies and the types of data that researchers commonly collect. While there are a few examples of published results where only the number of participants and how they are self-identified is reported, e.g. (Moemedi, 2010; Yang et al., 2014), in general the trend in the field is to include more information about the sampled population. Table 2 presents examples of representative papers, similar patterns may be found when examining larger surveys of prior evaluation studies, e.g., (Ebling and Glauert, 2015; Huenerfauth et al., 2008; Kipp et al., 2011b).

Common characteristics reported in studies include the age range of participants, the gender ratio of participants, and the ratio of participants identifying as deaf or hard-of-hearing (this is indicated by the term “describe” in Table 2). Studies vary in how they measure and report the level of sign-language skill of their participants, e.g., using “signing frequency” or “self-reported sign language skills.” We also see variability in whether researchers ask about participants’ exposure to technology; for example, (Hayward et al., 2010) included questions about “computer expertise.” Researchers in (Verlinden et al., 2001) noted that only those participants who were unfamiliar using the Internet had negative attitudes towards their avatar; the authors stated, “This suggests that acceptance of the avatar is greater for web-surfers and that this acceptance may increase as a person becomes more familiar with the Internet.” There is further variation in whether researchers asked participants about their attitude towards animated avatars (“attitude to avatar”) or their views about the future potential of signing animations in different real-world contexts (“animation usage”). When included, such questions provide insight into how participants may see this technology being applied, e.g. as an educational tool (Hayward et al., 2010) or for giving information in public spaces (Kennaway et al., 2007).

Table 2: Demographic and Technology Experience Characteristics Reported in Example User Studies.

Paper	*Demographic*	*Technology*	*Attitudes*
Hayward et al., 2010	age, gender, describe, profession	computer expertise	animation usage
Kennaway et al., 2007	age, gender, describe, signing frequency, preferred language		animation usage
Verlinden et al., 2001	age, gender, preferred language		attitude to avatar
Gibet et al.,2011	age, gender, describe, self-reported SL skills, location		attitude to avatar

Table 3: Demographic Profile of Participants in Prior Studies

Paper	*Age*	*Female:Male*	*Description*	*Assessing Signing Skills*
Hayward et al., 2010	35-50	4 : 1	Deaf	Deaf educators
Kennaway et al., 2007	16-66	"slightly less female"	"most were deaf, some were hard-of-hearing"	"all were good signers… all using signing on a daily basis"
Verlinden et al., 2001	20-53	5 : 4	deaf	Some had preference for sign language; others had no preference between signing or text.
Gibet et al.,2011	19-56	18 : 7	17 deaf, 8 hearing	8 "good," 6 "very good," 11 "native/expert"

Table 3 shows some values of the most commonly reported demographic characteristics from the papers in Table 2. We can see wide variation in the demographic characteristics of users in prior sign language animation studies. For example there is especially wide variation in how researchers assess the signing skills of participants to determine whether they have sufficient fluency or native-level skill to participate in the study, e.g., some described what language their participants preferred (Kennaway et al., 2007; Verlinden et al., 2001) and how often they used signing (Kennaway et al., 2007).

A key question arises from examining this table: Do these differences in the demographic characteristics of the population of users in the study have an impact on the comprehension scores or subjective judgments of the participants? Knowing the answer to this question would make it easier to compare the results across different studies (so that we would know whether a particular set of participants might have been pre-disposed to have positive or negative evaluations of ASL animations). Thus, the goal of our study in Chapter 7 is to identify demographic characteristics or technology experience/attitude factors that relate to user's scores in evaluation studies. Based on these results, we will propose a set of standard questions that could be asked of participants in a user study (and thereby reported by researchers in their publications) to facilitate comparison of results across papers.

Some studies have included anecdotal evidence of relationships between (a) certain participant characteristics and (b) the subjective judgments or comprehension scores for sign language animation: e.g., the "web-surfers" comment in (Verlinden et al., 2001). However, due to the relatively small sample size of most prior studies, researchers rarely present quantitative results for sub-populations. We are not aware of any prior study that conducted an exploration of whether a large variety of participant characteristics may relate to evaluation scores for sign language animation.

3.5.2 Acceptance of Multiple Signing Avatars

In this section we discuss prior work on acceptance of sign language avatars with a focus on the aspects that will be addressed by our user study in Chapter 7, which examines how demographic characteristics or technology experience of participants in a study may affect the results collected.

Kipp et al. (2011b) carried out the most comprehensive study to date with participants evaluating multiple sign language avatars. In focus groups, eight German Sign Language signers were presented with six avatars signing content in different sign languages. Participants commented on the stiff or unrealistic movements, lack of facial expression, etc. They also discussed how avatar appearance should suit the audience (e.g., cartoonish is OK for kids). One concern with this study design was that researchers showed German Sign Language users animations in American Sign Language (and other unfamiliar languages), which limited their ability to offer critiques. Another concern was that researchers showed participants some hand-animated avatars (produced through a painstaking process of carefully posing the character). Current sign language animation research focuses on synthesized animation, in which software automatically selects aspects of the movement to allow for generation of animations from a sparse input script. Hand-animated characters may appear high-quality, but they are time-consuming to produce and have fewer potential accessibility applications for deaf users. They are not an accurate representation of the state-of-the-art of sign language animation synthesis technologies, and thus they may be misleading to show to participants in a study. Taking into account these previous findings, our new study (Chapter 7) will be designed such that all the stimuli are consisted of utterances in one sign language only (ASL). None of our avatars will be hand-animated: to avoid setting unrealistic expectations of the state-of-the-art.

Kipp et al. (2011b) also conducted an online survey (N=317), in which participants rated three avatars (one was hand-animated) on a 5-point scale in regard to: comprehensibility, facial expression, naturalness, charisma, movements, mouthing, appearance, hand shapes, and clothing. The hand-animated avatar received higher scores. In our new study (Chapter 7), we will include objective comprehension questions to measure participants' understanding. Selfreports of understanding typically have low correlation to a participant's accuracy at answering comprehension questions (Huenerfauth et al., 2008).

Notably, in both the focus group and the online survey, the authors observed higher scores in response the questions "Do you think avatars are useful?" and "Do you think Deaf people would use avatars?" when asked at the end of the study (compared to the beginning). The authors speculate that additional exposure to animations influenced participants' responses. In our new study in Chapter 7, we will include a question about whether participants had previously seen computer animations of sign language (details Section 7.1.2).

Participants in (Kipp et al., 2011b) also suggested use-cases for signing avatars, including: public transit, movies/entertainment, government and educational websites, and other areas. In our new study, we will also ask participants to judge the usefulness of signing avatars in various contexts: information on websites, for public places (e.g. airport, train station), as a virtual interpreter in a face-to-face meeting, as a virtual interpreter for telephone relay, etc.

While (Kipp et al., 2011b) collected some demographics (gender, age, deaf/hard-of-hearing/hearing, and profession), they did not analyze the data to look for relationships between these factors and the survey responses. Our new study will include a regression analysis to identify demographic and/or experience factors that related to the participants' subjective responses and comprehension scores.

Given the online modality of (Kipp et al., 2011b), there is a possibility that participants were more technology-savvy than the general population. In our new study, we will conduct an in-person survey in which participants will evaluate sign language animations; researchers will be traveling to meet participants at convenient locations. Our goal is to encourage participation of less technology-savvy individuals and to allow for us to confirm that participants will meet our study criteria (and will be accurately reporting demographic data).

Chapter 4 Stimulus and Comprehension Question Design

As discussed in the previous chapter, while we must evaluate the quality of the facial expressions in an ASL animation to advance research in this field, it is difficult due to the subtle and complex manner in which facial expressions affect the meaning of sentences. It can be challenging to design experiments that probe whether human participants watching an ASL animation have understood the information that should have been conveyed by facial expressions.

The easiest characteristics of an animation stimulus to evaluate are those that represent categorical information: for instance, to determine whether (or not) a sentence with a facial expression was correctly interpreted as a question by a participant. In this case, it is possible to invent experiments to determine whether a human watching an animation interpreted it as a declarative sentence or as a question by asking appropriate evaluation questions. However, some ASL facial expressions convey information in matters of degree, e.g., an emotional facial expression can convey continuous degrees (by intensity of eye-brow movement, etc.). Measuring whether someone has successfully understood the correct degree is more difficult.

In the most challenging case, a facial expression may not affect the superficial meaning of a sentence but only the implications that can be drawn. For instance, when a signer performs “I didn’t order a soda” with a cluster of behaviors (including frowning and head tilt) during the sign “I,” it can indicate that the signer believes someone else ordered the soda. With facial prominence on the sign “soda,” it could indicate that the signer placed an order, but for something else. In either case, the basic information is the same: the signer did not order a soda, but a different implication can be made.

For inspiration as to how to use comprehension questions to evaluate stimuli, Huenerfauth et al. (2011) considered prior research on how humans interpret and understand speech with various prosody (Allbritton et al., 1996; Price et al., 1991) (speed, loudness, and pitch changes). Researchers designed sets of sentences that, in the absence of prosodic information, contain ambiguity in how they can be interpreted. When prosodic information is added, then one interpretation is clearly correct. Participants in the studies listen to audio performances of these sentences and answer questions about their meaning. These questions are carefully engineered such that someone would answer the question differently – based on which of the alternative possible interpretations of the spoken sentence they had mentally constructed. For example, someone who heard the sentence “I didn’t order a soda” (with prominence on “I”) may be more likely to respond affirmatively to a question asking: “Does the speaker think that someone else ordered a soda?” In later sections of this chapter, we will design stimuli and questions that make use of similar principles.

In this chapter, we investigate how to design sign language animation stimuli that contain meaningful facial expressions, with accompanying evaluation questions that enable us to measure whether a participant has successfully understood the intended meaning of the facial expression. During this chapter, we will discuss the design, evaluation, and release (to the research community) of a collection of experimental stimuli for use in evaluations of facial expressions in ASL animation.

4.1 Pilot Testing of Experimental Stimuli[4]

Through pilot testing of various stimuli and questions, our goal was to identify a methodology for designing stimuli and conducting experiments to measure the quality of facial expressions in an ASL animation. We want to evaluate whether facial expressions in an ASL animation enable participants who view the animations to identify the content of the sentences being performed. In the studies presented in this section, participants look at animations of a virtual human character telling a short story in ASL, and they answer questions about each story. Each story includes one category of facial expression (e.g., Yes-No questions, sadness, etc.). The animations displayed are one of two types: (i) with facial expressions carefully produced by a human animator or (ii) without appropriate facial expressions (i.e., the face doesn't move).

While the experiments in this section are only pilot studies used to confirm our methodology, in later work, when we begin to investigate facial expression animation synthesis, our experiments will contain a third type of animation: (iii) with facial expressions planned by our automatic synthesis software.

4.1.1 English-to-ASL versus ASL-Originated Stimuli Design

Deciding on these short stories shown in a user study, creating the animations and creating the comprehension questions for each story, is a process that we refer to as "stimuli design," and the manner in which this is done can affect the scores collected in the study. In this section we will present two alternative methods for "stimuli design" to determine which is best for

[4] This section describes joint work with Pengfei Lu, a Ph.D. student in Computer Science at CUNY, and Professor Matt Huenerfauth (Kacorri et al., 2013c).

conducting ASL facial expression user-studies. The result of their comparison will be discussed in Section 4.1.2.

For each of the alternative stimuli-design methods, there are several variables that we consider: (i) whether the stimuli stories originated in English or in ASL, (ii) the amount of involvement of a professional signer in designing the stimuli, (iii) the categories of facial expressions included in the stimuli, and (iv) the complexity of stimuli (i.e., number of words per story). In order to investigate how some of these variables affect participants' judgments in an evaluation study, we conducted a comparison of two stimuli sets, which we refer to as our "English-to-ASL" stimuli and our "ASL-originated" stimuli. The primary difference between the sets is the degree of involvement of a ASL signer in the stimuli-creation process (leading to more fluent ASL sentences in the "ASL-originated" set) and the categories of facial expressions included in each stimuli set. The English-to-ASL stimuli were first evaluated in a study in 2011 (Huenerfauth et al., 2011), and the "ASL-originated" stimuli were evaluated in a new study we conducted in 2013.

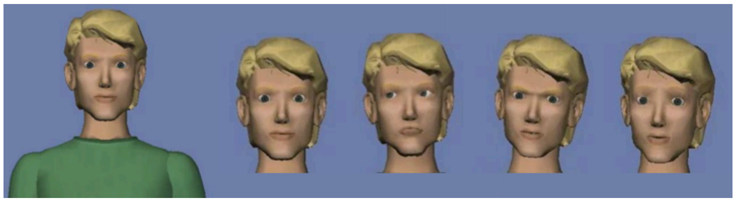
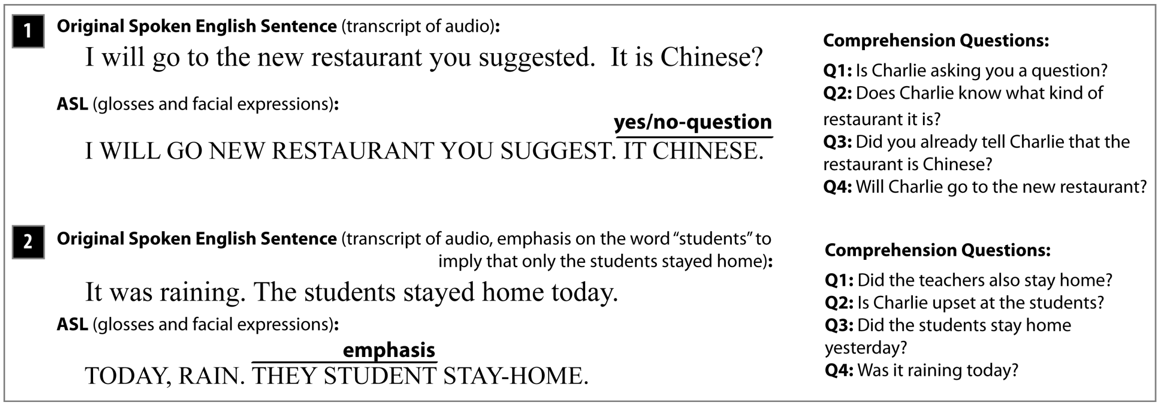


Figure 8: ASL character and some of the available facial expressions.

In both studies, ASL signers watched animations of a virtual human character, as shown in Figure 8, telling a short story in ASL. The story was either (i) with facial expressions added by a ASL signer or (ii) without facial expressions added. Then, participants answered comprehension questions carefully engineered to capture the possible confusion introduced by a misinterpretation of the face. An ASL signer, who is a professional interpreter, conducted all the instructions and interactions. The animations in both studies were created using identical commercial sign language animation software, Vcom3D Sign Smith Studio (VCOM3D, 2014).

Huenerfauth et al. (2011) methodology for creating the "English-to-ASL" stimuli was to begin with an English sentence whose meaning would change with/without prosody, and then the authors attempted to translate the sentence into ASL in a manner that would preserve this reliance of the prosodic information (conveyed by facial expression instead of spoken prosody). Specifically, they asked a signer (who works as a professional interpreter) to: (i) translate each of the English passages into ASL, (ii) use the Vcom3D Sign Smith Studio to produce the ASL animations, and (iii) choose from the available facial expressions repertoire the facial expressions that she thinks linguistically or naturally conveyed the prosodic information for the ASL stimuli. Each animation was produced in two versions: with and without facial expressions added.

There were a total of 28 stimuli with an average of 9 signs in length, and at least one facial expression per story. Figure 9 shows two examples of stimuli used, as original English and ASL translated transcriptions, and the corresponding comprehension questions. The bars over the script indicate the facial expression to be performed during some of the signs. The stimuli can be divided into 5 categories (Figure 10), based on the facial expression. The number of stimuli per category is given in parentheses.



1 **Original Spoken English Sentence** (transcript of audio)**:**

I will go to the new restaurant you suggested. It is Chinese?

ASL (glosses and facial expressions)**:**

yes/no-question

I WILL GO NEW RESTAURANT YOU SUGGEST. IT CHINESE.

Comprehension Questions:

Q1: Is Charlie asking you a question?
Q2: Does Charlie know what kind of restaurant it is?
Q3: Did you already tell Charlie that the restaurant is Chinese?
Q4: Will Charlie go to the new restaurant?

2 **Original Spoken English Sentence** (transcript of audio, emphasis on the word "students" to imply that only the students stayed home)**:**

It was raining. The students stayed home today.

ASL (glosses and facial expressions)**:**

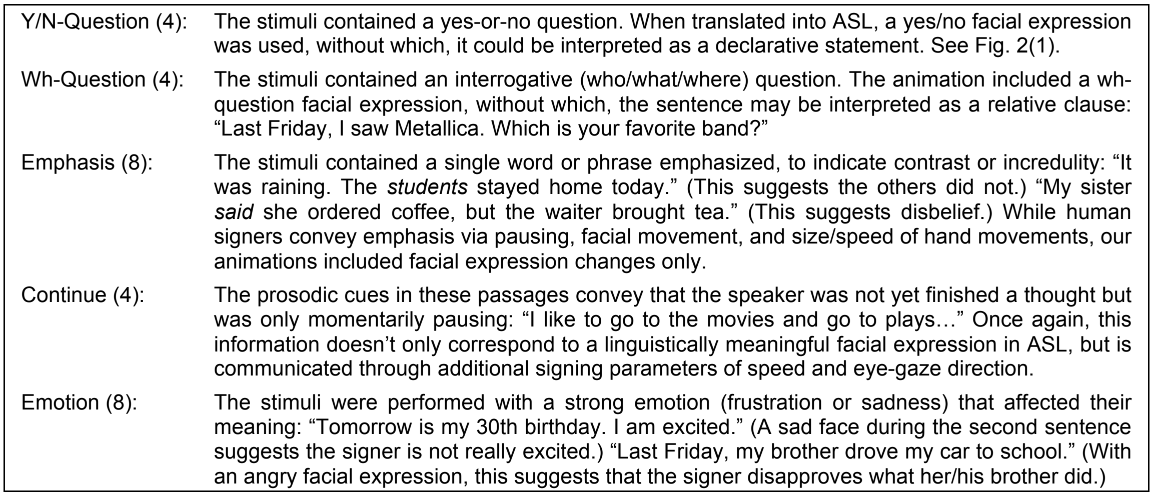
emphasis

TODAY, RAIN. THEY STUDENT STAY-HOME.

Comprehension Questions:

Q1: Did the teachers also stay home?
Q2: Is Charlie upset at the students?
Q3: Did the students stay home yesterday?
Q4: Was it raining today?

Figure 9: English-to-ASL Stimuli Set examples: (1) yes/no-question and (2) emphasis. The name "Charlie" refers to the animated character or human signer performing the passage.



Y/N-Question (4): The stimuli contained a yes-or-no question. When translated into ASL, a yes/no facial expression was used, without which, it could be interpreted as a declarative statement. See Fig. 2(1).

Wh-Question (4): The stimuli contained an interrogative (who/what/where) question. The animation included a wh-question facial expression, without which, the sentence may be interpreted as a relative clause: "Last Friday, I saw Metallica. Which is your favorite band?"

Emphasis (8): The stimuli contained a single word or phrase emphasized, to indicate contrast or incredulity: "It was raining. The *students* stayed home today." (This suggests the others did not.) "My sister *said* she ordered coffee, but the waiter brought tea." (This suggests disbelief.) While human signers convey emphasis via pausing, facial movement, and size/speed of hand movements, our animations included facial expression changes only.

Continue (4): The prosodic cues in these passages convey that the speaker was not yet finished a thought but was only momentarily pausing: "I like to go to the movies and go to plays…" Once again, this information doesn't only correspond to a linguistically meaningful facial expression in ASL, but is communicated through additional signing parameters of speed and eye-gaze direction.

Emotion (8): The stimuli were performed with a strong emotion (frustration or sadness) that affected their meaning: "Tomorrow is my 30th birthday. I am excited." (A sad face during the second sentence suggests the signer is not really excited.) "Last Friday, my brother drove my car to school." (With an angry facial expression, this suggests that the signer disapproves what her/his brother did.)

Figure 10: Five categories of facial expression in the English-to-ASL Stimuli Set.

Starting with English speech passages when creating stimuli for an ASL animation study may seem like a good approach given: (i) it is true to the goal of ASL animation synthesis, that is converting English text or speech to comprehensible ASL animations; (ii) it makes use of passages that are carefully engineered and successfully applied to collect users interpretation, and (iii) prosodic information in English are often conveyed by facial expressions in ASL. However, it can lead to various problems. First, the English influence might result in ASL stimuli following an English word order, e.g. the ASL sentence in Figure 9(1) has a rather English-like word order. Second, some of the categories like Emphasis and Continue are communicated by a cluster of behaviors, not a single ASL facial expression.

To overcome the above challenges, we designed a second set of stimuli with the help of a signer, who first wrote a script for each stimulus with the facial expressions indicated by bars over the glosses they appear. Then we recorded a second ASL signer performing these scripts in an ASL-focused lab-environment with little English influence. Next, another signer created animated versions of these stories by consulting the recorded videos. Again, both stories and questions were engineered in such a way that the wrong answers would indicate that the users misunderstood the facial expression displayed, as shown in Figure 11.

We initially created a total of 38 ASL stories, and the signer selected 21[5] of the most fluent animations (average of 9 signs per story). The resulting stimuli did not include any sentences in the categories of Emphasis and Continue used in the first set. They were replaced by new categories that actually correspond to particular types of facial expressions recognized by ASL linguists, such as: “topic,” “rhetorical question,” and “negative.” The stimuli can be divided in 6 categories (Figure 12).

[5] Appendix A provides more details on the stimuli and comprehension questions, whose codenames are: E1, E2, E4-E6, E8, W2-W4, Y3, Y4, Y6, R3, R5, R7, N1-N3, and T3-T5.

1 **ASL** (glosses and facial expressions)**:**

NEXT YEAR, YOUR SISTER ME VISIT WILL (SHAKE HEAD "YES").

yes/no-question

LIVE WASHINGTON_DC SHE(POINT)?

English Translation:

I will visit your sister next year. Does she live in Washington DC?

Comprehension Questions:

Q1: Is Charlie asking you a question?
Q2: Does Charlie think your sister lives in Washington?
Q3: Does Charlie know where your sister lives?
Q4: Will Charlie visit your sister?

2 **ASL** (glosses and facial expressions)**:**

EVERYDAY ME SCHOOL GO-GO. BACK-AND-FORTH ME TRAIN.

topic

BUS TAKE FOREVER.

English Translation (imply: Charlie don't take the bus):

I go to school everyday. I take the train back and forth.
The bus just takes forever.

Comprehension Questions:

Q1: Does train take a long time?
Q2: Does Charlie everyday ride the bus?
Q3: Does Charlie everyday ride the train?
Q4: Does it take for Charlie a really long time to get to school?

Figure 11: Example of an y/n-question (1) and a topic (2) stimulus in the ASL-originated set.

Y/N-Question (3):	The stimuli contained a yes-or-no question; without facial expression, it could be interpreted as a declarative statement. See Fig. 4(1).
Wh-Question (3):	The stimuli contained an interrogative question; without the wh-question facial expression, the human viewer could misunderstand the sentence, e.g. COMPUTER YOU BOUGHT WHERE? #SALLY FAVORITE SHOPPING CENTER. (In this paper, # indicates a finger-spelled word).
Rh-Question (3):	The stimuli contained a rhetorical question; without the facial expression, the sentence boundary may be unclear, e.g. "THIS YEAR ASL I LEARN HOW. I PRACTICE." (With a rhetorical-question face over the first sentence.)
Topic (3):	The stimuli contained a topic facial-expression indicating that some words at the beginning of a sentence are an important topic, e.g. Fig. 4(2).
Negation (3):	The stimuli contain a negative facial expression and head movement, e.g., "#ALEX TEND TAKE-UP MATH CLASS. NOW SEMESTER, SCHOOL HAVE SCIENCE CLASS. ALEX TAKE-UP TWO CLASS." with a negation facial expression over "HAVE SCIENCE CLASS" would indicate that a school does not offer science classes (the opposite meaning of the sentence).
Emotion (6):	The stimuli were performed with a strong emotion (frustration, sadness, or irony) that affected their meaning: "YESTERDAY, MY SISTER CAT BRING." (A sad face suggests that the signer is not being happy to receive a cat from her/his sister.)

Figure 12: Six categories of facial expression in the ASL-originated Stimuli Set.

4.1.2 Results and Comparison of the Studies

Two groups of ASL signers evaluate the ASL animations from the two stimuli sets, in each study, they viewed animations: (a) with facial expressions and (b) without facial expressions. A fully-factorial within-subjects design was used such that: (1) no participant saw the same story twice, (2) the order of presentation was randomized, and (3) each participant saw every story – in either version (a) or (b). ASL signers were recruited from ads posted on Deaf community websites in New York. All instructions and interactions were conducted in ASL by a signer (a professional interpreter). In (Huenerfauth, 2008; Huenerfauth et al., 2008), the authors discussed why it is important to recruit signers, and they list best practices to ensure that responses given by participants are as ASL-accurate as possible.

Twelve participants evaluated the **English-to-ASL** stimuli set: 8 participants used ASL since birth, 3 began using ASL prior to age 10 and attended a school using ASL, and 1 participant learned ASL at age 18. This final participant used ASL for over 22 years, attended a university with instruction in ASL, and uses ASL daily to communicate with a spouse. There were 7 men and 5 women of ages 21-46 (median age 32).

Sixteen participants valuated the **ASL-originated** stimuli set: 10 participants learned ASL prior to age 5, and 6 participants attended residential schools using ASL since early childhood. The remaining 10 participants had used ASL for over 9 years, learned ASL as adolescents, attended a university with classroom instruction in ASL, and used ASL daily to communicate with a significant other or family member. There were 11 men and 5 women of ages 20-41 (median age 31).

Figure 13 shows the results of the studies that compare English-to-ASL stimuli and ASL-originated stimuli, with the results of the "Emotion" category presented separately from the results from all other categories. Error bars indicate standard error of the mean; significant pairwise differences are marked with stars (ANOVA, $p<0.10$). Our goal is to identify "good" stimuli for use in studies evaluating ASL facial expression animations. Since a human animator has carefully produced the facial expressions for these studies, "good" stimuli should have a big difference in comprehension scores between the without-facial-expression and with-facial-expression versions. *It is important to note that the scores across studies can't be directly compared, since the sentences and questions may have been more difficult in one study.*

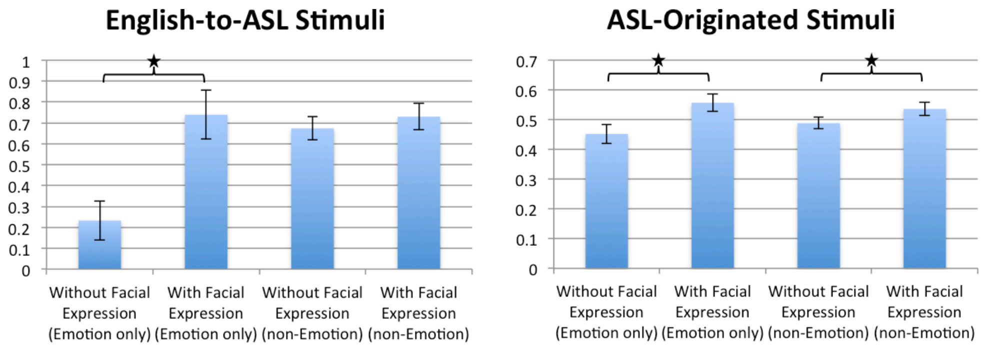


Figure 13: Comprehension question scores for both types of stimuli, showing results for animations with- and without-facial-expressions, with results for emotion and non-emotion categories.

For English-to-ASL stimuli, for the emotion category, adding facial expressions led to significantly higher comprehension scores. However, there was no benefit from adding facial expressions for the non-emotion categories, which didn't convey the subtle meaning differences that we had intended. Perhaps since the stimuli were first conceived as English stimuli with vocal prosody, something was "lost in translation" when the stimuli were converted into ASL animations with facial expressions.

For the ASL-originated stimuli, adding facial expressions led to significantly higher comprehension scores for both emotion and non-emotion categories. This is a desirable result because it indicates that the stimuli/questions allowed us to distinguish between animations with good or with bad facial expressions (in this case, no facial expressions at all). If we used these stimuli/questions in future studies, we could compare the performance of animations with facial expressions automatically synthesized by our software – to video recordings of signers, animations with facial expressions produced by a human animator, or animations without any facial expressions. Thus, we could track the performance of our facial-expression synthesis algorithms to guide our research.

4.2 Release of Experimental Stimuli and Questions[6]

After determining that an "ASL-Originated" stimuli design process produced sets of stimuli and questions that were more effective in studies evaluating sign language animations, we created a collection of stimuli for our future studies. These stimuli have been specifically engineered to enable us to evaluate the perception and understanding of facial expressions in ASL animations.

We have used these stimuli during the past several years of user studies on ASL facial expressions that convey grammatical syntax information (Kacorri et al., 2013a; Kacorri et al., 2013b; Kacorri et al., 2013c, Kacorri et al., 2014; Kacorri and Huenerfauth, 2014). This collection of scripted ASL multi-sentence single-signer passages and corresponding comprehension questions has enabled us to probe whether human participants watching these stimuli have understood the information that should have been conveyed specifically by the facial expressions. In (Huenerfauth and Kacorri, 2014) we shared this set of stimuli and questions with the research community to support research on non-manual linguistic phenomena.

[6] This section describes joint work with Professor Matt Huenerfauth (Huenerfauth and Kacorri, 2014).

4.2.1 Overview of the Collection

Our stimuli collection includes: 48 ASL passages performed by a (male) signer; 192 comprehension questions (4 questions for each passage, each question performed by 2 signers, male and female); a set of Likert-scale subject questions about the grammatical correctness, ease of understanding, and naturalness of movement of the passages; and a set of Likert-scale questions asking whether participants noticed specific categories of facial expressions. The collection consists of video recordings of the ASL signer, ASL transcriptions of each passage, English translation of the ASL passages and comprehension questions as plaintext files, and two sets of questionnaires with the Likert-scale questions. The English translations of the ASL stories includes both the indented meaning when the ASL facial expression is performed correctly and a second ambiguous meaning when the facial expression is not correctly perceived by the person viewing the story. (More details on the stimuli collection are available in Appendix A.)

Each stimulus focuses on a particular facial expression in one of the following categories listed below. Each is illustrated in Figure 14. The first five categories are described in Section 2.1, and the Emotional affect category is informally described below. Please consult ASL linguistics references for more detailed explanations, e.g., (Neidle et al., 2000).

- **Yes/No Questions, WH-Questions**, **RH-Questions**
- **Topic**
- **Negative**

- **Emotional affect**: These facial expressions are not linguistically governed, but they include several typical affective facial expressions that can indicate sadness, anger, frustration, etc. during a sentence.

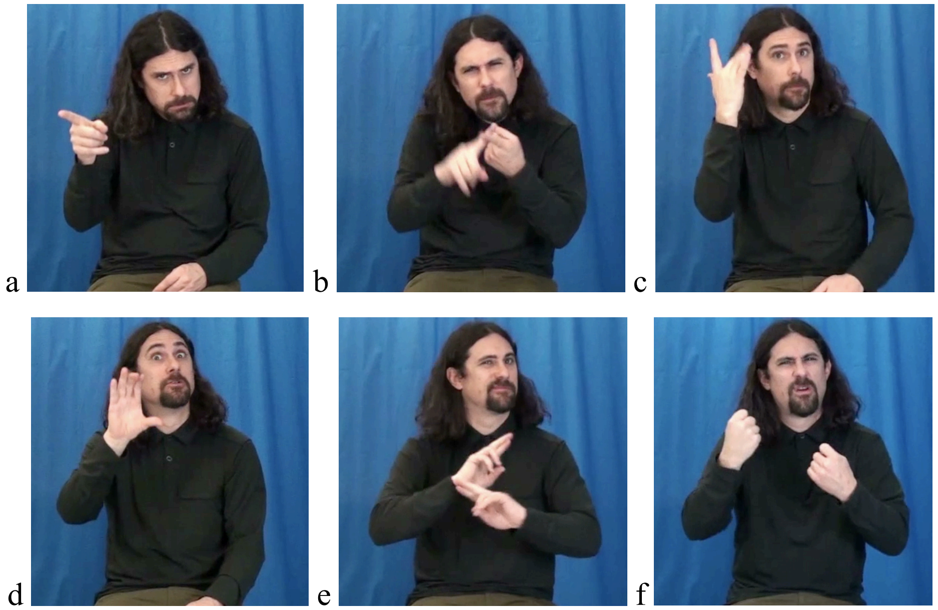


Figure 14: Still images taken from videos included in the stimuli collection, with each image illustrating a moment when a particular facial expressions is occurring: (a) YN-Question, (b) WH-Question, (c) RH-Question, (d) Topic, (e) Negative, and (f) Emotional Affect (an example of anger).

As discussed in Section 4.1, the value of this collection is that the stories and questions were carefully engineered so that the participant must perceive and understand the facial expression in order to answer the comprehension questions correctly. For each stimulus, if the manual portion of the performance were considered alone (without the facial expressions), then there would be an ambiguity or an alternative semantic interpretation possible for the stimulus. Our comprehension questions have been designed to detect when a participant has misunderstood the stimulus due to the facial expression not being successfully perceived or understood. Thus, these stimuli can be used to evaluate the quality of automatic animation-synthesis systems for generating animations of ASL with facial expressions. Table 4 provides a listing of the number of stimuli in the collection of each type.

Table 4: Collection Overview.

Facial expression type	**Number of stimuli** avg(#glosses/stimulus)	**Codenames of these stimuli in the collection**
WH-word Questions	9 stimuli (13)	W1, W2, W3, W4, W5, W6, W7, W8, W9
Yes/No Question	7 stimuli (9.29)	Y1, Y2, Y3, Y4, Y5, Y6, Y7
Topic	7 stimuli (10)	T1, T2, T3, T4, T5, T6, T7
Rhetorical Question	11 stimuli (11.82)	R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R11
Negative	6 stimuli (16.5)	N1, N2, N3, N4, N5, N6
Emotional Affect	8 stimuli (6.88)	E1, E2, E3, E4, E5, E6, E7, E8

4.2.2 Design of Stimuli and Comprehension Questions

As briefly mentioned in Section 4.1.1, to create the collection of ASL-Originated stimuli, prior to the design of the stimuli, an ASL signer was given 6 categories of facial expressions and was introduced to premise that the passage must be ambiguous in its meaning if the facial expression were not understood. The ASL signer invented, performed, and transcribed the ASL passages, and the passages were discussed and edited in collaboration with a team of other ASL signers at the laboratory. Next, the two ambiguous meanings were translated into English sentences. Consulting the ASL transcription and the two ambiguous English translations, a second ASL signer performed the ASL passages for the video recordings in our collection. Finally, linguistic researchers at our laboratory engineered the comprehension questions for each story such that they would receive different answers, depending on the perception and understanding of the facial expression. The collection includes a sample HTML form where the 4 comprehension questions are embedded in video format and the answers are collected on a 7-point Likert-scale from “definitely no” to “definitely yes.”

While the full collection of stimuli and questions is given in the Appendix A, this section explains a specific example of each category of stimuli to illustrate in detail how each stimulus can have alternative interpretations, if the facial expression were not correctly understood. In these examples the name “Charlie” refers to the animated character or the human signer performing the ASL passage. We also have included in the collection an introductory video recording were a signer describes the comprehension question task and introduces the avatar or the human signer as “Charlie”.

Topic Example: The following sentence is an example of a stimulus with a Topic facial expression (which should occur during the gloss “SWEET FOOD”): NEW RESTAURANT INCLUDE PASTA PIZZA SWEET FOOD MY SISTER COOK EXPERT. When the Topic face is perceived, then the stimulus has the approximate meaning: “The new restaurant has pasta and pizza. As for sweet foods (pastries), my sister is an expert chef.” We have intentionally designed the stimulus so that it is performed at a human conversational speed without any long pauses during the signing that would emphasize the sentence boundary before “SWEET.” This has been done so that the meaning of the stimulus is strongly affected by whether the viewer perceives the Topic facial expression. When the Topic face is not perceived, then the sentence boundary may be less clear (especially when the sentence is performed by an animated avatar that typically lacks the subtle acceleration and timing of a human signer). In such a case, the viewer may interpret “SWEET FOOD” as being the third item in the list of foods available at the restaurant; thereby the stimulus has the meaning: “The new restaurant has pasta, pizza, and sweet foods (pastries). My sister is an expert chef.” One of the comprehension questions for this stimulus is: Does the new restaurant have sweet foods?

The answer depends on whether the Topic facial expression was perceived and understood.

WH-Word Question Example: The following sentence is an example of a stimulus with a WH-Question facial expression (which should occur during the glosses “HER BIRTHDAY PARTY WHEN”): THAT MARY HER BIRTHDAY PARTY WHEN MARY DRUNK. When the WH-Question face is perceived, then the stimulus has the approximate meaning: “When is Mary's birthday party? Mary is drunk.” When the WH-Question face is not perceived, then it may be less clear to the viewer where the sentence boundary is located. In such a case, the viewer may interpret “WHEN MARY DRUNK” as a question (albeit in English-like word order); thereby the stimulus would have the meaning: “It is Mary's birthday party. When did Mary got drunk?” One of the comprehension questions for this stimulus is: Does Charlie know when the party is? (The signer appearing in the video is introduces as “Charlie” at the beginning of the study.) The participant is more likely to answer “no” to this question if the WH-Question facial expression was correctly perceived.

Rhetorical Question Example: The following sentence is an example of a stimulus with a RH-Question facial expression (which should occur during the glosses “WHY”): ALEX NOW GO-GO PARTIES WHY FINISH DIVORCE. When the RH-Question face is perceived, then the stimulus has the approximate meaning: “Alex is now often going to parties because he is divorced.” When the RH-Question face is not perceived, then the sentence boundary may be less clear. In such a case, the viewer may interpret “WHY FINISH DIVORCE” as a question; thereby the stimulus has the meaning: “Alex is now often going to parties. Why did he get divorced?” One of the comprehension questions for this stimulus is: Does Charlie know why Alex started going to parties? The answer depends on whether the RH-Question facial expression was perceived and understood.

Yes/No Question Example: The following sentence is an example of a stimulus with a Yes/No Question facial expression (which should occur during the glosses “ALL FOOD CHEAP POINT”): BOB'S DINER THAT YOUR SISTER HER FAVORITE RESTAURANT ALL FOOD CHEAP POINT. When the YN-Question face is perceived, then the stimulus has the approximate meaning: “Bob’s Diner is your sister’s favorite restaurant. Is all the food cheap?” When the YN-Question face is not perceived, then the final sentence could appear to be a declarative statement. Thus, the stimulus has the meaning: “Bob’s Diner is your sister’s favorite restaurant. All the food is cheap.” One of the comprehension questions for this stimulus is: Does Charlie know if the restaurant is expensive? If the YN-Question facial expression was correctly perceived and understood, then the participant is more likely to answer no to this question.

Negative Example: The following sentence is an example of a stimulus with a Negative facial expression (which should occur during the glosses “HAVE SCIENCE CLASS”): ALEX TEND TAKE-UP MATH CLASS. NOW SEMESTER, SCHOOL HAVE SCIENCE CLASS. ALEX TAKE-UP TWO CLASS.” When the Negative face is perceived, then the stimulus has the approximate meaning: “Alex usually takes math classes. This semester, the school doesn't have any science classes. Alex is taking two classes. ” When the Negative face is not perceived, then the meaning of the middle sentence is inverted: “This semester, the school has science classes.” One of the comprehension questions for this stimulus is: Does the school have science classes this semester? The answer depends on whether the Negative facial expression was perceived and understood.

Emotional Affect Example: The following sentence is an example of a stimulus with an emotional affect facial expression (this example includes an angry facial expression during the entire sentence): LAST FRIDAY, MY BROTHER TAKE MY CAR. DRIVE SCHOOL. When the emotional affect facial expression is perceived, then the stimulus has the approximate meaning: "Last Friday, my brother took my car to drive to school." (The sentence has the subtext that the signer is upset about this.) When the emotional affect face is not perceived, then this subtext is not conveyed. One of the comprehension questions for this stimulus is: Is Charlie angry at his brother? The answer depends on whether the emotional facial expression was perceived and understood.

4.2.3 Likert-scale Questions

In addition to the four comprehension questions that are designed specifically for each stimulus, this collection also includes a set of Likert scale questions that can be used to measure participants' self-reported evaluation of each. The set of Likert scale questions is identical for all of the stimuli, and it includes three subjective evaluation questions and four questions measuring whether participants' noticed a particular facial expression:

"Good ASL grammar?": A subjective evaluation question of how grammatically correct was the presented signing with answers on a 1-to-10 Likert scale where 1 indicates bad and 10 perfect.

"Easy to understand?": A subjective evaluation question on comprehensibility of the signed message with answers on a 1-to-10 scale where 1 indicates confusing and 10 clear.

"Natural?": A subjective evaluation question on how naturally moving the signer appeared with answers on a 1-to-10 scale where 1 indicates that the signer moves like a robot and 10 that the signer moves like a person.

"Did you notice a … facial expression?": Four questions in relation to how much participants noticed an emotional, negative, interrogative, or topic facial expression during the story with answers on a 1-to-10 scale from "yes" to "no".

The collection includes an HTML questionnaire with these Likert-scale questions and the options for the answers as radio buttons (shown in the Appendix A).

4.2.4 Availability of the Collection

As with prior ASL corpora resources released by our laboratory (Lu and Huenerfauth, 2009; Lu and Huenerfauth, 2012a), this stimuli collection is available for use by other sign language animation researchers, at http://latlab.ist.rit.edu/2014assets. We have invited members of the research community to provide feedback to us about the stimuli in this collection, and we welcome recommendations of additional stimuli designs or edits that would enhance the collection (which we would look forward to incorporating into a future release of this resource). Ultimately, the field of sign language animation synthesis may benefit from the community identifying a standard set of evaluation stimuli and questions for system evaluation, to better enable comparison of systems and progress in the field.

Chapter 5 Effect of Videos as Upper Baseline and Questions[7]

This chapter focuses on the methodological issue: how the presentation of upper baseline for comparison (either a high-quality animation or a video of a human signer) affects the responses recorded in a study. As discussed above, to evaluate our approaches for synthesizing animations of sign language, we typically compare an animation that has been synthesized using our current model to some other animation (e.g., synthesized using an earlier model or using some simplistic "lower baseline" technique). ASL signers answer subjective questions about the quality of our animations, and they answer comprehension questions about the animations' information content.

A challenge when interpreting the results of comprehension questions is that the score is based on factors beyond the animation itself, e.g., a question's difficulty, a participant's memory skill, etc. To make it easier to interpret the results from comprehension questions, we have generally added an "upper baseline" (a third type of animation shown during the study for comparison). A good upper baseline should represent an "ideal" output of the system, and it may consist of a high-quality computer animation or a video recording of a human signer (performing identical sentences to the virtual human in the animations). As discussed in Chapter 3, research groups have differed in their selection of an upper baseline for evaluations: some have used videos of humans, and some have used computer animations. There are trade-offs for either choice, as discussed below.

[7] This chapter describes joint work with Pengfei Lu, a Ph.D. student in Computer Science at CUNY, and Professor Matt Huenerfauth (Lu and Kacorri, 2012; Kacorri et al., 2013b).

As an upper baseline in prior studies in our lab, researchers investigating data-driven synthesis of manual movements used a computer animation of a virtual character (one that is visually identical in appearance to the virtual human in their model-synthesized animations). An ASL signer who is a skilled animator carefully controlled the movements of the character to produce the most fluent/natural animation possible – performing identical sentences to the virtual human in their model-synthesized animations. The rationale for this choice was that our lab does not investigate issues related to the photorealism of virtual human animations, but instead, it investigates models of the movements of the virtual human character. Thus, an animator-controlled high-quality animation represents an "ideal" output of what our lab's software could achieve. Further, a potential problem with including videos of real humans in the experiments is the concern that participants would focus on the differences in appearance between the human and the virtual human – and thereby they might attend less to the movement subtleties of the virtual human character, which was their research focus.

On the other hand, an intuitive upper baseline for an experiment would be a video of a human ASL signer. Our study participants are used to seeing humans performing ASL all the time; they are less familiar with seeing computer animations of ASL. Further, a video of an ASL signer performing ASL would likely have higher fluency/clarity than any animation; so, it could be considered a truer "ideal." Another advantage is that non-experts can interpret the results from the experiment; since a video of a human is more familiar than an animation of a virtual human signing, it is easier to understand the results of the experiment, relative to a human video upper baseline. A downside of a human video upper baseline is that it may be an impossibly high ideal – the state of the art of sign language computer animation may be decades away from producing something with similar quality to a video of an actual human. So, a video upper baseline could yield scores that are so "off the scale" higher that they could make it difficult to obtain meaningful evaluation scores for the other animations in the study.

Despite these various trade-offs and despite prior research groups making different choices as to their upper baselines (details in Chapter 3), we have not found any prior methodological research on the effect of selecting each of these different types of upper baseline. While it is intuitively plausible that a computer animation being evaluated may receive different evaluation scores in an experiment – depending on whether it is being compared to another animation or compared to a video of a human, the specific empirical effect of selecting an upper baseline has not been quantified. This means that currently there is no reliable way to compare the empirical results across evaluation studies conducted by different research groups (who have used different upper baselines), and there is no guidance for future researchers as to the best approach to use when designing their evaluation studies. The goal of this chapter is to fill this gap in the methodological literature and to provide a useful foundation for future empirical research in the growing field of sign language computer animation synthesis.

This chapter begins with a set of research hypotheses that relate to how the selection of a video or animation upper baseline affects the results of a user study (Section 5.1). Next, Section 5.2 describes our first experimental study, in which we replicate a prior study (Huenerfauth and Lu, 2010) and replace the upper baseline in that study with a video of a human signer. Section 5.3 describes a new pair of experiments, focused on facial expressions during sign language, with a similar structure: performed once with an animation upper baseline and then performed again with a video upper baseline. Finally, Section 5.4 presents a third set of experiments that explores a related issue: whether presenting the comprehension questions in an experiment in the form of animation or video affects the evaluation results.

5.1 Research Methodology and Hypotheses

To compare the results of prior studies that used different upper baselines (Section 3.2), we need to quantify how the upper baseline affects the evaluation scores collected. To do so, we need to conduct an identical study in two ways: (1) once using videos of human signers as an upper baseline and (2) once using computer animations as an upper baseline. If the other animations shown in the study (aside from the upper baseline) remain constant, then any differences in their evaluation scores could be attributed to difference in the upper baseline used. In this manner, we can examine several research questions:

- Do video upper baselines receive higher comprehension question scores than animation upper baselines do?
- If a study uses a video upper baseline (instead of an animation upper baseline), then are the comprehension scores for the other animations in the study affected?
- Do video upper baselines receive higher Likert-scale subjective evaluation scores than animation upper baselines do?
- If a study uses a video upper baseline (instead of an animation upper baseline), then are the Likert-scale subjective evaluation scores for the other animations affected?

It is important to note that, for these research questions, the scientific aim of any individual study (determining if the mathematical/linguistic model under consideration produces good ASL animations) is not important: Instead, we are only focused on whether changing the upper baseline in the experiment causes measurable differences in the evaluation scores for the upper baseline and for the other animations being evaluated in that experiment.

In order to formulate some hypotheses in regard to these questions, we considered a pair of prior experiments at our lab that were nearly (but not exactly) identical: with one experiment using an animation upper baseline and the other using a video upper baseline. In (Lu and Huenerfauth, 2010), an experiment was conducted to evaluate the quality of some computer animations of sign language, with an animation produced by a human animator as an upper baseline. In (Lu and Huenerfauth, 2014), a similar (but not identical) set of computer animations of sign language were evaluated, but a video of a human signer was used as an upper baseline. In both studies, ASL signers who saw the animations/videos answered comprehension questions and Likert-scale subjective evaluation questions. Unfortunately, this prior pair of studies was not a perfect test of our research questions: The script of the ASL stories in the two studies was not identical (so the stories might have been harder in one of the studies). Further, the human in the videos used as an upper baseline in (Lu and Huenerfauth, 2014) wore some motion-capture equipment; so, he may have been harder to understand. Regardless, by considering how the scores in these (approximately) identical studies differed, we gain insight into the effect of different upper baseline – and can formulate some hypotheses.

Changing the upper baseline did not produce a difference in the comprehension question scores for the other stimuli in the study (the motion-capture-based animations), which had similar scores in both studies. For the Likert-scale subjective scores (1-to-10 scales for naturalness of movement, perception of understandability, and grammatical correctness of the animations), in the later study (with the video upper baseline), the other animations received lower scores than they had in the prior study. We speculate that seeing a video of a human as one of the stimuli led participants to assign lower Likert-scale subjective scores to the animations (which looked worse by comparison to a video of a real human). Based on these prior studies, to investigate **RQ2** we hypothesize the following:

H1: A human video upper baseline will receive higher comprehension question scores than an animated-character upper baseline produced by a human animator.

H2: The upper baseline used (human video or animated character) will not affect the comprehension questions accuracy scores for the other stimuli shown in the study.

H3: A human video upper baseline will receive higher Likert-scale subjective scores than an animated-character upper baseline.

H4: Using a human video upper baseline will depress the subjective Likert-scale scores that participants assign to the other stimuli in the study.

As mentioned at the end of Section 3.3, videos of human signers could appear during a user study in other capacities, aside from appearing as an upper baseline. Specifically, videos of human signers might be displayed to study participants as the comprehension questions that participants are asked after viewing each of the sign language animations. Displaying videos of humans asking questions in ASL could also have an effect on participants' subjective ratings of the animations in the study. Therefore, Section 5.4 will evaluate the following two additional hypotheses, which investigate **RQ3,** and thus, focus on a comparison between presenting questions as videos of human signers or as animations of sign language:

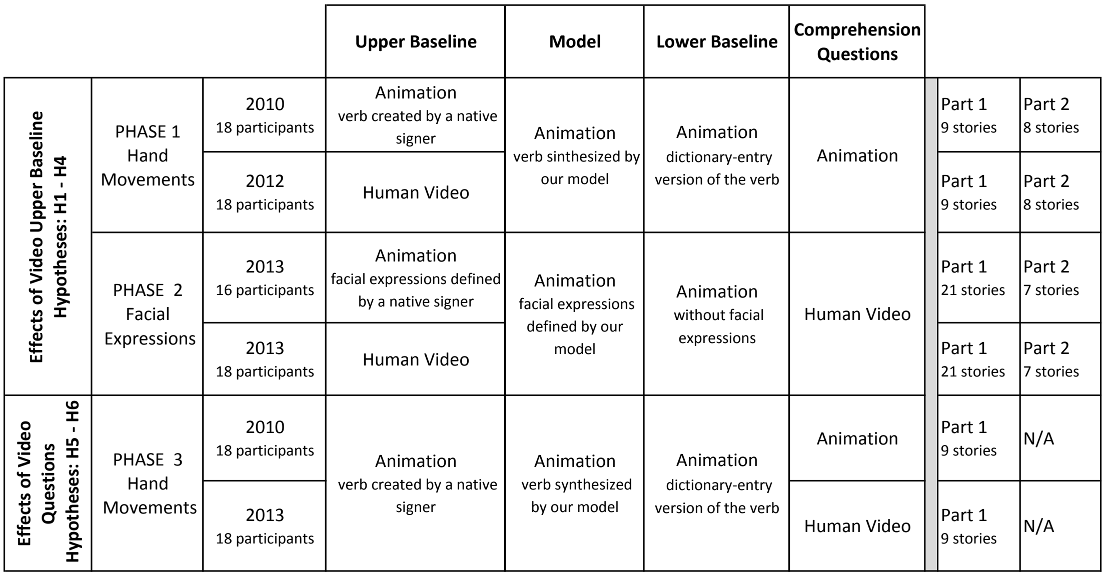
H5: Displaying comprehension questions as videos of a human signer or as a high-quality animation will not affect the comprehension questions accuracy scores.

H6: Displaying comprehension questions as videos of a human signer or as high-quality animations will not affect the subjective Likert-scale scores that participants assign to the animations in the study.

5.1.1 Three Phases of This Research

Our research methodology consists of three phases, which are summarized in Table 3.

Table 5: Summary of Three Phases of Experiment.



			Upper Baseline	**Model**	**Lower Baseline**	**Comprehension Questions**		
Effects of Video Upper Baseline Hypotheses: H1 - H4	PHASE 1 Hand Movements	2010 18 participants	Animation verb created by a native signer	Animation verb sinthesized by our model	Animation dictionary-entry version of the verb	Animation	Part 1 9 stories	Part 2 8 stories
		2012 18 participants	Human Video				Part 1 9 stories	Part 2 8 stories
	PHASE 2 Facial Expressions	2013 16 participants	Animation facial expressions defined by a native signer	Animation facial expressions defined by our model	Animation without facial expressions	Human Video	Part 1 21 stories	Part 2 7 stories
		2013 18 participants	Human Video				Part 1 21 stories	Part 2 7 stories
Effects of Video Questions Hypotheses: H5 - H6	PHASE 3 Hand Movements	2010 18 participants	Animation verb created by a native signer	Animation verb synthesized by our model	Animation dictionary-entry version of the verb	Animation	Part 1 9 stories	N/A
		2013 18 participants				Human Video	Part 1 9 stories	N/A

In Phase 1, to evaluate hypotheses H1-H4, we conducted an identical pair of experiments, with the only difference being the upper baseline used. Participants evaluated animations of short stories that contained ASL verbs with complex hand movements. Section 5.2 portrays the challenges in recording videos of a human performing an identical script of signs as an animated character. The clearest results were for hypotheses H3 and H4. The video upper baseline received higher Likert-scale subjective scores (H3 supported), and it led to lower Likert-scale subjective scores for the other stimuli during the side-by-side comparison part of the study (H4 supported).

In Phase 2, an additional pair of studies was conducted: one with a video upper baseline and one with an animation upper baseline. The focus of this thesis is the synthesis of linguistically meaningful facial expressions for sign language animations; so, for Phase 2, we used animations of ASL sentences with various grammatical and emotional facial expressions. The studies in Phase 2 also contained a larger number of comprehension questions, to enable us to better evaluate hypotheses H1 and H2, which were not adequately addressed in Phase 1. The video upper baseline received higher comprehension scores than the animation upper baseline (H1 supported) and led to a small increase in the comprehension scores for the other stimuli (H2 not supported).

In Phase 3, to evaluate hypotheses H5 and H6, we conducted a final pair of studies: one with comprehension questions presented as human videos and one with comprehension questions presented as high-quality ASL animations. Section 5.4 describes how the choice of video or high-quality animation had no effect on the comprehension (H5 supported) or Likert-scale scores we collected (H6 supported).

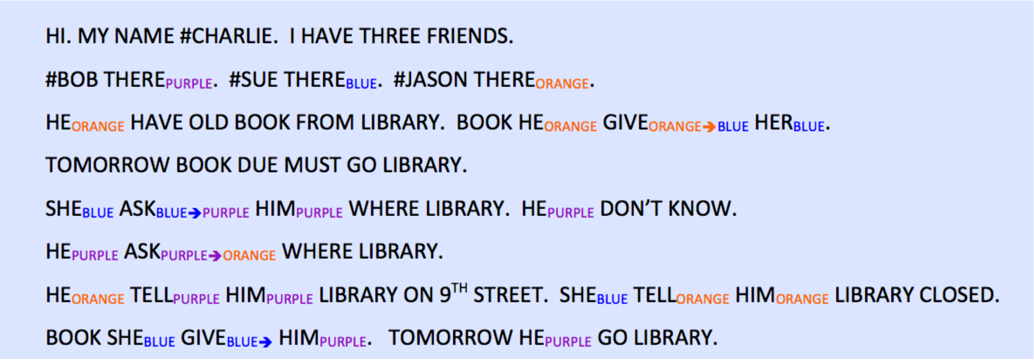
5.2 Phase 1: Animation vs. Video Upper Baselines

To clearly evaluate hypotheses H1-H4, we needed to compare the results of two experiments that were identical, aside from the upper baseline used. Since researchers in our lab had previously conducted a study in which computer animations were used as an upper baseline (Huenerfauth and Lu, 2010), we decided to replicate that study. We replaced the upper baseline with a video of human signer performing identical ASL stories as the animated character. This section describes the challenges we faced when producing a video of a human that performed identical signs to our animated character upper baseline, and it presents the results of our 2012 replication of the original 2010 study.

5.2.1 Design of Evaluation Studies for Phase 1

In (Huenerfauth and Lu, 2010), a model was evaluated for synthesizing the movements of "inflected" ASL verb signs whose movements depend on locations in the space around the signer where the verb's subject and object have been previously set up. In order to evaluate the understandability of animations in which the verbs were produced using the new model, three versions of animations were compared: (1) a lower baseline consisting of the simple dictionary-entry versions of the verb signs (where the hand movement doesn't indicate subject/object), (2) a middle case consisting of animations of the verbs synthesized by their model, and (3) an upper baseline consisting of animations of inflected versions of each verb produced by a human animator, who was an ASL signer.

The experimental study consisted of two parts: In part 1, participants were asked to view 9 animations of a virtual human character telling a short story in ASL. Each story included instances of the inflected verbs. A fully-factorial design was used such that: (1) no participant saw the same story twice, (2) order of presentation was randomized, and (3) each participant saw 3 animations of each version: i) lower baseline, ii) model, or iii) upper baseline. Figure 15 shows a story transcript, the English translation for this transcript is “Hi, my name is Charlie. I have three friends. Bob, Sue and Jason. Jason has an old book from the library; he gives the book to Sue. The book is due tomorrow and it must go to the library. Sue asks Bob where the library is. Bob doesn't know. Bob asks Jason where the library is. Jason tells Bob the library is on 9th street. Sue tells Jason the library is closed. She gives the book to Bob. Tomorrow Bob will go to the library.” Colors indicate locations around the signer where the verb’s subject/object are located. After watching each story animation (of one of three types: lower baseline, model-synthesized, or upper baseline) one time, participants answered 4 multiple-choice comprehension questions. Questions focused on whether they understood and remembered the subject and object of each verb. Participants also responded to three 1-to-10 Likert-scale questions about how grammatically correct, easy to understand, or naturally moving the animation appeared.



HI. MY NAME #CHARLIE. I HAVE THREE FRIENDS.

#BOB THERE$_{\text{PURPLE}}$. #SUE THERE$_{\text{BLUE}}$. #JASON THERE$_{\text{ORANGE}}$.

HE$_{\text{ORANGE}}$ HAVE OLD BOOK FROM LIBRARY. BOOK HE$_{\text{ORANGE}}$ GIVE$_{\text{ORANGE}\rightarrow\text{BLUE}}$ HER$_{\text{BLUE}}$.

TOMORROW BOOK DUE MUST GO LIBRARY.

SHE$_{\text{BLUE}}$ ASK$_{\text{BLUE}\rightarrow\text{PURPLE}}$ HIM$_{\text{PURPLE}}$ WHERE LIBRARY. HE$_{\text{PURPLE}}$ DON’T KNOW.

HE$_{\text{PURPLE}}$ ASK$_{\text{PURPLE}\rightarrow\text{ORANGE}}$ WHERE LIBRARY.

HE$_{\text{ORANGE}}$ TELL$_{\text{PURPLE}}$ HIM$_{\text{PURPLE}}$ LIBRARY ON 9$^{\text{TH}}$ STREET. SHE$_{\text{BLUE}}$ TELL$_{\text{ORANGE}}$ HIM$_{\text{ORANGE}}$ LIBRARY CLOSED.

BOOK SHE$_{\text{BLUE}}$ GIVE$_{\text{BLUE}\rightarrow}$ HIM$_{\text{PURPLE}}$. TOMORROW HE$_{\text{PURPLE}}$ GO LIBRARY.

Figure 15: Example script for a story shown in the study.

Part 2 of the study involved a side-by-side comparison methodology, as described in detail in (Lu and Huenerfauth, 2012; Huenerfauth et al., 2008): participants viewed three versions of an animation of a single ASL story side-by-side on one screen, as depicted in Figure 16(a). A total of 8 stories (three versions of each) were viewed by each participant. The sentences shown side-by-side were identical, except for the version of the verb that appeared in each, which was either: the dictionary-entry version of the verb animation (the "lower baseline"), the verb animation synthesized by their model, or the verb carefully created by an ASL signer using animation software (Vcom3D, 2014) ("upper baseline"). The participants could re-play each animation as many times as they wished. Participants were asked to focus on the verb and respond to a single 1-to-10 Likert-scale question about the quality of the sentence animation.

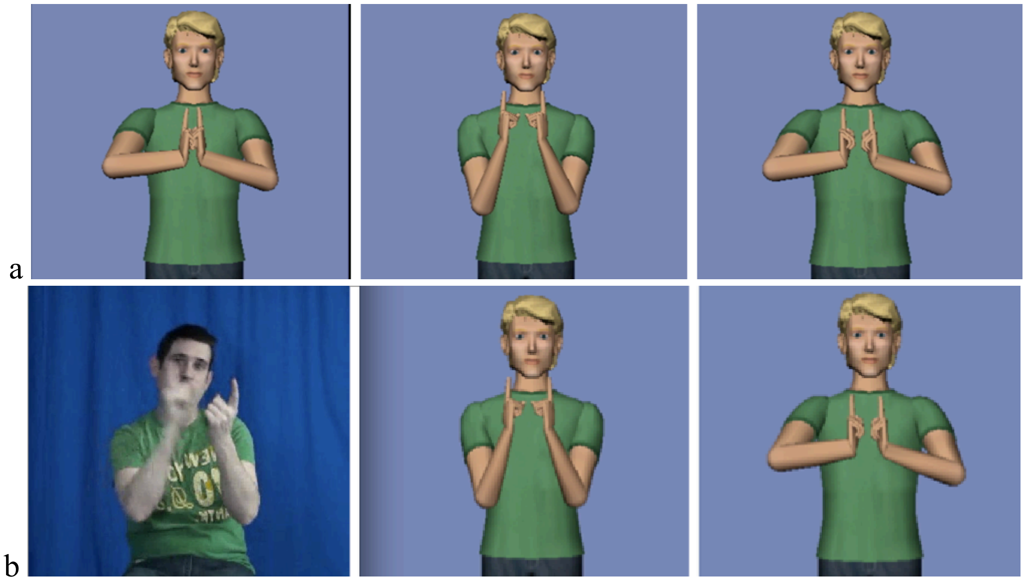


Figure 16: Screenshots from the side-by-side comparison, as seen by participants in (a) 2010 or (b) 2012.

During our replication of the study, in 2012, we replaced the upper baseline animations with videos of a human signer. The top row in Figure 16(a) shows an example of what the participants saw in the part 2 (side-by-side comparison) in 2010 and the lower row in Figure 16(b) shows what they saw in 2012. All the other animations and their sequencing in this pair of studies were identical.

5.2.2 Recording the Human Video Upper Baseline for Phase 1

To produce the human video upper baseline, we recorded an ASL signer in our studio. For part 1, we needed to record a human performing the 9 short stories (with inflected versions of the verbs), and for part 2, we recorded a human performing the 12 sentences (with inflected versions of the verbs). Producing a video recording of a human that "matched" the animations being shown in the study was challenging. We wanted to "control" as many of the variables of the ASL performance between the upper baseline and the animation under primary evaluation (our model-synthesized animation) as possible, so that it would serve as an effective upper baseline for comparison. To match the background color, the human sat in front of a blue curtain. The human signer also wore a green t-shirt on the day of the recording, which was similar to the virtual human. To maintain the same viewing perspective, we placed the camcorder facing the signer at his head height, which matched the "virtual camera position" in the ASL animations. We cropped and resized the video files to match the height/width of the 2010 upper baseline animations – and to approximate the same placement of a human in a the video frame as how the virtual human character had appeared in the animation in 2010. The frame rate and the resolution of the video were identical to the animation from 2010.

To ensure that the human signer performed fluent ASL signing, all of the instructions and interactions for the recording session were conducted in ASL by another signer sitting behind the camcorder. We needed the human signer to perform the same "script" of signs as the other (animation) stimuli shown in the study; so, we placed a large monitor in front of the signer (just below the camera) to display the story scripts (a script example is shown in Figure 15). The signer had time to memorize and practice each of the scripts prior to the recording session – so that he would not need to read the script while signing. Because the stories were a bit complicated (an average of 55 signs in length, included 3-5 main characters set up at various locations in the signing space, with 3-5 inflected verbs per story), the signer had to practice in order to perform each story fluently. Unfortunately, due to the complexity of the stories and the need for accuracy, the signer found it very difficult to avoid glancing at the script occasionally during the performance. While we asked the signer to attempt to memorize the script and not look at the script during the recording process, several of our recordings contained infelicitous moments when the signer's eyes glance at the monitor displaying the story script.

On one hand, we wanted to control as many variables as possible so that they were held constant between our upper-baseline video and our animation being evaluated; on the other hand, we wanted to record a natural, fluent version of the sentences from the human signer. If the human's performance looks artificial, then it would not be an ideal of fluency and naturalness. Since ASL has no standard written form, and multiple signs can have similar meaning and use the same notation used in our notation, we had to explain our notation scheme to the signer and occasionally demonstrate which ASL sign was indicated by a particular word in the script. The script notation does not capture all of the subtleties of performance that are part of ASL; it is merely a loose sketch of what must be signed. Although there was a script, the entire ASL performance was still underspecified, leaving room for the human, who was a Deaf ASL interpreter, to fill in the remaining elements of the performance based on his linguistic expertise in ASL.

While we gave the signer some “artistic freedom” in the performance of the stories, for the sake of naturalness and fluency, we did have to ask the signer to control the speed of his signing and some aspects of facial expressions, torso movement, head movement, etc., that could not be supported by the current animation tool and were not included in the animations being evaluated. To produce the same time duration in the videos as the upper baseline animations that had been used in 2010, we asked the signer to practice several times before the recording, and we used a stopwatch to measure how many seconds he took for each story during the practice and recording. After making several recordings of each story, we picked the one video recording of the story with the closest time duration to the upper baseline animation from 2010. We also asked the signer not to add too many theatrical embellishments, e.g., additional emotional facial expressions, which hadn’t appeared on our virtual human character’s face. This coaching and scripting process that was required in order to produce a good-quality human video upper baseline was surprisingly time-consuming, and it often felt like a delicate “balancing act” between guiding/controlling the human’s performance while still allowing freedom in the performance so that the result would be natural and fluent. Even with this complex process outlined above, our resulting videos may not have been completely natural and fluent. For instance, some of the participants noticed problems in the human video, e.g., commenting that the “person signs well but need[s] little [more] facial expression.”

One aspect of ASL that is especially difficult to capture in a script notation is specifying where in the space around a signer's body someone should point to refer to entities under discussion. Because the stories in this experiment contained many characters or objects that were assigned locations in space, with some verb signs in the stories changing their hand movements based on these locations, we needed our human performer to point to particular locations in space during the story (that identically matched the locations where our virtual human character points in the animations). Figure 17 illustrates how we set up small colored paper squares around the studio (with colors that matched the script in Figure 15) to guide the human where to point or where to aim the motion path of inflecting verb signs during the recording session. At 30-degree intervals around the signer, the squares were arranged in an arc in the following order (from the signer's left to the signer's right): purple, white, red, green, blue, orange, and yellow.

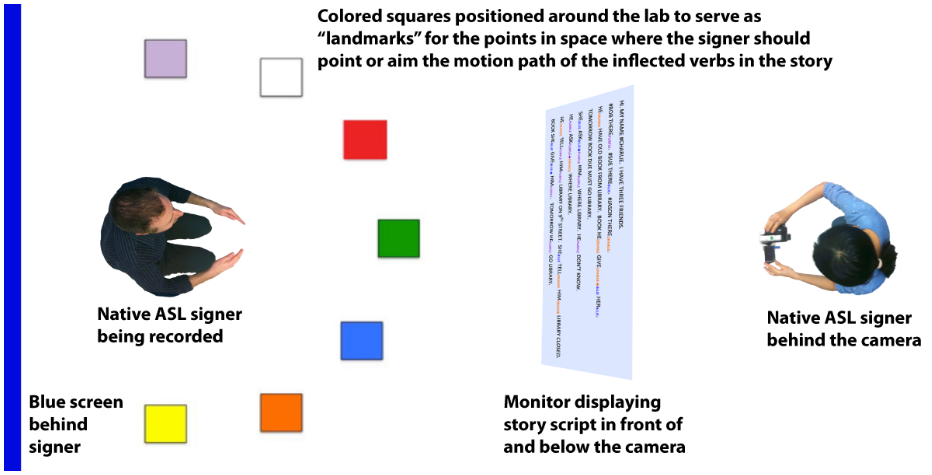

Colored squares positioned around the lab to serve as "landmarks" for the points in space where the signer should point or aim the motion path of the inflected verbs in the story
Native ASL signer being recorded
Native ASL signer behind the camera
Blue screen behind signer
Monitor displaying story script in front of and below the camera

Figure 17: Diagram of an overhead view of recording studio.

5.2.3 Data Collection and Results for Phase 1

We note that, given the goal of this study, it is not possible to test our hypotheses with a fully within-subjects design. Once a participant has seen an upper baseline video of a human signer, then they cannot participate in the animation upper baseline portion of the study. There is a carry-over effect: the participants cannot “un-see” or forget the video upper baseline once it has been seen. Further, there may be a practice effect when viewing animations of ASL and answer comprehension questions, and since it would not be possible to counterbalance the order in which participants participate in each study, we could not control for this order effect. Fortunately, we were able to design the experiments to control some variability due to individual differences in participants’ skill. Identical recruitment procedures were followed in both 2010 and 2012, and very similar demographics were observed in the participants in both studies.

To ensure that responses given by participants are as linguistically accurate as possible, our lab screens participants to ensure that they are native or fluent ASL signers and controls the experimental environment so that it is ASL-focused; details of these methods appear in (Huenerfauth et al., 2008). An ASL signer conducted all of the interactions for our studies. Ads were posted on New York City Deaf community websites asking potential participants if they had grown up using ASL at home or had attended an ASL-based school as a young child. The 2010 study included 18 participants: 12 learned ASL prior to age 5, and 4 attended residential schools using ASL since early childhood. The remaining 2 participants used ASL for over 15 years, learned ASL as adolescents, attended a university with instruction in ASL, and used ASL daily to communicate with a family member. There were 12 men and 6 women of ages 20-56 (average age 30.5). The 2012 study included 18 participants: 16 learned ASL prior to age 5, and 10 attended residential schools using ASL since early childhood. The remaining 2 participants used ASL for over 13 years, learned ASL as adolescents, attended a university with instruction in ASL, and used ASL on a daily basis to communicate with a family member. There were 12 men and 6 women of ages 22-49 (average age 32.8).

The results collected include the comprehension-question and Likert-scale scores in part 1 of the studies (after a participant viewed a story one time) and the Likert-scale scores collected in part 2 of the studies (in which participants assigned a score to each of the three sentences which they viewed side-by-side). In Figure 18Figure 19Figure 20, which display the results, the thin error bars display the standard error of the mean. Animator10 and Video12 were the upper baselines, Lower10 and Lower12 were the lower baselines with the dictionary-entry version of the verbs, and Model10 and Model12 were the versions of the animations produced using our lab's verb model. It is important to note that Lower10 and Lower12 were identical stimuli; the only difference was that the evaluation scores were collected in either the 2010 or 2012 study – likewise for Model10 and Model12.

To evaluate some of our hypotheses, we needed to consider the union of the responses for the Model and Lower animations in each study; these are displayed as "Model+Lower10" and "Model+Lower12" in Figure 18Figure 19Figure 20. Thus, "Model+Lower10" includes all of the data for Lower10 and Model10 combined. For the sake of clarity, we have included two graphs in each figure so that we would never display "Model+Lower" in the same graph as "Model" or "Lower" (since the latter is the combination of the data of the former two). In each figure, the graph on the left includes data for the upper baselines and for the "Model+Lower" data, and the graph on the right includes the individual results for Lower, Model, and the upper baselines.

One-way ANOVAs were used for comprehension-question data to check for statistical significance, and Kruskal-Wallis tests, for Likert-scale scores (because the Likert-scale data was not normally distributed). Statistical significance ($p<0.05$) for any of our planned comparisons has been marked with a star in Figure 18Figure 19Figure 20. The following comparisons were planned and conducted: (1) all three values from 2010, (2) all three values from 2012, (3) Video12 and Animator10, (4) Model12 and Model10, (5) Lower12 and Lower10, (6) Model+Lower10 and Model+Lower12, (7) Animator10 and Model+Lower10, and (8) Video12 and Model+Lower12. The reader should note that comparisons (1), (2), (7), and (8) are not needed to evaluate the specific hypotheses in this chapter. A researcher evaluating the quality of an animation relative to upper and lower baselines would traditionally perform these comparisons, and so we have presented them here for the benefit of future researchers who want to compare their results to ours.

Since H2, H5, and H6 are null hypotheses, it is appropriate to conduct "equivalence testing" to determine if pairs of values are indeed statistically equivalent, and we have therefore followed the two one-sided test (TOST) procedure (Schuirmann, 1987), which consists of: (1) selecting an equivalence margin *theta*, (2) calculating appropriate confidence intervals from the observed data, and (3) determining whether the entire confidence interval falls within the interval (-*theta*, +*theta*). If it falls within this interval, then the two values can be deemed equivalent. We've noticed that in our prior work, when we found significant differences between groups of sign language animation, we've generally seen differences of at least 1.5 Likert-scale units or 15% comprehension-question accuracy scores (e.g. Lu and Huenerfauth, 2014). Thus, we've selected equivalence margin intervals of (-1.5, +1.5) for Likert-scale scores and (-0.15, +0.15) for comprehension scores. Having selected an alpha-value of 0.05, then according to the TOST procedure for a two-sided analysis, we use a 90% confidence interval. Equivalence testing has been performed for the following pairs of values: (a) Video12 and Animator10, (b) Model12 and Model10, (c) Lower12 and Lower10, and (d) Model+Lower10 and Model+Lower12. Confidence intervals were determined using t-tests (for comprehension-question data) or Mann-Whitney U-tests (for Likert-scale data).

The data analysis and creation of the graphs for Phase 2 and Phase 3 have been conducted in an analogous manner for those studies, and therefore the above details are not repeated again in Sections 5.3 and 5.4 of this chapter.

Figure 18 illustrates the comprehension-question accuracy scores from the 2010 and 2012 studies. Hypothesis H1 would predict that Video12 would have significantly higher scores than Animator10, but this was not supported by the data. This was an interesting result: our videos of a human signer did not achieve higher comprehension scores than the upper baseline animations we produced in 2010 of a virtual human with the verbs carefully planned by a human animator. We speculate that our challenges in recording the human video may have led to some understandability problems in Video12 stimuli, or this may indicate that our upper baseline animations from 2010 were of good quality.

Hypothesis H2 would predict that the comprehension scores for the Model and Lower stimuli would be unaffected by changing the upper baseline from an animation in 2010 to a video in 2012. The following confidence intervals were calculated for TOST equivalence testing: (-0.154, 0.014) for Model+Lower10 vs. Model+Lower12, (-0.160, 0.067) for Model10 vs. Model12, and (-0.216, 0.031) for Lower10 vs. Lower12. Given that these intervals are not entirely within our equivalence margin interval of (-0.15, +0.15), we cannot determine whether pairs are equivalent. Thus, H2 is inconclusive. In Phase 2, we will conduct a study with more response data to better investigate H2.

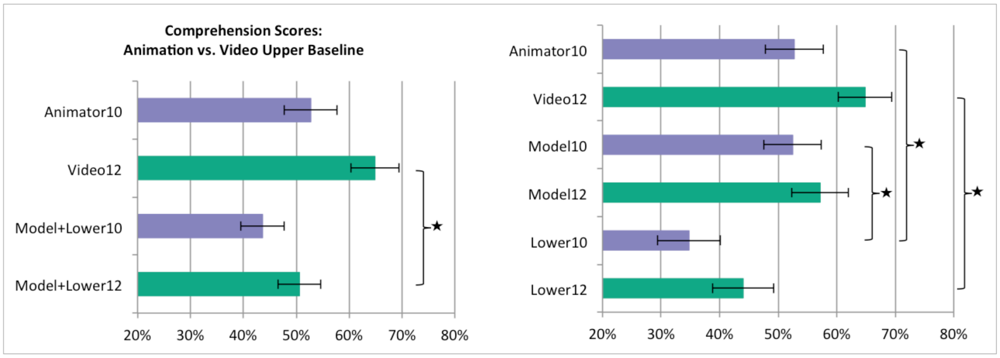


Figure 18: Results of comprehension scores in Phase 1.

Figure 19 illustrates the 1-to-10 Likert-scale subjective scores for grammaticality, understandability, and naturalness that participants answered after watching the short stories in the studies. Hypothesis H3 was supported by the data in Figure 19; the video upper baseline received higher subjective Likert-scale scores than the animation upper baseline. All three scores had the same pattern: Video12 had significantly higher grammaticality, understandability, and naturalness scores than Animator10.

Hypothesis H4 was that the use of a video upper baseline would lead to a change in the Likert-scale subjective scores for the other stimuli in the study. The data in Figure 19 did not support hypothesis H4; there was no significant change in the Likert-scale scores for Model or Lower when we used the video upper baseline in 2012. When we examine the Likert-scale scores obtained during side-by-side comparisons in Figure 20, we will see some contradictory results in regard to Hypothesis H4.

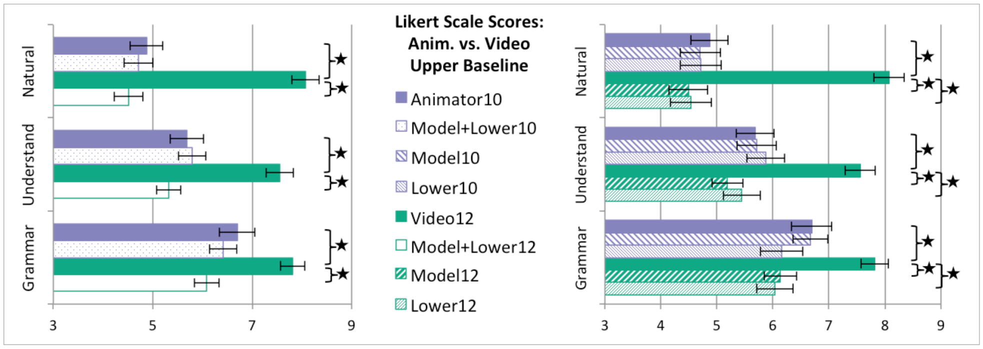


Figure 19: Results of grammaticality, understandability, and naturalness scores in Phase 1.

Figure 20 illustrates the scalar subjective scores collected from participants during part 2 of the studies (the side-by-side comparison of identical sentences, with different versions of the verb in each, which could be replayed many times). Video12 is significantly higher than Animator10, further supporting Hypothesis H3 that video upper baselines would get higher Likert-scale subjective scores than animation upper baselines.

The results in Figure 20 supported Hypothesis H4: Using a human video upper baseline depressed the subjective Likert-scale scores that participants gave to the animations. The 2012 values for Model+Lower, Model, and Lower were all significantly lower than their 2010 counterparts. The magnitude of this depression is 10%-20%. This is not a surprising result; when looking at videos of humans in direct comparison to animations of a virtual human character, it is reasonable that participants would feel that the animations are lower quality. What is surprising is that we had not observed any significant depression in Figure 19 when looking at the Likert-scale data from part 1, in which participants assigned a Likert-scale subjective score to a story that they had just watched. We speculate that the depressive effect may depend on whether participants are assigning Likert-scale subjective scores to videos in a side-by-side direct comparison (as in part 2, Figure 20) or sequentially throughout a study (as in part 1, Figure 19). Perhaps the side-by-side setting forced participants to be more comparative – with the video "standing out" from the other two stimuli, which were both animations. Another possible explanation for this result may be that during part 1 of the study, when watching a story one time and then answering the comprehension questions, the participants may have been very focused on the task of trying to understand and remember as much information as possible from the stories. Thus, they may have been less focused subjectively on the superficial appearance of the animations/videos. We will explore H4 further in Phase 2 in this chapter.

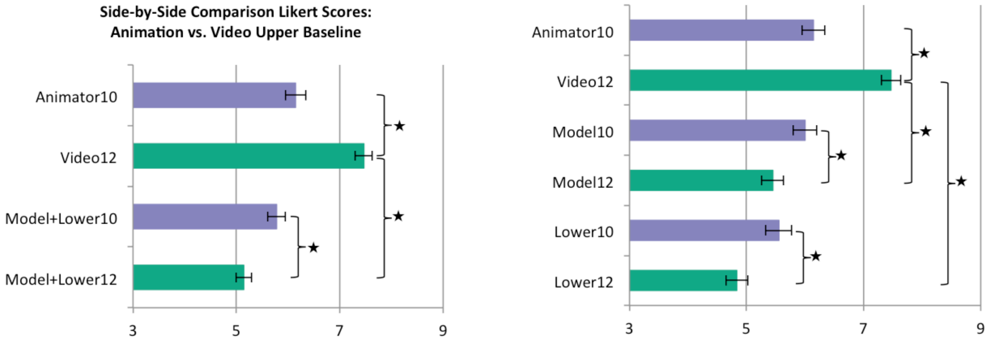


Figure 20: Results of side-by-side comparison scores in Phase 1.

As discussed in Section 5.2.2, when creating baselines for comparison to animations in a study, a balance must be achieved between matching the content of the stimuli across versions and allowing for natural signing. Some of the comments of participants in the study indicated that in a few cases, we were not successful at this. Specifically, when producing the script for the human to perform in the video recordings, we included every sign that was performed by the virtual human character in the upper baseline animations from 2010. When a signer sets up points in space to represent entities under discussion, the signer may refer to these items later in the conversation by pointing to them. Because the movement path of an inflected ASL verb indicates the location around the signer where the subject and object of the verb are established, it is common (but not required) for signers to omit pointing to the subject/object before/after the verb (because the location in space that represents those entities is already indicated by the motion-path of the verb). The human animator who produced our upper baseline animations in 2010 still included some extra "pointing" to these locations, and so we included them in the script given to the human signer in 2012. In the feedback comments in 2012, some participants said: "Most verbs shouldn't end with the pointing of the finger (or direction) as the action already indicated that much," "too many endings were a pointing, it threw off my attention a lot," etc. What is interesting is that no participants criticized this in 2010; thus, when they saw a human signer performing this extra pointing movement, it felt more unnatural and warranted a comment at the end of the study.

5.3 Phase 2: Animation vs. Video Baselines with Facial Expression

While focusing in facial expression synthesis, for Phase 2, we decided to conduct another round of experiments with ASL animations with facial expressions. To better investigate some of the partially supported hypotheses in Phase 1, we conduct a study with a larger number of comprehension question responses recorded. Here, we'd expect that a human video upper-baseline would receive even higher comprehension scores than animation upper-baseline produced by an animator for the following reasons: First, animating facial expressions accurately is too difficult with the use of current ASL animation technology (Elliott et al., 2008; Filhol et al., 2010; Fotinea et al., 2008; Vcom3D 2014). Handling complex aspects of facial expressions such as the exact face, timing of the intensity with the hands, simultaneous performance, and transitions is beyond the state of the art of current ASL systems (Huenerfauth et al., 2011). Thus, what an animator can achieve as an upper baseline might lack the naturalness and the quality of video of a human signer. Second, facial expressions introduce new challenges in achieving that delicate "balancing act" between a natural, fluent version of the stories from the human signer and the control of important variables between the upper-baseline and the animation being evaluated. Finally, deaf viewers tend to focus on signers' faces, as shown in (Emmorey et al., 2009), which could result to an audience sensitive to facial expressions errors.

5.3.1 Design of Evaluation Studies for Phase 2

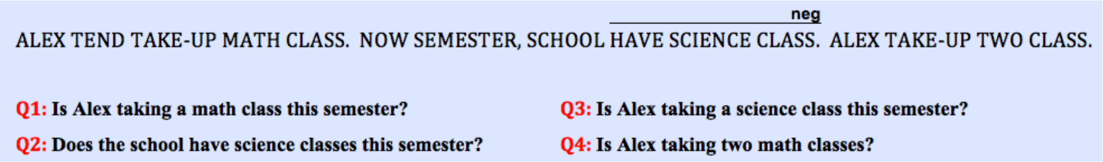
In Phase 2 we evaluate a model for performing facial expressions in ASL animation, whose movements of the face depend on the grammar or emotions of the stories being displayed. For this study, we investigated six categories of meaningful facial expressions: yes/no question; rhetorical question; negative; topic; wh-word question (e.g., who, what, and how); and emotions (e.g., frustration, sadness and irony). For this pair of studies, because we were primarily interested in the affect of different upper baselines, it was not important for the "model" being evaluated to be very sophisticated. Thus, our model was a simplistic rule: apply a facial expression from the above six categories over a whole sentence with the same grammar/meaning, e.g. a y/n-question face over the entire ASL sentence asking a question that can be answered with a yes or no.

We want to evaluate the understandability of ASL animations of three types: (1) a lower baseline consisting of stories with a static face (no facial expressions), (2) animations with facial expressions that follow the simplistic "model" above, and (3) an upper baseline. To evaluate hypotheses H1-H4, we conducted a pair of studies: (i) one in which the upper baseline was a video of a human ASL signer and (ii) one in which the upper baseline was an animation whose facial expressions were produced by an ASL signer using some animation software (Vcom3D, 2014). Since the same model is being evaluated in both studies, the simplicity of the model does not affect the comparison of the different upper baselines used in each study. The design and conduct of these studies was similar to the pair of studies in Phase 1.

Our studies consisted of two parts: In part 1, participants viewed animations of a virtual human character or human videos telling a short story in ASL. Each story included one of the above six categories of facial expressions (whereas the stories in Phase 1 focused on complex ASL verb signs). The stories[8] were selected from the stimuli collection in Chapter 4. Figure 21 shows a story transcript (corresponding to the stimuli with codename N2 in the collection); the bars with the abbreviations over the script indicate the required facial expression to be performed during some of the signs. An English translation of this story would be: "Alex usually takes Math classes. This semester, the school doesn't have any science classes. Alex is taking two classes." Participants watched each story, which was one of three types: lower baseline (no facial expression), animation with facial expressions based on our simple-rule model, and upper baseline (which was either a human video or an animator-produced

[8] Their codenames in the collection are: E1, E2, E4-E6, W2-W4, Y3, Y4, Y6, R3, R5, R7, N1-N3, and T3-T5. Appendix A provides more details on their script and comprehension questions.

animation). Then, participants answered 4 yes/no comprehension questions (Figure 21). As discussed in Chapter 4, stories and questions were engineered in such a way that the wrong answers would indicate that the users misunderstood the facial expression displayed. In Figure 21, if the "negative" facial expression was not noticed by a participant, then the participant would think that the school does not offer science classes, which would affect the participant's answers to the questions. For each story viewed, participants also responded to 1-to-10 Likert-scale questions about how grammatically correct, easy to understand, naturally moving the hands and the face of the animation/human signer appeared. These Likert-scale questions were identical to those used in Phase 1.



ALEX TEND TAKE-UP MATH CLASS. NOW SEMESTER, SCHOOL $\overset{\text{neg}}{\overline{\text{HAVE SCIENCE CLASS.}}}$ ALEX TAKE-UP TWO CLASS.

Q1: Is Alex taking a math class this semester?

Q2: Does the school have science classes this semester?

Q3: Is Alex taking a science class this semester?

Q4: Is Alex taking two math classes?

Figure 21: Example script (originally shown in Fig. 12, repeated here for convenience) and corresponding comprehension questions for a story shown in the study.

In part 2 of the studies, participants viewed three versions of a single ASL sentence side-by-side on one screen. The sentences shown side-by-side were identical, except that they were of different versions: lower baseline animation, model-based animation, and upper baseline (either a video or an animation, depending on the study). Figure 22(a) contains a screenshot of what a participant would see side-by-side in the study with the animation upper baseline, and Figure 22(b) depicts what was seen in the study with the video upper baseline. The participants could re-play each animation as many times as they wished. Participants were asked to focus on the facial expressions and respond to a 1-to-10 Likert-scale question about the quality of each of the three versions of the sentence. The methodology used here is similar to studies in Phase 1.

For both studies, we selected a total of 28 ASL stories, distributed as follows: y/n-question (4), rh-question (4), negative (4), topic (4), wh-question (4), and emotions (8). The stories were split in the two parts of the studies in a 3:1 ratio, resulting at 21 stories for the part 1 and 7 stories for the part 2. Beside the upper-baselines used, all the other animations and their sequencing in our pair of studies were identical. A fully-factorial design was used such that: (1) no participant saw the same story twice, (2) order of presentation was randomized, and (3) each participant saw 7 animations of each version: i) lower baseline, ii) model, or iii) upper baseline. Again, all of the instructions and interactions for both groups were conducted in ASL by a Deaf signer, who is a professional interpreter. Part of the introduction, included in the beginning of the experiment, and the comprehension questions in part 1 of both studies were presented by a video recording of the interpreter.

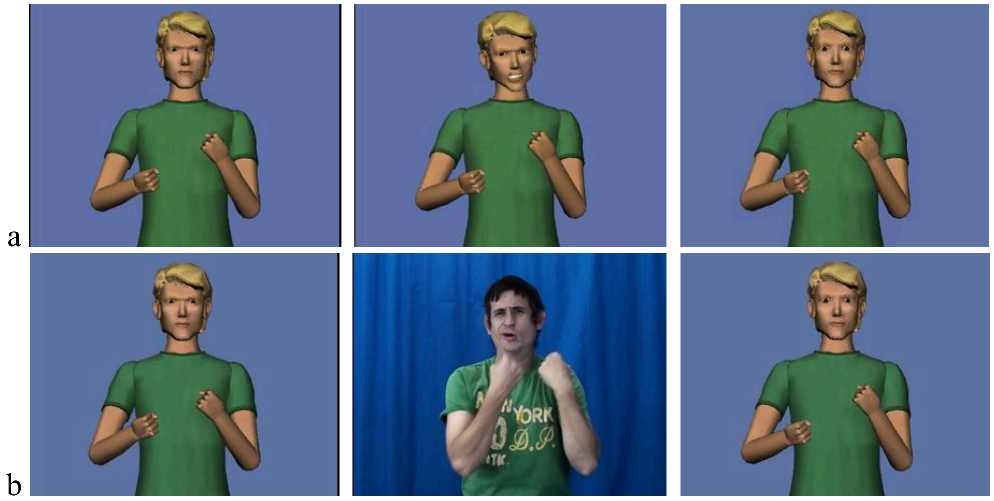


Figure 22: Screenshots of the side-by-side comparison portion of the studies as shown to participants in (a) animator-upper-baseline study and (b) video-upper-baseline study.

5.3.2 Creation of Human Video Upper Baseline for Phase 2

To produce the human video upper baseline, as done in Phase 1, we recorded an ASL signer (the same person as in Phase 1) in our studio, sitting on a stool in front of a blue curtain, wearing a green t-shirt. The camera placement was identical to the recording process in Phase 1, and the same monitor was placed below the camera with the story scripts (like the example story shown in Figure 21). As before, the signer had time to memorize and practice each of the scripts prior to the recording session. All of the instructions and interactions for the recording session were conducted in ASL by another signer (same person as in Phase 1) sitting behind the camcorder. The cropping, placement of the signer in the video frame, video size, resolution, and framerate were identical to the animations in this study and the human video in Phase 1.

For the recorded video to serve as an effective upper baseline for comparison, we again wanted to "control" as many of the variables of the ASL performance as possible. For this study, we primarily care about how the facial expressions differ between the upper baseline and our model animation under evaluation; so, we would prefer for the other aspects of the performance to be identical. While in Phase 1, the animations being evaluated pre-existed the video upper baseline (because we were replicating an earlier study), during Phase 2, we were able to record the video upper baseline prior to producing our animations. Thus, the human signer had fewer constraints on the performance because they did not need to mimic the animation. However, producing such a video of a human was still challenging. Because our stories and comprehension questions were carefully engineered, the signer had to perform a specific "script" of signs and the correct category of facial expression during a story, since a difference in the facial expressions could result in a different meaning (e.g. negating a statement). Since ASL has no standard written form, we had to explain our notation scheme (Figure 21) to the participant being recorded. The stories were somewhat complicated and were engineered to cause confusion if the wrong facial expressions were applied (as discussed in Chapter 4). They were an average of 9 signs in length, with 1-4 main characters set up at various locations in the signing space, with at least one facial expression per story. So, the human signer required several minutes of practice in order to perform each story smoothly. During the recording process, the signer glanced at the script occasionally; so, the videos include some moments when his eyes glance between the monitor displaying the story transcripts and the camcorder.

While the signer had greater freedom in performance than in Phase 1, we still had to let the signer know about some restrictions, in regard to: (i) pausing between stories and positioning the hands down at a default pose at the beginning and end of each story, (ii) controlling the intensity of the facial expression in the story (so as not to be comically exaggerated in an unnatural manner), and (iii) avoiding embellishments, e.g., additional emotional facial expressions on top of grammatical facial expressions. Since we were not investigating co-occurring facial expressions in our experiment, it would be undesirable for the human signer to add such embellishments to the video. We also asked the signer to avoid using ASL signs that were influenced by English, such as alphabet-letter-initialized signs. After practicing for a few minutes, the signer attempted to perform each story multiple times until we produced an acceptable recording. An average of 3 attempts per story were recorded. From an overset of 39 ASL story scripts (available in our stimuli collection) that were used during the recording session, we selected to include in the study only the recordings of the 28 ASL stories that best met the above criteria.

Even though we started with the human signing the scripts, the coaching and scripting process, described in Section 5.2.2 as a delicate "balancing act," was still challenging for this study. The human signer needed to exercise control over micro-movements of his face, which is an acting skill that is beyond that needed in spontaneous signing. Further, it was difficult for the research team to evaluate the quality of these micro-movements in real time; so, it was necessary to replay videos during the recording process to assess the quality. Moreover, our standards for the video quality were higher in Phase 2 because we expected our study participants to be very sensitive to unnatural facial expressions. Prior research on Deaf ASL signers has indicated that they visually fixate primarily on the face of the person who is signing (Emmorey et al., 2009).

5.3.3 Data Collection and Results of Phase 2

Similar methods were used as in Phase 1 to ensure that participants were native or fluent ASL signers and that the study environment was ASL-focused with little English influence. Ads were posted on New York City Deaf community websites asking potential participants if they had grown up using ASL at home or had attended an ASL-based school as a young child. The study with the video upper baseline included 18 participants: 15 participants learned ASL prior to age 5, and 8 participants attended residential schools using ASL since early childhood. The remaining 10 participants had been using ASL for over 12 years, learned ASL as adolescents, attended a university with classroom instruction in ASL, and used ASL daily to communicate with a significant other or family member. There were 9 men and 9 women of ages 22-45 (average age 31.6). The study with the animation upper baseline included 16 participants: 10 participants learned ASL prior to age 5, and 6 participants attended residential schools using ASL since early childhood. The remaining 10 participants had been using ASL for over 9 years, learned ASL as adolescents, attended a university with classroom instruction in ASL, and used ASL daily to communicate with a significant other or family member. There were 11 men and 5 women of ages 20-41 (average age 31.2).

Figure 23Figure 24Figure 25 display the results from the video-upper-baseline and animation-upper-baseline studies, including the following response data: comprehension-question scores and Likert-scale scores collected in part 1 of the studies (after a participant viewed a story) and the Likert-scale scores collected in part 2 of the studies (in which participants assigned a score to each of the three sentences viewed side-by-side). Labels ending with the letter "A" indicate data collected in animation-upper-baseline study, and labels ending in the letter "V" indicate data collected in video-upper-baseline study. See Section 5.2.3 for additional details (not repeated here) about error bars, the pooling together of the Lower and Model data to produce the "Model+Lower" category, layout of the graphs, statistical tests performed, planned comparisons, and the use of stars to indicate statistical significance in the graphs.

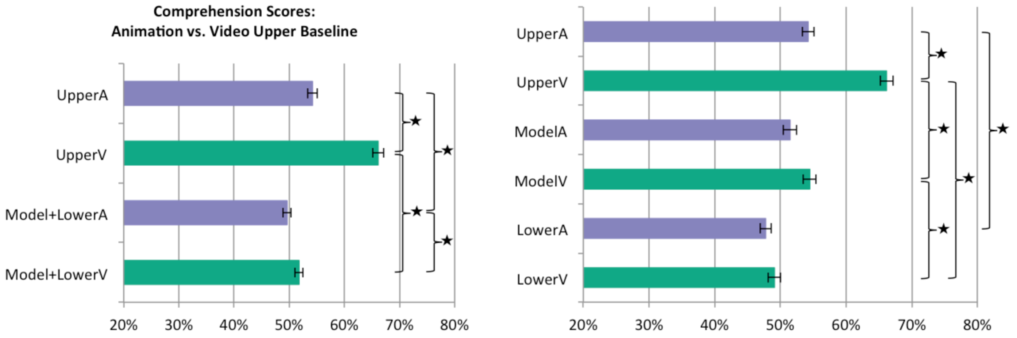


Figure 23: Results of comprehension scores in Phase 2.

Figure 23 displays the comprehension-question accuracy scores from part 1 of the studies. We see that there was a significant difference between UpperA and UpperV; so, here hypothesis H1 was supported. This is a different outcome than was observed in Phase 1; here, the videos of a human signer achieved higher comprehension scores than the animations of a virtual human with the facial expressions carefully animated by a human. Given that handling complex aspects of facial expressions is beyond the state of the art of current ASL synthesis systems, it was not surprising that the upper baseline created by an ASL animator received lower scores than the human video.

In Phase 2, Hypothesis H2 was not supported. In fact, contrary to our hypothesis, Model+LowerV actually had *significantly higher* comprehension scores than Model+LowerA (Mann Whitney U-test, alpha=0.05). However, the magnitude of this difference was quite modest (approximately 2% higher), and TOST equivalence testing indicated that the values are actually "equivalent" according to our (-15%, +15%) margin. While no story was displayed more than one time during the study, we speculate that seeing a video of a human performing some of the ASL stories may have helped participants grasp the overall genre of the stories in the study. Perhaps participants were able to realize that all of the stories contained a subtle ambiguity that depended on the facial expression and that the facial expressions were of different types (yn-question, emotion, etc.). This may be why there were slightly higher comprehension scores for Model+Lower in the video-upper-baseline study.

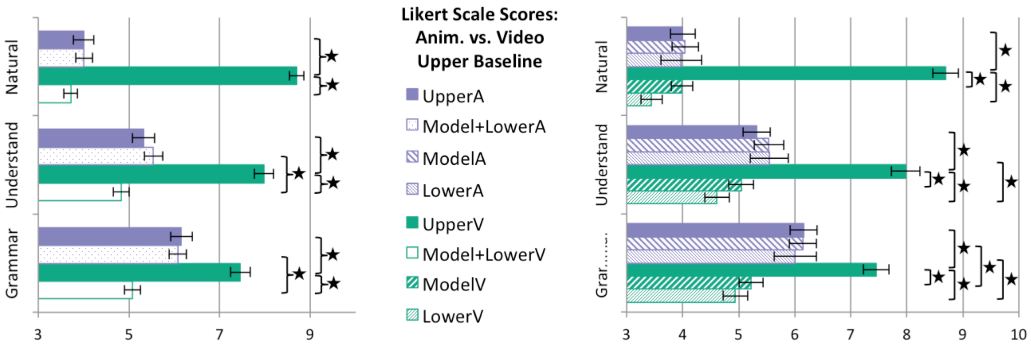


Figure 24: Results of grammaticality, understandability, and naturalness scores in Phase 2.

Figure 24 displays the 1-to-10 Likert-scale subjective scores for grammaticality, understandability, and naturalness from part 1 of the studies. UpperV had significantly higher grammaticality, understandability, and naturalness scores than UpperA – thereby supporting hypothesis H3, that video upper baseline would get higher subjective Likert-scale scores than an animation upper baseline.

The results in Figure 24 partially support hypothesis H4, that the use of a video upper baseline would affect the Likert-scale subjective scores for the other stimuli in the study. For grammaticality and understandability, we see a significant difference between “Model+LowerA” and “Model+LowerV.” This is a different outcome than was observed in Phase 1, when no significant difference was observed for Likert-scale data in part 1. We speculate that the videos of a human with facial expression was perceived so much better than animations by our participants in Phase 2 that it may have affected the participants "calibration" of their Likert-scale responses, resulting in lower Likert-scale subjective scores (for grammaticality and understandability) for the non-video stimuli.

Figure 25 displays the Likert-scale subjective scores collected from participants during part 2 of the studies (the side-by-side comparison of identical sentences, with different versions of the facial expressions that could be replayed multiple times). UpperV is significantly higher than UpperA, further supporting hypothesis H3 that human upper baselines would get higher Likert-scale subjective scores than animation upper baselines.

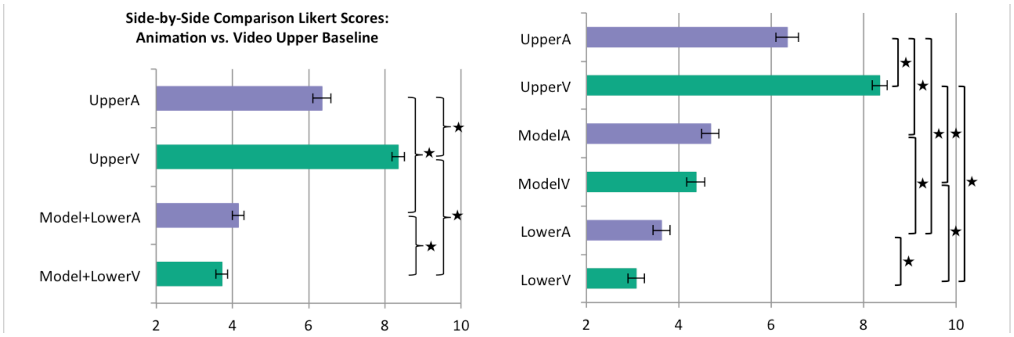


Figure 25: Results of Side-by-side comparison scores in Phase 2.

In Figure 25, hypothesis H4 was again supported: Using a human video upper baseline depressed the subjective Likert-scale scores that participants gave to the animations. Mode1+LowerV was significantly lower than Model+LowerA (same for the detailed pairs ModelV/ModelA and LowerV/LowerA). The magnitude of this depression is 10%-20%. As in Phase 1, this is not a surprising result; when looking at videos of humans in direct comparison to animations of a virtual human character, it is reasonable that participants would feel that the animations are less natural/grammatical.

It is notable that we did not observe a significant depression for naturalness in Figure 24.

As mentioned in Section 5.2.3, we speculate that the depressive effect of displaying video upper baselines may depend on whether participants are assigning Likert-scale subjective scores to videos in a side-by-side direct comparison (as in part 2, Figure 25) or sequentially throughout a study (as in part 1, Figure 24). Perhaps in the side-by-side setting, a greater "comparativeness" is triggered in the participants, and the visual distinctness of the video "stands out" in comparison to animations – thereby resulting in a stronger depressive effect on the Likert-scale scores for the other stimuli in the study.

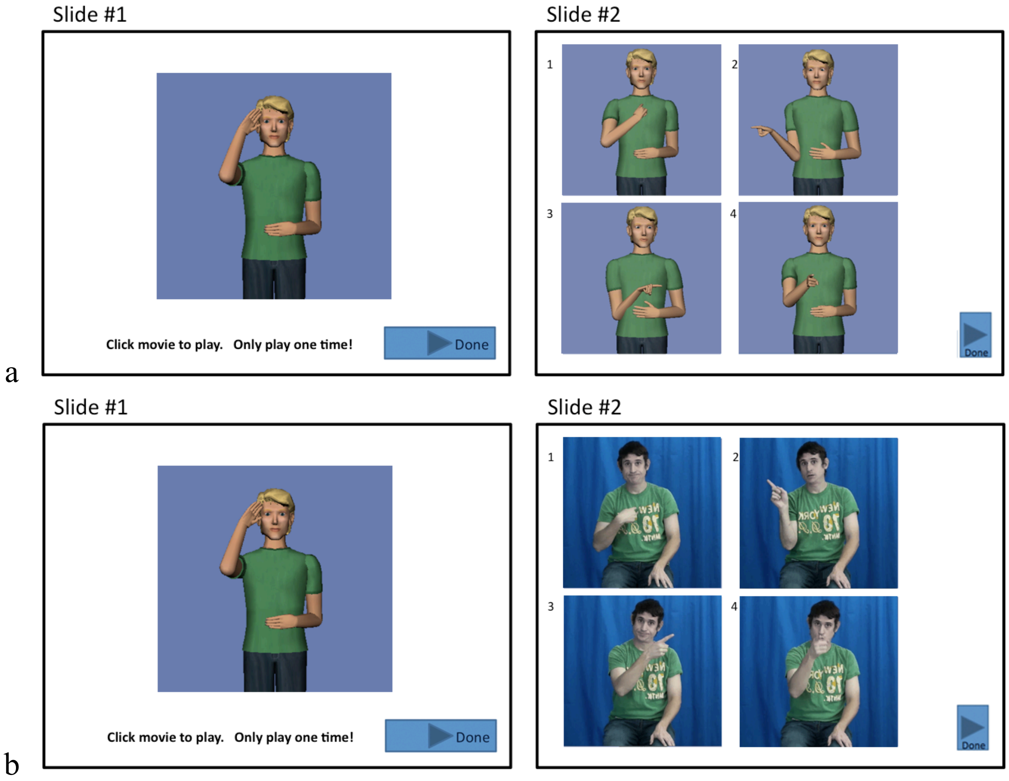
5.4 Phase 3: Animation vs. Video Comprehension Questions

Aside from being used as an upper baseline, a video of a human signer could appear within the software interface presenting animations in a user study (as mentioned in Section 3.3). Specifically, comprehension questions that participants are asked after viewing each of the sign language animations might be displayed in different modalities (human video or animation). In this study we investigate whether this choice affects the comprehension (Hypothesis H5) and subjective Likert-scale (Hypothesis H6) scores collected in a study with deaf participants. To evaluate these hypotheses, we conducted a final pair of studies: one with comprehension questions presented as high-quality ASL animations and one with comprehension questions presented as human videos.

5.4.1 Design of Evaluation Studies for Phase 3

Researchers at our lab had previously conducted a user study that presented the comprehension questions in the form of animations of a virtual human character (Huenerfauth and Lu, 2010);

for that study, an ASL signer with animation experience carefully produced the animations using sign language animation software (Vcom3D, 2014). For Phase 3, we decided to replicate that study; we replaced the animations containing the comprehension questions with videos of human signer performing identical ASL questions. Because we wanted to isolate the effect of showing a human video for the comprehension questions, we decided to use an animation as the upper baseline in both the 2010 and 2013 study. So, the only difference between the 2010 and 2013 studies is whether the comprehension questions were presented in the form of: (1) video in which a human signer asked questions in ASL about information contained within the stories being presented or (2) animation in which an virtual human character asked identical questions. It is important to note that the stimuli shown in the two studies were identical; the only difference was how the comprehension questions were presented. Figure 26(a) illustrates what the participants saw in that earlier animation-comprehension-questions study in 2010 (with an ASL story shown on one slide and four questions displayed on the next slide), and Figure 26(b) illustrates what participants saw in our video-comprehension-questions study in 2013. We are interested in how participants' scores on comprehension questions might change (Hypothesis H5), and we are also interested in whether there might be an effect on the scores for the Likert-scale questions about the naturalness, understandability, and grammaticality of the ASL stories (Hypothesis H6). Unlike the studies described in Phase 1 and 2 of this chapter, we did not conduct the part 2 side-by-side comparison of the stories in our Phase 3 experiment because no aspect of that part of the study differed between 2010 and 2013, since there were no comprehension questions asked in that part of the study. Similar to the study in 2010, participants were asked to view a total of 9 short stories in ASL. Again, a fully-factorial design was used such that: (1) no participant saw the same story twice, (2) order of presentation was randomized, and (3) each participant saw 3 animations of each version: i) lower baseline, ii) model, or iii) upper baseline.



Click movie to play. Only play one time!

Click movie to play. Only play one time!

Figure 26: Screenshots of two forms of comprehension questions presented in 2010 study (a) and 2013 study (b).

To produce the human video comprehension questions, we recorded the videos of an ASL signer, with the same blue background, camera angle, and other details, as described in Section 5.2.2. We used one large monitor in front of the signer to display the comprehension-question scripts (the studio setup is similar as shown in Figure 17).

5.4.2 Results of Phase 3

This section presents the results of our 2013 replication of an original 2010 study, and it compares the results of these two studies. The data collected include the comprehension-question and Likert-scale scores after a participant viewed a story one time. Details of the participants in the 2010 study were described in Section 5.3.3. The 2013 study included 18 participants: 17 participants attended residential schools using ASL since early childhood, and the 18th participant used ASL since birth, attended mainstream schools, and attended a university with instruction in ASL. 15 participants learned ASL prior to age 5. There were 12 men and 6 women of ages 20-37 (average age 28.8).

In the results illustrated in Figure 27 and Figure 28, the thin error bars display the standard error of the mean. Animator10 and Animator13 were upper baselines, Lower10 and Lower13 were lower baselines, Model10 and Model13 were versions of the animations produced using our lab's verb inflection model (details in Section 5.2), and Model+Lower10 and Model+Lower13 were the combined data from the Model and Lower groups. It is important to note that all of the stimuli were identical in 2010 and 2013 (i.e., Animator10 and Animator13, etc.); the only difference in those two studies was that comprehension questions used to collect the evaluation scores were presented in different forms: as animations in the 2010 study and as videos of a human signer in the 2013 study.

As mentioned in Section 5.2.3, since H5 and H6 hypothesize no differences, TOST equivalence testing was performed. With alpha=0.05, 90% confidence intervals were calculated (via t-tests for comprehension question scores and via Mann-Whitney U-tests for Likert-scale scores) for the following pairs of values: (a) Animator13 and Animator10, (b) Model+Lower13 and Model+Lower10, (c) Model13 and Model10, and (d) Lower13 and Lower10. Section 5.2.3 explained how we selected an equivalence margin interval of (-

1.5,+1.5) for Likert-scale scores and (-0.15,+0.15) for comprehension question scores. According to the TOST procedure, whenever the entire confidence interval falls within our equivalence margin interval, then the pair of scores is deemed equivalent.

While it was not necessary for examining H5 and H6, we also conducted one-way ANOVAs (for comprehension question scores) and Mann-Whitney U-tests (for Likert-scale scores) for the following pairs of values: (1) all three values from 2010, (2) all three values from 2013, (3) Animator13 and Model+Lower13, and (4) Animator10 and Model+Lower10. Statistical significance (p<0.05) for any of these comparisons has been marked with a star in Figure 27 and Figure 28. As discussed in Section 5.2.3, while not necessary for testing our hypotheses, we believe it is useful for us to present these statistical tests in our thesis, for reference, since future researchers evaluating the quality of an animation relative to upper and lower baselines would traditionally perform these comparisons.

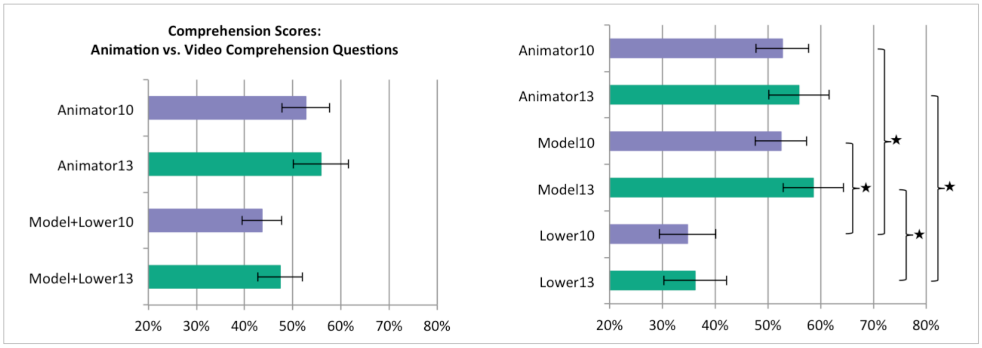


Figure 27: Results of comprehension scores in Phase 3.

Hypothesis H5 would predict that the mode of presentation of the comprehension questions (animation vs. video) would not affect the comprehension scores collected in the study. This was mostly supported by the results. TOST equivalence testing indicated that the following pairs of values were equivalent: Animator10 and Animator13, Lower10 and Lower13, and Model+Lower10 and Model+Lower13. (The results were inconclusive for Model10 vs. Model13.) It should be noted that the animations used in 2010 to present the comprehension questions were carefully produced by a human animator and were of good quality; we predict that if low-quality animations had been used, then we would have seen lower comprehension scores in 2010 (due to confusion from participants who did not understand the question being asked).

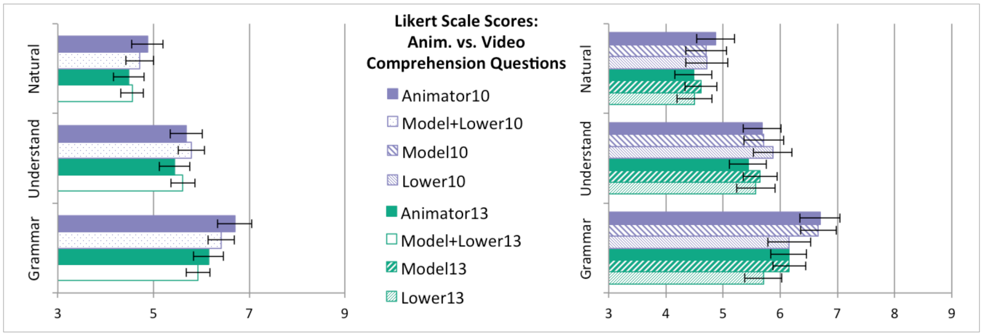


Figure 28: Results of grammaticality, understandability, and naturalness scores in Phase 3.

Figure 28 illustrates the 1-to-10 Likert-scale subjective scores for grammaticality, understandability, and naturalness from the studies in Phase 3. Hypothesis H6 would predict that the mode of presentation of the comprehension questions (video vs. animation) would have no effect on the Likert-scale subjective evaluation scores in the study, and this was mostly supported by TOST equivalence testing. (The scores for Naturalness for Animator10 and Animator13 were inconclusive; all other compared values were determined to be equivalent.) Thus, we conclude that H6 was supported.

5.5 Conclusions and Recommendations for Future Researchers

This chapter has investigated several methodological issues (formulated as RQ2 and RQ3 in the Prologue) that are important for researchers in the growing field of sign language animation research, who are conducting experimental studies with sign language users evaluating their animations. Specifically, we examined whether certain choices in experiment design affect the comprehension and subjective scores collected in a study. We quantified the effects of changing the mode of presentation (video vs. animation) of two elements of a study: the upper baseline stimuli and the comprehension questions. Awareness of such effects is important so that future researchers can make informed choices when designing new studies and so that they can fairly compare their results to previously published studies, which may have made different methodological choices.

In order to investigate these issues, we conducted several replications of experiments in which most of the stimuli were held constant, and we were able to measure if there was a difference in the scores collected from participants when we changed the upper baseline or modality of presentation for the comprehension questions. To make our results useful to a wide variety of researchers, we included a variety of study designs and question formats, including: comprehension questions responses, Likert-scale subjective scores of a single stimulus, and Likert-scale subjective scores of multiple stimuli presented side-by-side. The three pairs of experiments conducted allowed us to evaluate six hypotheses:

H1: Video upper baselines received higher comprehension scores than animation upper baselines. This difference was significant in Phase 2 experiments, which included

ASL animations with facial expressions, but it was not significant in Phase 1. We speculate that the effect may occur when there is a greater quality-difference between the animations and videos, as there is when they include facial expressions, which are not handled well by current animation tools. An alternative may be that the effect could be observed in Phase 2 because the studies contained more stories, and thus a larger number of comprehension question responses were collected.

H2: When video upper baselines were used instead of animation upper baselines, the comprehension question scores for the other stimuli in the study increased slightly. This increase was significant in Phase 2, but it was inconclusive in Phase 1. We speculate that the video upper baseline shown during the study may have given the participants some additional advantage in answering the comprehension questions for the other stimuli because it allowed them to better understand the genre of stories being presented, the relationship between the stories and the comprehension questions, or the types of facial expressions that they should expect to see in the other (animation) stimuli. We had not hypothesized that we would observe such an effect: H2 had been that changing the upper baseline from animation to video would have no effect on the comprehension question scores for the other stimuli. The presence of this effect may depend on the degree to which participants in the study are able to "learn something useful" from watching the high-quality video upper baseline stimuli that generalizes to the other stimuli in the study; thus it may be magnifying a "learning effect" that was inherent to the design of a study.

H3: Video upper baselines received higher Likert-scale subjective scores than an animation upper baseline. This hypothesis was supported by statistically significant differences observed in both Phase 1 and Phase 2. This effect was present in both Likert-scale subjective scores collected after sequential presentation of an individual stimulus (as done in part 1 of the studies in Phases 1 and 2) and in Likert-scale subjective scores collected during simultaneous presentation of multiple stimuli side-by-side (as in part 2 of the studies). Given the state of the art of sign language animation technologies (and that there are still many unsolved challenges to address in the field), it is not surprising that videos of humans would receive higher subjective evaluation ratings than animations.

H4: This hypothesis is best considered if it is split into two sub-cases: (H4a) for sequential presentation of stimuli, as in part 1 of the studies in Phase 1 and 2, and (H4b) for simultaneous side-by-side presentation of stimuli, as in part 2.

H4a: When video upper baselines were used instead of animation upper baselines, the Likert-scale subjective evaluation scores for the other stimuli in the study decreased during sequential presentation. This difference was significant for some of the Likert-scale categories in Phase 2 (grammaticality and understandability), but it was not significant for any of the categories in Phase 1. So, sub-hypothesis H4a is partially supported by the results presented in this chapter – during the experiments with sign language animations with facial expressions.

H4b: When video upper baselines were used instead of animation upper baselines, the Likert-scale subjective evaluation scores for the other stimuli in the study decreased during simultaneous (side-by-side) presentation of stimuli. This difference was significant in both Phase 1 and Phase 2. As discussed earlier, we speculate that participants felt a greater sense of "comparativeness" when the stimuli were shown side-by-side, and this may have strengthened the depressive effect on the Likert-scale scores for the animation stimuli when a video upper baseline was shown.

H5: Displaying comprehension questions as videos of a human signer or as a high-quality animation will not affect the comprehension questions accuracy scores. Statistically equivalent comprehension scores were collected in the studies with video or with high-quality animations used to present comprehension questions in Phase 3.

H6: Displaying comprehension questions as videos of a human signer or as high-quality animations will not affect the subjective Likert-scale scores that participants assign to the animations in the study. The choice of video or high-quality animation led to statistically equivalent Likert-scale scores collected in both studies in Phase 3.

The two main contributions of this chapter are: (a) providing methodological guidance for future researchers who are conducting studies with sign language and (b) facilitating fair comparisons of the results of sign language animation evaluation studies, in which the authors have made different methodological choices.

5.5.1 Recommendations for Future Researchers

This section discusses how the conclusions outlined above can be translated into concrete methodological guidance for researchers conducting evaluation studies with sign language animations. First of all, while not a major focus of this chapter, the conclusions above and the prior research in our lab (Huenerfauth et al., 2008) have indicated that comprehension question scores and Likert-scale subjective evaluation scores for sign language animations often do not have identical results, and we recommend to future researchers that they include both forms of evaluation in any future studies, since they may be measuring different aspects of sign language animations. Of course, the primary focus of this chapter has been on the mode of presentation for two aspects of a sign language evaluation study: the upper baseline and the comprehension questions. In particular, two forms of presentation were examined: videos of human signers and high-quality animations of a virtual human. The results in this chapter do not give a single “correct answer” for the best choice that future researchers should make when designing their studies; indeed, either choice (video or animation) is potentially valid. The selection should be based on the research goals of the study, the practical challenges in producing animations or videos, and the expected effect of these methodological choices on the data collected.

5.5.1.1 Upper Baseline Based on Goals of the Study

Researchers considering which form of upper baseline to use should consider the research questions they want to explore. Given our slightly different results in Phase 1 and Phase 2 of this chapter, researchers may want to consider whether animations in their study are closer to those in Phase 1 (hand movements during verb signs) or Phase 2 (facial expressions), when considering our results. Each choice of upper baseline has trade-offs that must be considered in regard to the requirements of the study design:

Video upper baselines would be preferable for researchers studying computer graphics issues relating to the visual appearance of a virtual human for sign language animations, since this would serve as an "ideal" of photorealism.

Researchers who want to convey to a lay audience the overall understandability of their sign language animations (i.e., the current state of the art) may wish to use videos of humans as an upper baseline (because they are more familiar than animations as a basis for comparison). Of course, researchers would need to explain the limitations of current sign language animation technologies to manage the expectations of a lay audience being presented their results.

Researchers who are studying particular linguistic aspects of sign language animations (e.g., the speed/timing of signs or the timing of facial micro-expressions in relation to hand movements) may find an animated-character baseline more useful to their research because it is possible to control the variables of the character's movements precisely. A human in a video may not be able to voluntarily and consistently control these aspects of the performance as necessary for a study.

For study designs in which it is important that the participants cannot easily determine which stimuli are the upper baselines, animation (with a virtual human identical in appearance to the one in the synthesized animations) are desirable.

5.5.1.2 Challenges in Producing Videos or Animation

Researchers should also consider the practical challenges they may face in producing videos or animations for use as upper baselines or as comprehension questions:

Challenges in producing video upper baseline: We found that it is harder than many researchers may expect to produce a human video that is a good upper baseline (see Sections 5.2.2 and 5.3.2). Scripting and coaching was needed to ensure that our human videos had the same sign sequence, point locations, facial expressions, speed, and other performance variables as our other stimuli. If the study requires that some performance variable is held constant that is very detailed (e.g. precise millisecond timing of speed/pauses, exact height of the eyebrows, etc.), then this may be too difficult for a human to perform voluntarily and consistently. To avoid producing an artificial-looking result, a delicate "balancing act" (as discussed in Section 5.2.2) was needed between controlling the human's performance and providing freedom so that the result is fluent and natural. The researchers must decide what level of embellishments or improvisation they will tolerate from the human signer. If they prevent the signer from performing aspects of signing (because those aspects are outside the repertoire of the animation system being evaluated), then the video upper baselines may be artificially limited in how natural/fluent they appear.

Challenges in producing animation upper baseline: There are also challenges in producing a good-quality animation that is an effective upper baseline. Depending on the specific animation system/tool used by the researchers, the ease with which a human animator can control their virtual human to produce high-quality signing may vary. If it is possible to blend software-controlled with human-animator-controlled elements of the performance for the virtual human, then it may be easier to produce an upper baseline with variables that are held constant between the upper baseline and the other stimuli. In our studies, we found that our animation tool made it easy for the human animator to add novel hand movements, but the set of facial expression controls was limited. This may have resulted in some of our upper baseline animations in Phase 2 having lower quality than we would have preferred.

Challenges in producing video comprehension questions: While we did not find it especially difficult to produce video recordings of a human signer performing comprehension questions for use in our study in Phase 3, we did need to provide scripts and coaching during the process. Since there is no standard written form for ASL, it was necessary to explain our script notation to the signer. Further, there are regional/dialectical variations in how certain signs are performed in ASL, and we needed to ensure that the same variant of a sign used in our ASL story stimuli was used during the comprehension questions, to avoid confusion during the study.

Challenges in producing animation comprehension questions: In Phase 3, there was no significant difference in the comprehension or Likert-scale scores when we presented our comprehension questions either as videos or as animations. However, our comprehension question animations were high-quality animations produced by an ASL signer who had experience using our animation tool (Vcom3D, 2014). If future researchers were using an animation tool of lower quality (or asking comprehension questions that required some linguistic/performance aspect that was beyond the repertoire of their animation tool), then the resulting comprehension questions may be difficult for participants to understand. In that case, we would expect that the comprehension question scores collected in the study would be lower, due to the participants' difficulty in understanding the question being asked.

5.5.1.3 Effects on Collected Scores from Methodological Choices

Future researchers may also consider how the responses they collect will be affected by their choice of upper baselines and the mode of presentation for comprehension questions. Given that video upper baselines received higher comprehension and Likert-scale scores than animation upper baselines in our studies (Hypotheses H1 and H3), researchers should expect that if they use a video, their upper baseline scores would be higher. Thus, the other stimuli might appear relatively worse by comparison (to a naïve reader of their study who only considers the relative values of the raw scores). Similarly, when using videos as an upper baseline, we observed a depressive effect on the Likert-scale scores for the other stimuli in the study during side-by-side comparisons and, in some cases, during sequential evaluation of stimuli (Hypothesis H4).

Given that researchers may have an interest in the sign language animations they synthesize appearing more successful, this might suggest that there is an incentive for researchers to avoid video upper baselines. Given the advantages of video upper baselines for some study designs and the ease-of-interpretation of the results (mentioned in the "Goals of the Study" section above), it would be inappropriate to avoid the use of video upper baselines merely because they may make animations appear less understandable by comparison. It is for this reason that we believe methodological studies such as this thesis are important for the research community because it can provide a resource for future researchers who can explain how the results of their study should be interpreted when video upper baselines have been used for comparison.

While the use of video upper baselines clearly led to larger differences in the Likert-scale scores between the upper baseline and the other stimuli (since the upper baseline scores were higher and the other stimuli were lower), the results were more complex for comprehension question scores. In Phase 2, we observed an across-the-board increase in comprehension question scores for all of the stimuli in the study, when video upper baselines were used (Hypothesis H2). In our discussion of those results, we speculated that the result could have been due our participants learning something from watching the video upper baseline that generalized to the other stimuli in the study. Future researchers designing studies in which there could be a similar learning effect may see a similar resulting increase in comprehension scores when video upper baselines are used.

Based on the results in Phase 3, we did not see any significant difference in the scores collected in studies that used video or used animation to present comprehension questions. However, as noted above, this was evaluated using animations that were high-quality, and researchers may see lower comprehension question scores if they use difficult-to-understand animations to present their comprehension questions in a study.

Chapter 6 Evaluation of Facial Expressions with Eye-Tracking[9]

This chapter focuses on another important methodological issue: how eye tracking can be used in user-based experimental studies of sign language animations. As explained in Chapter 4, it is challenging to design experimental stimuli and questions that effectively measure whether participants have understood the information being conveyed by facial expressions. Signers may not consciously notice a facial expression during an ASL passage, and the subtle and complex ways in which facial expressions can affect the meaning of ASL sentences can make it difficult to invent stimuli and questions that effectively probe a participant's understanding of the information conveyed by the signer's face.

As discussed in related work (Section 3.4), researchers in various fields have used eye tracking to unobtrusively probe where participants are looking during an experiment (and in some cases, to infer the cognitive processes or task-strategies of those users). In fact, Sections 3.4.1 and 3.4.2 discuss how researchers have successfully used these methods with participants who are deaf, to investigate perception, reading, and sign language comprehension (of videos of humans). This chapter focuses on the research questions **RQ4-RQ5** and examines whether these methods can be effectively adapted to the evaluation of sign language animations; specifically, we ask:

- Does the eye-movement behavior of ASL signers participating in an experiment differ depending on whether they are watching a video of a human signer or an ASL

[9] This chapter describes joint work with Allen Harper, a Ph.D. student in Computer Science at CUNY, and Professor Matt Huenerfauth (Kacorri et al., 2013a; Kacorri et al., 2014).

animation?

- …whether they are viewing ASL animations with some facial expressions or ASL animations with no facial expressions?
- Does the eye-movement behavior of these participants correlate to their responses to subjective evaluations questions or comprehension questions about the videos/animations?

Section 6.1 describes how we selected which eye-tracking metrics to study by considering prior research. Also, it discusses how we formulated some hypotheses (more specific than the research questions listed above) about how the eye-movements of ASL signers relate to the quality of the ASL video/animation and to participants' responses to subjective and comprehension questions. Section 6.2 describes our experiments recording the eye movements of ASL signers who view animations/videos and then answer subjective and comprehension questions. Finally, Sections 6.3 and 6.4 describe our initial and extended analyses and their results, respectively.

Notably, for the research questions being investigated in this chapter, we are not primarily concerned with determining the level of quality of any particular ASL animation (which has traditionally been the focus of our prior experiments). Instead, we are focused on whether the eye movements of ASL signers reveal information about the quality of the ASL video/animation being viewed. We compare videos and animations under the supposition that very high-quality ASL animations may lead to similar eye-movement patterns as videos. If a correlation can be found between eye-movement metrics and certain types of ASL videos (or participants' responses to evaluation/comprehension questions about the stimuli), then this relationship could be utilized when designing future evaluation studies of ASL animations. Those metrics could be used as an additional or alternative form of evaluation for ASL animations. In some experimental contexts, it may be desirable not to interrupt participants with questions, or asking specific questions might artificially draw attention to aspects of the animation that could lead to unnatural interactions (e.g., if we wanted to study the effect of different eye-brow movements for our animated signer, if we ask too many questions about the eye-brows, then signers may stare at them, instead of simply watching the animations for their information content). In other contexts (such as for ASL animations with facial expression), it can be difficult to engineer large numbers of stimuli and questions that effectively probe whether the animation is of high quality or well understood.

6.1 Eye-gaze Metrics and Hypotheses

While our laboratory has investigated the calibration and use of motion-capture equipment (including eye-trackers) for recording sign language performances from human signers (Lu and Huenerfauth, 2009; Lu and Huenerfauth, 2010), no previous studies had been conducted that used eye-tracking technology to record signers while they viewed animations of ASL (nor did we find prior published work in which this was done). Therefore, we consider prior work on ASL *videos* (reviewed in Section 3.4.2) to determine the eye-tracking metrics we should examine and the hypotheses we should test.

Although Muir and Richardson (2005) did not study sign language *animation*, they observed changes in proportional fixation time on the face of signers when the visual difficulty of videos varied. Thus, we decided to examine the proportional fixation time on the signer's face. Since there is some imprecision in the coordinates recorded from a desktop-mounted eye-tracker, we decided not to track the precise location of the signer's face at each moment in time during the videos. Instead, we decided to define an AOI that consists of a box that contains the entire face of the signer in approximately 95% of the signing stories. (We never observed the signer’s nose leaving this box during the stories.) Details of the AOIs in our study can be found in Section 6.3.1.

The problem with examining only the proportional fixation time metric is that it does not elucidate whether the participant: (a) stared at the face for a long time and then stared at the hands for a long time or (b) often switched their gaze between the face and the hands during the entire story. Both types of behaviors could produce the same proportional fixation time value. Thus, we also decided to define a second AOI over the region of the screen where the signer's hands may be located, and we record the number of “transitions” between the face AOI and the hands AOI during the sign language videos and animations.

Since prior researchers have recorded that native signers viewing understandable videos of ASL focus their eye-gaze almost exclusively on the face, we make the supposition that if a participant spends time gazing at the hands (or transitioning between the face and hands), then this might be evidence of non-fluency in our animations. It could indicate that the signer’s face is not giving the participant useful information (so there is no value in looking at it), or it could indicate that the participant is having some difficulty in recognizing the hand shape/movement for a sign (so participants need to direct their gaze at the hands). As discussed in Section 3.4, in (Emmorey et al., 2009) less skilled signers were more likely to transition their gaze to the hands of the signer. If we make the supposition that this is a behavior that occurs when a participant is having greater difficulty understanding a message, then we would expect more transitions in our lower-quality or hard-to-understand animations or videos. While (Emmorey et al., 2009) also noted eye-gaze at locative classifier constructions by both skilled and unskilled signers, the stimuli in our study do not contain classifier constructions (complex signs that convey 3D motion paths or spatial arrangements).

Based on these prior studies, we hypothesize the following:

H1: There is a significant difference in signers' eye-movement behavior between when they view *videos* of ASL and when they view *animations* of ASL.

H2: There is a significant difference in signers' eye-movement behavior when they view animations of ASL *with* some facial expressions and when they view animations of ASL *without* any facial expressions.

H3: There is a significant correlation between a signer's eye movement behavior and the scalar *subjective scores* (grammatical, understandable, natural) that the signer assigns to an animation or video.

H4: There is a significant correlation between a signer's eye movement behavior and the signer *reporting having noticed* a facial expression in a video or animation.

H5: There is a significant correlation between a signer's eye movement behavior and the signer correctly answering *comprehension questions* about a video or animation.

Each hypothesis above will be initially examined in terms of the following two eye-tracking metrics: proportional fixation time on the face and transition frequency between the face and body/hands. Based on the results of H1, we will determine whether to consider video separately from animations for H3 to H5. Similarly, results from H2 will determine if animations with facial expressions are considered separately from animations without, for H3 to H5.

6.2 User Study

To evaluate hypotheses H1-H5, we conducted a user study, where participants viewed short stories in ASL performed by either a human signer or an animated character. In particular, each story was one of three types: a "video" recording of an ASL signer, an animation with facial expressions based on a "model," and an animation with a static face (no facial expressions) as shown in Fig. 29. Each "model" animation contained a single ASL facial expression (yes/no question, wh-word question, rhetorical question, negative, topic, or an emotion), based on a simple rule: apply one facial expression over an entire sentence, e.g. use a rhetorical-question facial expression during a sentence asking a question that doesn't require an answer. Additional details of the facial expressions in our stimuli appear in Chapter 4 and Appendix A.

As discussed in Chapter 4, an ASL signer wrote a script for each of the 21 stories[10], including one of six types of facial expressions. To produce the video stimuli, we recorded a second signer performing these scripts in an ASL-focused lab environment, as illustrated in

[10] Their codenames in the stimuli collection are: E1, E2, E4-E6, W2-W4, Y3, Y4, Y6, R3, R5, R7, N1-N3, and T3-T5. Appendix A provides more details on their script and comprehension questions.

Figure 17. Then another signer created both the model and no facial expressions animated stimuli by consulting the recorded videos and using some animation software (VCom3D, 2014). The video size, resolution, and frame-rate for all stimuli were identical.

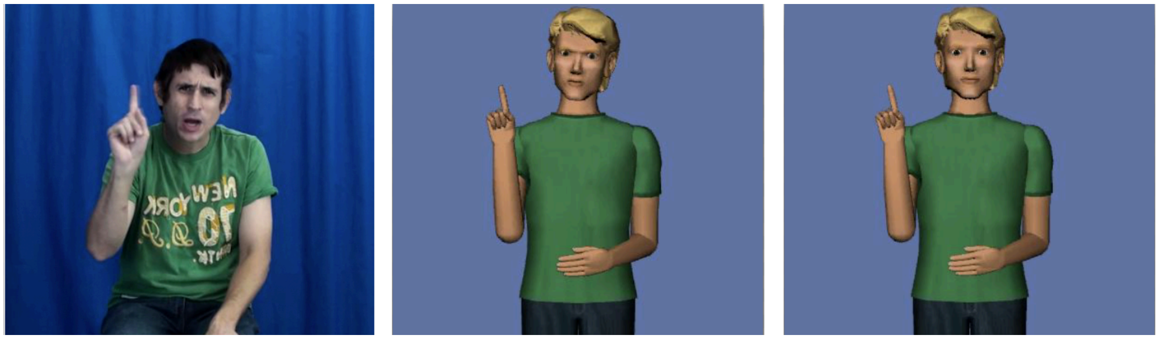
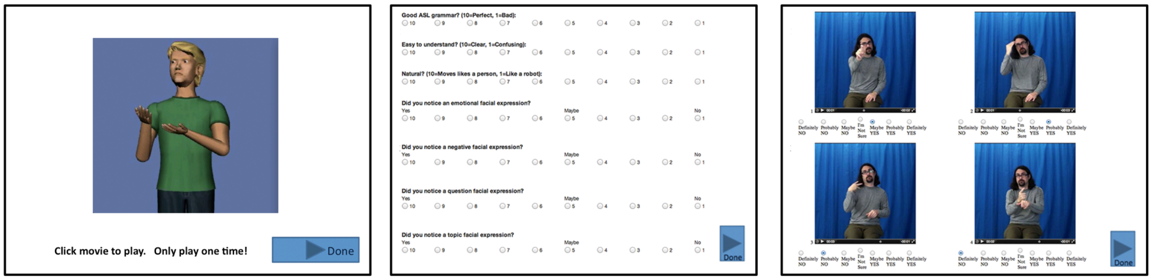


Fig. 29: Screenshots from the three types of stimuli: i) video of human signer, ii) animation with facial expressions, and iii) animation without facial expressions.

During the study, after viewing a story, each participant responded to three types of questions. All questions were presented onscreen (embedded in the stimuli interface) as HTML forms, as shown in Fig. 30, to minimize possible loss of tracking accuracy due to head movements of participants between the screen and a paper questionnaire. On one screen, they answered 1-to-10 Likert-scale questions: three subjective evaluation questions (of how grammatically correct, easy to understand, and naturally moving the signer appeared) and a "notice" question (1-to-10 from "yes" to "no" in relation to how much they noticed an emotional, negative, questions, and topic facial expression during the story). On the next screen, they answered four comprehension questions on a 7-point Likert scale from "definitely no" to "definitely yes." Given that facial expressions in ASL can differentiate the meaning of identical sequences of hand movements (as discussed in Section 2.1), both stories and comprehension questions were engineered in such a way that the wrong answers to the comprehension questions would indicate that the participants had misunderstood the facial expression displayed (Chapter 4). E.g. the comprehension-question responses would indicate whether a participant had noticed a "yes/no question" facial expression or instead had considered the story to be a declarative statement.



Click movie to play. Only play one time!

Good ASL grammar? (10=Perfect, 1=Bad):
○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Easy to understand? (10=Clear, 1=Confusing):
○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Natural? (10=Moves likes a person, 1=Like a robot):
○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Did you notice an emotional facial expression?
Yes ○ 10 ○ 9 ○ 8 ○ 7 ○ 6 Maybe ○ 5 ○ 4 ○ 3 ○ 2 No ○ 1

Did you notice a negative facial expression?
Yes ○ 10 ○ 9 ○ 8 ○ 7 ○ 6 Maybe ○ 5 ○ 4 ○ 3 ○ 2 No ○ 1

Did you notice a question facial expression?
Yes ○ 10 ○ 9 ○ 8 ○ 7 ○ 6 Maybe ○ 5 ○ 4 ○ 3 ○ 2 No ○ 1

Did you notice a topic facial expression?
Yes ○ 10 ○ 9 ○ 8 ○ 7 ○ 6 Maybe ○ 5 ○ 4 ○ 3 ○ 2 No ○ 1

1 ○ Definitely NO ○ Probably NO ○ Maybe NO ○ I'm Not Sure ● Maybe YES ○ Probably YES ○ Definitely YES

2 ○ Definitely NO ○ Probably NO ○ Maybe NO ○ I'm Not Sure ○ Maybe YES ● Probably YES ○ Definitely YES

3 ○ Definitely NO ● Probably NO ○ Maybe NO ○ I'm Not Sure ○ Maybe YES ○ Probably YES ○ Definitely YES

4 ● Definitely NO ○ Probably NO ○ Maybe NO ○ I'm Not Sure ○ Maybe YES ○ Probably YES ○ Definitely YES

Fig. 30: An example of a stimulus used in the study: story, subjective and notice questions, and comprehension questions.

An initial sample story familiarized the participants with the experiment and the eye tracking system. All of the instructions and interactions were conducted in ASL; Likert scale questions were explained in ASL. Part of the introduction, included in the beginning of the experiment, and the comprehension questions were presented by a video recording of an ASL signer.

In this study a desktop-mounted eye-tracker (Applied Science Labs D6 system) is used, where the cameras and the illuminator are in a small device (placed directly below the computer screen that displays the visual stimuli). The participant is seated in front of the 19-inch computer screen (resolution 1440x900) at a typical viewing distance (with their eye approximately 60cm from the eye-tracker device). The participant is able to make head movements (up to 30cm) during the study and the eye-tracker software tracks the participant's head location and orientation to compensate for these movements.

6.3 Initial Analysis and Results

Ads were posted on New York City Deaf community websites asking potential participants if they had grown up using ASL at home or had attended an ASL-based school as a young child. Of the 11 participants recruited for the study: 7 learned ASL since birth, 3 learned ASL prior to age 4, and 1 learned ASL at age 8. This final participant attended schools for the deaf with instruction in ASL from age 8 to 18, and she uses ASL daily at home and at work. There were 4 men and 7 women of ages 24-44 (average age 33.4). We recorded eye-tracking data once for each story that was shown to participants (prior to the participant being asked Likert-scale or comprehension questions about the story). Because the eye-tracker could occasionally "lose" the pupil of the participant's eye during tracking (e.g., if the participant rubbed their face with their hand during the experiment), we needed to filter out any eye-tracking data in which there was a loss of tracking accuracy. Therefore, we decided to analyze only those recordings that meet both of these criteria:

- The eye-tracker was able to identify the participant's head and pupil location for at least 50% of the story time.
- The eye-tracker recorded that participant was looking at the video/animation for at least 50% of the time. (This criterion may fail if the participant looked away from the screen or there was a tracking calibration problem for that story. Eye-trackers must be calibrated periodically during use so that they know how the observed eye angles correspond to screen coordinates.)

The threshold values of 50% in these two criteria were selected after consulting a histogram of the relevant eye-tracking values to determine a natural boundary in the data. Applying these filtering criteria reduced the number of recordings from 231 to 181.

6.3.1 Areas of Interest in the Stimuli

Figure 31 illustrates how we defined the "Face" and "Hands" areas of interest (AOIs) for the videos of the human signer and the animations of the virtual character. Identical AOIs were used for the animations with or without facial expressions. Note that the region of the screen where the hands may be located could potentially overlap with where the face is located (signers may move their hands in front of their face when signing), but our AOIs are defined so that they do not overlap. We made the simplifying assumption that the face should take precedence, and that is why the Hands AOI has an irregular shape to accommodate the Face AOI. Thus, if a participant were looking at our signer's hands when they moved in front of the signer's face, we would count that moment of time as a "face" fixation. This is a limitation of our study, but it simplified the eye-tracking analysis, and we believe that it had a minimal effect on the results obtained, given that the signer's hands do not overlap with the face during the vast majority of signing. While the Face AOIs have different horizontal/vertical ratios to accommodate the different head shapes and movements of the signers, the area (length x width) of the Face AOI for the animated character is identical to the area of the Face AOI for the human. The human signer performed some torso movements when signing, such as bending forward slightly, therefore the region of the screen where his hands tend to occupy is a little lower compared to the animated character. So, we set the borders of the Hands AOI lower for the human signer; to preserve fairness, we kept the area of the two Hands AOIs as similar as possible. The area of the animation Hands AOI is 99.3% of the area of the video Hands AOI.

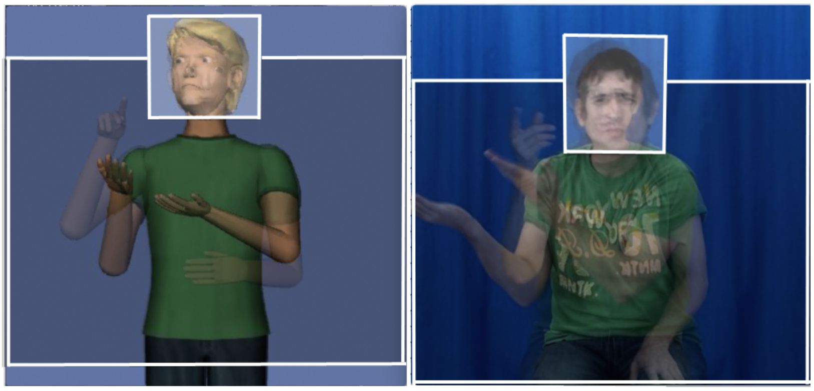


Figure 31: Screen regions for the face and hands AOIs of the animated character and the human signer.

6.3.2 Results and Comparisons

Section 6.1 discussed how we considered two eye-tracking metrics during our analysis:

- FacePFT: Proportional fixation time on the face AOI (i.e., total time of all fixations on the face AOI divided by story duration).
- TransFH: Number of transitions between the face AOI and hands/body AOI, divided by the story duration (in seconds).

Proportional fixation time and transition frequencies are typically not normally distributed, and Shapiro-Wilk tests confirmed this on the data collected in our study. For this reason, we used non-parametric tests of statistical significance (Kruskal-Wallis) and correlation (Spearman's Rho) during our analysis.

Figure 32 shows a box plot of the FacePFT values for the video, animation with facial expressions (“Model”), and animation without facial expression (“Non”). Box edges indicate quartiles, whiskers indicate minimum/maximum values (all with values of 0 or 1), and the centerline indicates the median. Stars indicate significant pairwise differences (Kruskal-Wallis, p<0.05).

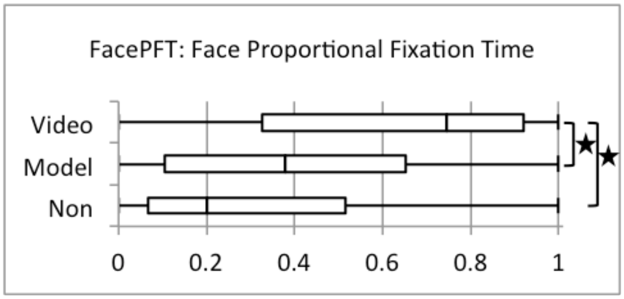


Figure 32: Proportional Fixation Time on the Face

In Figure 32, Hypothesis H1 was supported: participants spent significantly more time looking at the face AOI of the videos. We did not observe any support for H2: no significant difference was observed between the animations with and without facial expression. However, the median of the FacePFT values for Non, was (not significantly) lower than the median of the FacePFT values for Model. In a future study, we may record a larger number of participants to determine if the lack of significance here was due to an insufficient number of participants.

Figure 33 shows TransFH values. Note that due to the preponderance of zero values in the TransFH data, the boxes and whiskers of the plots are against the zero axis. These results show a similar (but inverse) pattern as the FacePFT values. When watching videos, participants moved their eyes between the face AOI and the body/hands AOI less frequently, than when watching animations. H1 was again supported: TransFH was lower for Video. H2 was not supported: no significant difference was observed between the animations with and without facial expressions (Model vs. Non).

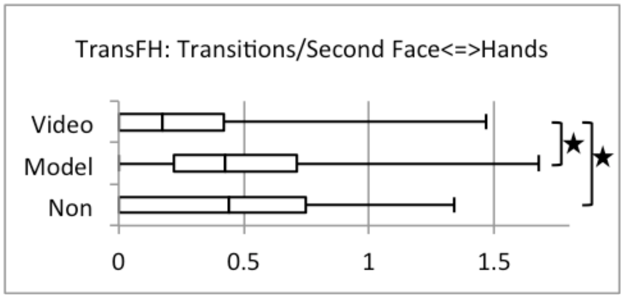


Figure 33: Transition Frequency Between the Face and Hands

Given that H1 was supported, when we are examining the results for hypotheses H3-H5, it is logical to consider the results for videos separately from animations. However, we will group the Model and Non videos together during the correlation analysis, since H2 was not supported. Table 6 displays the Spearman's *Rho* correlation values for FacePFT and TransFH. Values for which the p(uncorrelated) value is below 0.05 have been marked with an asterisk; the *Rho* is shown between the eye-metric and each of the responses recorded during the experiment:

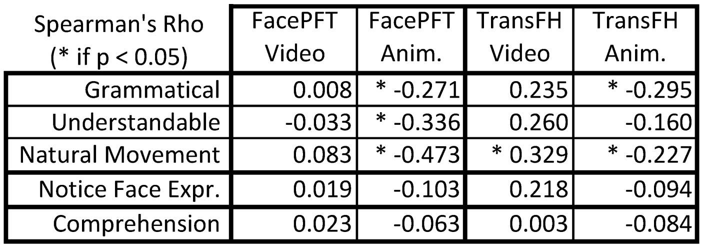
- Likert-scale subjective responses for whether the story was: grammatical, understandable, and had natural movement;
- Likert-scale response as to whether the participant noticed the particular facial expression in that story; and
- participant's accuracy on the comprehension questions.

Hypothesis H3 was supported for animations: there was a correlation between the subjective evaluation questions and the eye metrics. Specifically, FacePFT was significantly correlated with all three subjective scores, and TransFH, with grammaticality and naturalness of movement. The H3 results for videos are inconclusive: while TransFH was significantly correlated with naturalness of movement, it was not significant for the other two.

H4 was not supported by our results: none of the correlations were significant between eye metrics and the responses to the question about whether participants noticed the facial expression.

H5 was not supported by our results: There was no significant correlation between the eye metrics we examined and the accuracy of participants on comprehension questions.

Table 6: Correlations between Eye Metrics and Responses



Spearman's Rho (* if p < 0.05)	FacePFT Video	FacePFT Anim.	TransFH Video	TransFH Anim.
Grammatical	0.008	* -0.271	0.235	* -0.295
Understandable	-0.033	* -0.336	0.260	-0.160
Natural Movement	0.083	* -0.473	* 0.329	* -0.227
Notice Face Expr.	0.019	-0.103	0.218	-0.094
Comprehension	0.023	-0.063	0.003	-0.084

6.4 Extended Analysis and Results

The initial analysis (Section 6.3) did not reveal any significant correlation between the FacePFT and TransFH metrics and participants reporting having noticed a particular facial expression nor their comprehension questions scores (hypothesis H4 and H5) and the results were only partially supported for the scalar subjective scores (hypothesis H3).

One limitation of our initial analysis in Section 6.3 was that we did not distinguish between the upper (above nose) and lower face of the signer in the video. As discussed in Section 3.4.2, Muir and Richardson (2005) had distinguished between these parts of the face, and they found changes in proportional fixation time on the face of signers when the visual difficulty of videos varied. Since many grammatically significant ASL facial expressions consist of essential movements of the eyebrows, in this section, we separately analyze the upper and lower face.

A second limitation of our initial analysis in Section 6.3 was that we did not examine the path length of the eye gaze of the participant, which may have a relationship to the quality of the stimulus. In fact, Cavendar et al. (2005) had found a relationship between the path length of eye gaze the quality of *videos* of human signers (discussed in Section 3.4.2). In this analysis, we also measure a new metric, called Total Trail Distance, which is the aggregated distance between fixations normalized by the stimuli duration. Given that Emmorey et al. (2009) found that less skilled signers transitioned their gaze to the hands of the signer more frequently, we predict that there will be longer "trail lengths" of the eye gaze in our lower-quality animations, which are harder to understand.

6.4.1 Areas of Interest in Stimuli

In our initial analysis, only two areas of interest (AOIs) were considered for the analysis of participants' eye gazing behavior: "Face" and "Hands". In this section, we divide the "Face" AOI to "Upper Face" and "Lower Face" AOI based on the signers' nose-tip height. Figure 34 illustrates these areas of interest for the animations of the virtual character (with or without facial expressions) and for the videos of the human signer.

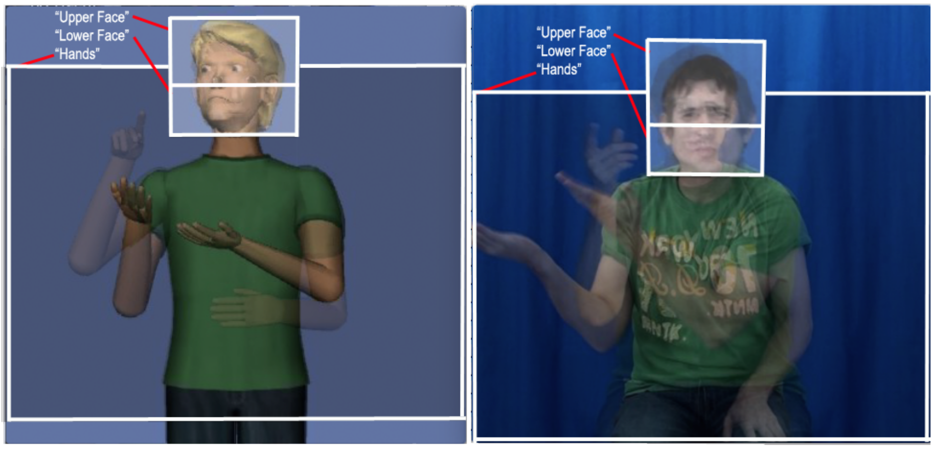


Figure 34: Screen regions for the upper face, lower face, and hands AOIs.

The AOIs were defined identically for all animations (with and without facial expressions). While the area (width x height) of the face AOIs were preserved, the vertical-horizontal ratio was slightly different for human videos: The human would often bend forward slightly, therefore the region of the screen where his head tend to occupy is a little lower compared to the animated character. So, we set the nose-tip line slightly lower for the human signer; to preserve fairness, we kept the area of the "Upper-Face" and "Lower-Face" AOIs as similar as possible between the animated character and human signer (97.6% for the upper and 102.6% for the lower portion).

6.4.2 Results and Comparisons

This section presents the results of the eye-tracking data analysis from the eleven participants, and the discussion is structured around three types of metrics:

- Transition frequency (i.e., the number of transitions between pairs of AOIs, divided by story duration in seconds) between the upper-face AOI and the hands-body AOI and between lower-face AOI and hands-body AOI.

- Proportional Fixation Time on the upper-face AOI or on the lower-face AOI (i.e., the total time of all fixations on the AOI, divided by story duration)
- Time-Normalized Total Trail Length (i.e., the sum of the distances between all of the participant's fixations, divided by the story duration in seconds).

Transition frequencies are displayed as a box plot in Figure 35, with the min/max values indicated by whiskers, quartiles by the box edges, and median values by a center line (not visible in Figure 35(a) because the median value was zero). On the basis of Kruskal-Wallis tests, significant differences are marked with stars ($p<0.05$). The three groups displayed include "Video" of a human signer, a "Model" animation with facial expressions, and a "Non" animation with no facial expressions. In Figure 35, there was a significant difference between the transition frequency between upper-face and body-hands, comparing Video and animations (Model + Non), which supports hypothesis H1, but not hypothesis H2.

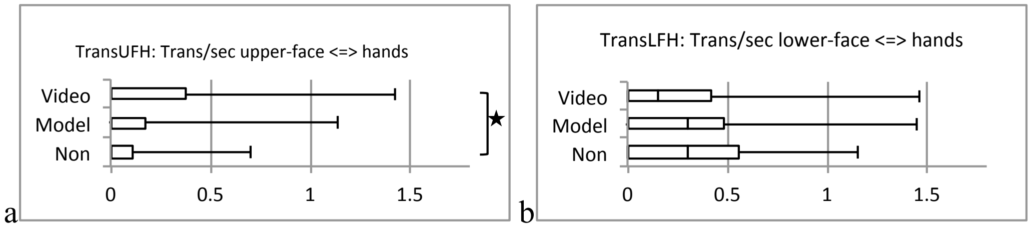


Figure 35: Transitions per second between: (a) the hands-body AOI and the upper-face AOI ("TransUFH") and (b) the hands-body AOI and the lower-face AOI ("TransLFH").

In order to better understand where participants were looking during the videos or animations, we also calculated the proportion of time their eye fixations were within the upper-face or lower-face AOIs; the results are shown in Figure 36. In this case, a significant difference was shown between Video and both types of animation (Model and Non) when considering the lower-face AOI in Figure 36(b). Only the pair Video vs. Non was significantly different when considering the upper-face AOI in Figure 36(a).

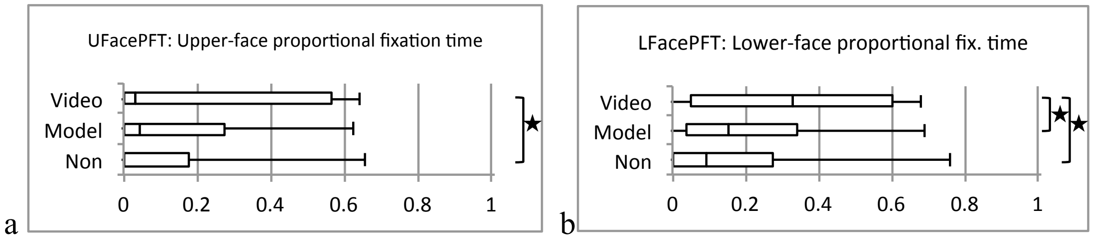


Figure 36: Proportional fixation time on: (a) the upper-face AOI (labeled as "UFacePFT") and (b) the lower-face AOI (labeled as "LFacePFT").

Since H2 was not supported, Model and Non were grouped together when calculating correlations to investigate Hypotheses H3, H4, and H5. Spearman's Rho was calculated, with significant correlations ($p<0.05$) marked with stars in Figure 37. Overall, the metrics using the upper-face AOI were more correlated to participants' responses to questions about the animations; most notably, Figure 37(a) shows significant correlations between the proportional fixation time on the upper-face AOI ("UFacePFT") and participants' responses to Likert-scale subjective questions in which they were asked to rate the grammaticality, understandability, and naturalness of movement of the animations. This result supports hypothesis H3 for animations, but not for Videos of human signers. No significant correlations were found between the eye metrics and the other types of participants' responses: questions about whether they noticed facial expressions and comprehension questions about the information content of videos or animations. Based on these results, H4 and H5 were not supported.

a

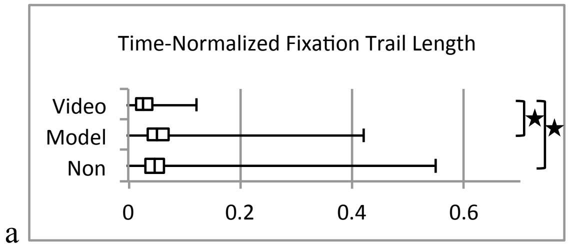
Spearman's Rho (* if p < 0.05)	UFacePFT Video	UFacePF T Anim.	TransUFH Video	TransUFH Anim.
Grammatical	0.149	* -0.340	0.166	* -0.305
Understandable	0.056	* -0.346	0.161	-0.145
Natural Movement	0.073	* -0.402	0.191	* -0.213
Notice Face Expr.	0.060	-0.101	0.058	-0.099
Comprehension	-0.001	-0.086	-0.064	-0.090

b

Spearman's Rho (* if p < 0.05)	LFacePFT Video	LFacePFT Anim.	TransLFH Video	TransLFH Anim.
Grammatical	0.087	-0.092	0.189	-0.090
Understandable	0.147	-0.156	0.217	-0.660
Natural Movement	0.093	* -0.215	* 0.277	-0.029
Notice Face Expr.	0.023	-0.239	0.198	-0.003
Comprehension	-0.018	-0.047	-0.030	0.027

Figure 37: Correlations between participants responses (rows) and eye metrics (columns), including proportional fixation time and transition frequency for upper-face and lower-face.

The final eye metric considered in this analysis is the time-normalized total trail length, which is shown in Figure 38. There was a significant difference between Video and both types of animation (Model and Non) in Figure 38(a), further supporting hypothesis H1. The correlations between this metric and the participants' responses are shown in Figure 38(b). This metric had significant correlations with the greatest number of types of participant responses, as indicated by the stars in Figure 38(b). While there was still no support for hypotheses H4 or H5, based on the results in Figure 38(b), hypothesis H3 was supported for both videos of human signers and animations of virtual humans.



b

Spearman's Rho (* if p < 0.05)	TrailLen Video	TrailLen Anim.
Grammatical	* 0.369	0.096
Understandable	* 0.378	* 0.227
Natural Movement	* 0.440	* 0.292
Notice Face Expr.	0.119	0.084
Comprehension	-0.035	-0.035

Figure 38: Fixation trail length for each type of stimulus (a) and correlations to responses (b).

6.5 Conclusions and Recommendations for Future Researchers

This chapter has identified how eye-tracking metrics are related to participants' judgments about the quality of ASL animations and videos. We have investigated and characterized differences in participants' eye-movement behavior when watching human videos or virtual-human animations of ASL. The results of our user study are useful for future researchers who wish to measure the quality of ASL videos or animations: eye-tracking metrics that can serve as complimentary or alternative methods of evaluating such stimuli. These metrics can be recorded while participants view stimuli, without asking them to respond to subjective or objective questions, providing flexibility to researchers in designing experimental studies to measure the quality of these stimuli.

To investigate RQ4 and RQ5 in this thesis, we formulated five hypotheses and conducted a user study. H1 and H2 hypothesized that there is a significant difference in signers' eye-movement behavior between when (a) they view *videos* of ASL and when they view *animations* of ASL, and (b) they view animations of ASL *with* some facial expressions and when they view animations of ASL *without* any facial expressions, respectively. H3, H4, and H5 hypothesized that there is a significant correlation between a signer's eye movement behavior and (a) the scalar *subjective scores* (grammatical, understandable, natural) that the signer assigns to an animation or video, (b) the signer *reporting having noticed* a facial expression in a video or animation, and (c) the signer correctly answering *comprehension questions* about a video or animation, respectively.

We found that hypotheses H1 and H3 were supported and hypotheses H2, H4, and H5 were not supported. There was a significant difference in the eye movement metrics when participants viewed ASL videos (as compared to when they viewed ASL animations), and some eye movement metrics were significantly correlated with participants' subjective judgments of video and animation quality (grammaticality, understandability, and naturalness of movement).

Specifically, the most notable findings in this chapter are:

- If using proportional fixation time to distinguish between ASL videos and animations, the upper-face AOI should be considered; if using transitions/second, the lower-face AOI should be considered. Since our initial analysis had not analyzed the eye-tracking data in such a fine-grained manner (i.e., the upper-face and lower-face AOIs had been clumped together into a single "face" AOI), this distinction between them in regard to the significance of transitions per second or proportional fixation time was not initially identified.

- If seeking an eye metric that correlates with participants' subjective judgments about ASL videos or animations, the time-normalized fixation trail length metric (described in this chapter) should be utilized. (The only exception would be for predicting participants' grammaticality judgments for ASL animations: the upper-face proportional fixation time was the best correlated.) These animation results for H3 may be the most useful finding in this chapter for future researchers; this is the first published result that indicates a relationship between eye-tracker metrics and participants' subjective judgments of sign language animation quality.

In short, the results presented in this chapter indicate that eye tracking analysis is valid for use as a complimentary form of measurement in a user-study to evaluate animations of sign language. Researchers who are studying computer graphics issues relating to the appearance of a virtual human for sign language animations and who are interested in obtaining participants' responses to subjective evaluations of the animation-quality may use eye-gazing metrics as an alternative form of measurement. This may be useful in experimental contexts in which the researchers cannot (or prefer not) to interrupt participants with questions while they are viewing a sequence of ASL animations. Additionally, researchers could directly compare eye movement of the participants between videos (that would serve as an “ideal” of photorealism) and their animations. If researchers obtain eye metric results that are similar for both videos of human signers and for their ASL animations, this may serve as evidence that their ASL animations are high-quality.

Sign language animation researchers who are considering using eye-tracking approaches with deaf users should consider some practical issues: First, they should minimize the need for the participants to look away from the screen during the experiment, to reduce eye tracking data loss and promote accuracy. Unlike hearing subjects who may ask questions and receive answers without taking their eyes off the computer screen, deaf participants would need to look away from the stimulus to communicate with the researcher. We recommend: (i) embedding the instructions and the questionnaires in the stimuli application, (ii) familiarizing the participants with a sample case initially, and (iii) positioning the ASL-signing researcher giving instructions to the participant opposite to the participant and behind the screen. If the researcher is at the participant’s side, the participant may tend to shift their head towards the researcher occasionally during the study, to monitor for communication or confirmation. Second, a delicate balance is needed when selecting the size of the video/animation on the computer screen. While a bigger video/animation permits for fine-grained (distinct) AOIs, the participant should be able to see the human/animated character in full, without the need for head movements. Also, when stimuli are so large that they approach the edges of the computer screen, there can be a loss in eye-tracker accuracy: when a participant’s eye is rotated farther from its neutral position, some eye-trackers “lose” the pupil or see reflection artifacts on the white sclera of the eye.

Chapter 7 Demographics and Other Factors Influencing Acceptance of ASL Animation[11]

Researchers generally evaluate the quality of their software by: generating animations using some current version of their software, setting up an experiment in which deaf participants view and evaluate the animations, and comparing the scores of animations produced using the software (to some baselines or to prior versions). However, the field lacks consensus about the set of demographic data that should be reported about the participants. Thus, it is difficult to compare the results across studies because some variation in comprehension or subjective evaluation scores in studies may be explained by demographic characteristics of the participants, rather than by true differences in the quality of the animations being evaluated.

In this chapter we examine the use of demographic and technology-experience variables as predictors of participants' responses to (a) subjective measures of animation quality and (b) objective measures of comprehension of the content. We present a study in which ASL signers were shown ASL animations (using a variety of avatars) and were asked questions of type (a) and (b). In addition, participants were asked questions about: (i) demographic characteristics and (ii) their technology experience/attitudes. Multiple regression analysis was used to determine whether independent variables (i-ii) relate to participants' responses (a-b).

[11] This chapter describes joint work with Professor Matt Huenerfauth, and graduate students working at LATLAb: Sarah Ebling, Kasmira Patel, Mackenzie Willard (Kacorri et al., 2015).

7.1 Collecting Independent Variables

The goal of our work is to examine whether metrics relating to participants' demographics (e.g., age, gender) or technology experience/attitudes can explain some of the subjective-judgment and comprehension-question scores collected in experiments to measure the quality of sign language animation systems. This section explains the design of our questionnaire for recording these independent variables, which will be used in our multiple regression models in Section 7.2.2. This section will also explain the origin of any questions that were adapted from survey instruments that were presented in prior work of other authors, e.g. (Rosen et al., 2013).

While some researchers have explored the design of fully online surveys of deaf users containing both ASL and English, e.g. (Tran et al., 2010), our survey was conducted in-person, with a human signer asking questions in ASL on a laptop screen and a paper answer sheet (with questions redundantly appearing in English, to aid the participant in aligning the video and paper). Given that our study included hard-of-hearing participants, the inclusion of English was considered important, and given our aim to include older participants in the study, a "low tech" paper answer sheet was considered preferable. Many of our questions were adapted from pre-existing English surveys (Section 3.5); so a professional ASL interpreter (with a bachelor's degree in interpreting and postgraduate coursework in information technology) translated items into ASL. Deaf members of the research team checked the videos for fluency and that subtleties of meaning were preserved. Several takes of each question were recorded so that we could select the best version for the questionnaire. Example videos appear here: http://latlab.ist.rit.edu/assets2015/.

7.1.1 Demographic Questions

We selected demographic questions by assembling items that were asked in our prior experimental studies, e.g. Chapter 4, and questions asked in studies surveyed in Section 3.5. Below, we list the demographic questions, preceded by the "codename" of the response variables used in our regression models in Section 7.2.2.

- **Gender**: What is your gender? (male, female, other)
- **Age**: How old are you?
- **Describe**: How do you describe yourself? (deaf/Deaf, hard-of-hearing, hearing, other)
- **WhenBecome**: At what age did you become deaf/Deaf or hard-of-hearing? (Note: No hearing participants were in this study.)
- **WhenLearn**: At what age did you begin to learn ASL? (Note: all participants in this study were ASL signers.)
- **ParentsAre**: Are your parents deaf/Deaf? (yes, no)
- **ParentsUse**: Did your parents use ASL at home? (yes, no)
- **SchoolType**: What type of school did you attend as a child? (residential school for deaf students, daytime school for deaf students, or a mainstream school)
- **SchoolASL**: Did you use ASL at this school? (yes, no)
- **Education**: Which describes your current level of education? (did not graduate high school, graduated high school, graduated college, have bachelor's degree, have graduate degree)
- **HomeASL**: Do you use ASL at home? (yes, no)

- **HomeEnglish**: Do you use English at home? (yes, no)
- **WorkASL**: Do you use ASL at work? (yes, no)
- **WorkEnglish**: Do you use English at work/school? (yes, no)

Note: After collecting data from participants (Section 7.2.1), we noticed a gap in the Age range 35-42 so instead of treating Age as a continuous variable, we binned it into three groups: 18 to 24, 25 to 34, and 43 to 59, and we relabeled the variable as AgeGroup.

7.1.2 Technology Experience and Attitudes

To measure participants' frequency of technology use, we adopted the InternetSearch and MediaSharing subscales from the Media and Technology Usage and Attitudes Scale (Rosen et al., 2013); scoring is based on the participant's response (e.g., Never, Monthly, Weekly, Once a day, etc.) to how frequently they engaged in various activities (listed below) on computers, laptops, tablets, or mobile phones:

- **InternetSearch**: Search the Internet for news. …for information. …for videos. …for images or photos.
- **MediaSharing**: Watch TV shows, movies, etc. Watch video clips. Download media files from other people. Share your own.

Using the same scoring, we created an ASLChat subscale:

- **ASLChat**: Have a signing (ASL) conversation with someone using a video phone. Have a signing (ASL) conversation with someone using a computer, laptop, tablet, smartphone.

We asked participants to indicate how often they played video games by selecting one of three frequency ranges (below), which we coded as "advanced," "intermediate," and "beginner."

- **GameGroup:** How often do you play games on a computer, game console, or phone? (several times a day, between once a day and once a week, less than once a week)

Next, participants were asked about their perceptions of the benefits of technology, using the PositiveAttitudes subscale of (Rosen et al., 2013), in which the score is the average of responses to individual statements listed below (Strongly agree = 5, Agree = 4, Neither agree no disagree = 3, Disagree = 2, Strongly disagree = 1):

- **PositiveAttitudes**: It is important to be able to find any information whenever I want to online. It is important to be able to access the Internet any time I want. It is important to keep up with the latest trends in technology. Technology will provide solutions to many of our problems. With technology anything is possible. I accomplish more because of technology.

Participants' impression of computer complexity was measured using two Computer Questionnaire questions from the October 2014 PRISM survey (CREATE, 2015), using identical Likert scoring as above.

- **ComputerComplex:** Computers are complicated. Computers make me nervous.

Finally, at the end of the questionnaire, users were asked a series of questions to evaluate their overall attitude of the usefulness of ASL animations in a variety of contexts. They were also asked if they had previously seen computer animations of ASL:

- **AnimationAttitude**: Computer animations of sign language could be used to give information on a website. Computer animations of sign language could be used to give information in a public place (e.g., airport, train station). Computer animations of sign language could be used as an interpreter in a face-to-face meeting. Computer animations of sign language could be used as an interpreter for a telephone relay. I would enjoy using computer animations of sign language. Other people would enjoy using computer animations of sign language.
- **SeenBefore**: Before today, had you ever seen a computer animation of sign language? (yes, no)

7.2 Collecting Dependent Variables

Section 3.5.2 described how related work, and in particular (Kipp et al., 2011b), displayed animations of multiple sign languages and animations that were hand-animated; and explained why we decided to display only **synthesized** animations of **ASL** in our current study. However, there is one type of "diversity" from (Kipp et al., 2011b) that we did preserve in our study design: We wanted the results of this study to be applicable to a variety of ASL signing avatars, with different appearance, rendering technologies, automation capabilities, and motion synthesis. Thus, we decided to display animations of three avatars synthesized by different state-of-the-art animation platforms (Jennings et al., 2010; our new virtual human platform in Chapter 11; and Vcom3D, 2015).

In addition, in the (Kipp et al., 2011b) study, each avatar performed a different message. To control for this in our study, we selected three short ASL stories from our stimuli and comprehension question collection (Chapter 4). Specifically, we selected three stimuli including a negative facial expression, a yes-no question, and a wh-question with codenames N2, W2, and Y3[12], respectively. These stimuli had been rated as being the most understandable when comparing two animation platforms (Section 11.4). Example stimuli from the current study may be viewed here: http://latlab.ist.rit.edu/assets2015/.

EMBR: Animations shown in Figure 1(a) were generated using the open source EMBR platform (Heloir et al., 2011), which we have extended with ASL handshapes and detailed upper-face controls compatible with the MPEG-4 Facial Animation standard (Chapter 11). The hand movements of the avatar were created by ASL signers, who selected key-poses to define each sign in the lexicon. Facial expressions and head movements were automatically driven by the video recordings of an ASL signer performing the stimulus and are part of the initial pilot data collected in this thesis (Section 10.2). To extract the MPEG-4 facial features and head pose of human signers, we used an automatic face tracking software (Section 10.1), whose output was converted to EMBR script with our MPEG4-to-animation pipeline (Section 11.2.2)

JASigning: Animations shown in Figure 1(b) were produced using the free Java Avatar Signing (JASigning) system (Jennings et al., 2010). All signs in the stimuli were notated in the Hamburg Notation System (HamNoSys, Prillwitz et al., 1989) by a Deaf researcher while consulting the video recordings of an ASL signer performing the stimulus. HamNoSys, which serves as an input for JASigning, has around 200 symbols describing the components: handshape, hand position, location, and movement. Information about the non-manual

[12] More details are available on Appendix A.

components (e.g., eyebrow movement and head movement) is included in the SiGML code (Hanke, 2001), an XML representation for HamNoSys, but time-alignment of non-manuals with the manual signs requires careful manual adjustment, e.g. (Ebling and Glauert, 2015).

VCOM: Animations shown in Figure 1(c) were generated using the commercially available ASL authoring tool, VCom3D Sign Smith Studio (Vcom3D. 2015), which allows users to produce animated ASL sentences by arranging a timeline of animated signs from a prebuilt or user-defined vocabulary. The software includes a library of facial expressions that can be applied over a single sign or multiple manual signs. Both the hand movements and facial expressions of the avatar for the three stimuli were created by ASL signers at a key-pose level. The VCOM and EMBR animations shared similar hand movements.

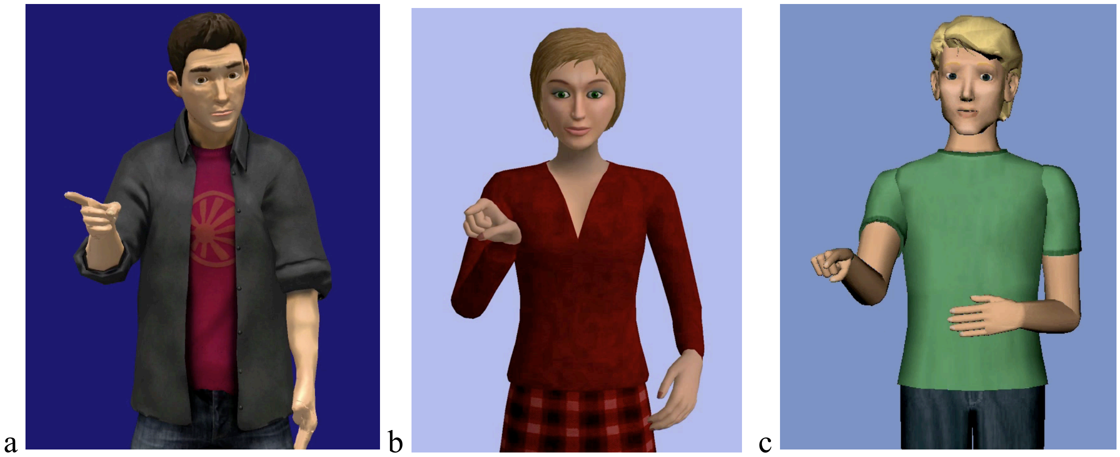


Figure 39: Screenshots from the three avatars shown in the study: (a) EMBR, (b) JASigning, (c) VCom3D.

During our study, after participants answered the demographic and technology-experience questions described in Section 7.1, they viewed a sample animation, to become familiar with the experiment setup and the questions they would be asked about each animation. (This sample animation used a different avatar than the other animations shown in the study.) Next, after viewing each of the three main animations, an onscreen video of an ASL signer asked participants four fact-based comprehension questions about the information conveyed in the animation. Participants responded to each question on a 7-point scale from "definitely no" to "definitely yes." As described in Chapter 4, a single "Comprehension" score for each animation can be calculated by averaging the scores of the four questions.

Next, the participants were asked to respond to a set of questions that measured their subjective impression of the animation, using a 1-to-10 scalar response. Each question was conveyed using ASL through an onscreen video, and the following English question text was shown on the questionnaire:

(a) Good ASL grammar? (10=Perfect, 1=Bad)

(b) Easy to understand? (10=Clear, 1=Confusing)

(c) Natural? (10=Moves like person, 1=Like robot)

(d) Was the signer friendly? (10=Friendly, 1=Not)

(e) Did you like the signer? (10=Love it, 1=Hate it)

(f) Was the signer realistic? (10=Realistic, 1=Not)

Questions (a-c) have been used in many of our prior experimental studies and were included in the collection of standard stimuli and questions that was released to the research community (with details available in Chapter 4 and Appendix A). Questions (d-f) were inspired by (Kipp et al., 2011b). To calculate a single "Subjective" score for each animation, the scalar response scores for the six questions were averaged.

7.2.1 Recruiting and Data Collection

Prior research, e.g. (Huenerfauth et al., 2008), has discussed the advantages of having Deaf researchers conduct experimental studies in ASL. In this study, a Deaf researcher (co-author) and two Deaf undergraduate students (ASL signers) recruited and collected data from participants, during meetings conducted in ASL. Potential participants were asked if they had grown up using ASL at home or had attended an ASL-based school as a young child. Initial advertisements were sent to local email distribution lists and Facebook groups. Our study (N=62) was completed during a four-week data collection period, a short timeframe made possible due to the many people who are deaf and hard-of-hearing associated with our university or in our city. It was easier for us to identify younger participants (especially college-aged students); the process of recruiting older participants took additional time and effort. The research team used personal contacts in the Deaf community to identify participants, especially older adults, who were less likely to be recruited through electronic methods. The advertisement included contact information for a Deaf researcher, including an email address, videophone, and text messaging (mobile phone). Research team members also attended local Deaf community events (e.g., the Deaf Club) to advertise the study.

Researchers met participants around the city to conduct the 70-minute survey, using a laptop with video questions in ASL. Of the 62 participants recruited for the study, 43 participants learned ASL prior to age 5, 16 had been using ASL for over 9 years, and the remaining 3 learned ASL as adolescents, attended a university with classroom instruction in ASL, and used ASL daily to communicate with a significant other or family member. There were 39 men and 23 women of ages 18-59 (avg. 25.73). For participants over age 43 (avg. 53.14), there were 4 men and 2 women who learned ASL prior to age 9, 5 self-reported to be deaf/Deaf and hard-of-hearing.

7.2.2 Analysis and Results

The goal of our analysis was to examine how demographic factors relate to participants' responses to subjective and comprehension questions about ASL animations. In addition, we wanted to know whether variance in scores could be explained by participants' technology experience and attitudes. We therefore used multiple regression to analyze the data. Our independent variables included all of the "Demographic" and "Technology" metrics, listed in Section 7.1. Our dependent variables included the "Comprehension" and "Subjective" scores described in Section 7.2.1. Many researchers, e.g., (Crabb and Hanson, 2014), follow the recommendation of (Gelman, 2008) that continuous-value variables be normalized by dividing the individual participant metrics by two times the group standard deviation, to facilitate easier comparison among coefficients of scalar and binary predictors. We have followed this procedure for all of the continuous independent variables in this study.

We trained two separate models for each of our dependent variables (Subjective and Comprehension): Model 1 was based upon Demographic variables only, and Model 2 was based upon both Demographic and Technology variables. The rationale for this choice is that while some prior authors have reported limited Demographic data about the participants in their studies, the set of Technology questions presented in this study is novel. Since we had recorded many Demographic and Technology variables (Section 7.1), it was important to explore combinations of variables in a systematic manner. We used the ‘leaps’ package (Lumley, 2009) to build models of all possible subsets of features to identify the model with the highest adjusted R-squared value, which indicates what percent of the total variability is accounted for by the model. For Model 1, the input to ‘leaps’ was all Demographic variables only. For Model 2, the input to ‘leaps’ was all Demographic and all Technology variables. For all models, we evaluated the collinearity of the independent variables (that were selected by ‘leaps’) by verifying that their variance-inflation was less than 2 (Fox and Monette, 1992).

Table 7: Multiple Regression Model – Comprehension

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	Estimate	*Std. Error*	*t score*
Model 1: Demographic	**Model 1:** Adj. R^2=0.256 (p<0.005)		
AgeGroup[25,34]	-0.344	0.195	-1.768 .
AgeGroup[35,)	-0.094	0.207	-0.452
Describehard-of-hearing	-0.242	0.149	-1.629
WhenBecome	0.204	0.126	1.624
WhenLearn	0.164	0.152	1.081
ParentsAreyes	0.252	0.166	1.516
SchoolASLyes	0.336	0.183	1.838 .
HomeASLyes	-0.177	0.147	-1.204
WorkEnglishyes	0.292	0.152	1.923 .
SchoolTypeMainstream	-0.092	0.146	-0.630
SchoolTypeResidential	0.575	0.169	3.407 **
Model 2: Demographic & Tech.	**Model 2:** Adj. R^2=0.382 (p<0.0001)		
Gendermale	0.273	0.126	2.168 *
Describehard-of-hearing	-0.317	0.135	-2.338 *
WhenBecome	0.217	0.117	1.857 .
HomeASLyes	-0.207	0.125	-1.655
SchoolTypeMainstream	-0.029	0.140	-0.208
SchoolTypeResidential	0.662	0.151	4.380 ***
InternetSearch	-0.493	0.140	-3.513 ***
PositiveAttitudes	0.249	0.118	2.105 *
ASLChat	0.181	0.129	1.402
GameGroupBeginner	-0.307	0.129	-2.377 *
GameGroupIntermediate	-0.283	0.202	-1.399
SeenBeforeyes	0.162	0.119	1.355

Table 7 summarizes the regression analysis for Comprehension, where ‘Estimate’ column reports the regression coefficient for the variable, i.e. how the output varies per unit change in variable, ‘Std. Error’ refers to how wrong the model is on average using the variable units, where smaller values indicate that the observations are closer to the fitted line, and ‘t score’ is the test statistic to calculate the p-value for significance testing. In Model 1 (demographic variables only), the type of school that the participant attended had the largest coefficient (see “Estimate” column): attending a residential school for deaf students had a positive relationship with the participant’s success at answering comprehension questions. Model 2 contained both demographic and technology variables, and a relationship between SchoolType and Comprehension is still present. Gender, Describe, InternetSearch, PositiveAttitudes, and GameGroup were also key components of Model 2. This suggests that when considering the results of studies that evaluate participants’ comprehension of synthesized ASL animations, some variance in participants’ scores can be explained by demographic and technology characteristics of each participant, e.g., their use of the Internet, positive attitude towards technology, and video game exposure. (Section 7.3 includes additional discussion of these factors.)

Table 8: Multiple Regression Model – Subjective

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	Estimate	*Std. Error*	*t score*
Model 1: Demographic	**Model 1:** Adj. R^2=0.153 (p< 0.02)		
Gendermale	-0.527	0.501	-1.05
Describehard-of-hearing	0.652	0.576	1.13
WhenLearn	-0.834	0.542	-1.54
HomeASLyes	-1.557	0.591	-2.63 *
SchoolTypeMainstream	0.659	0.584	1.13
SchoolTypeResidential	-0.538	0.643	-0.84
Model 2: Demographics & Tech.	**Model 2:** Adj. R^2=0.335 (p<0.0001)		
WhenLearn	-0.589	0.486	-1.21
HomeASLyes	-1.431	0.499	-2.87 **
SchoolTypeMainstream	0.685	0.517	1.32
SchoolTypeResidential	-0.030	0.590	-0.05
ComputerComplex	0.628	0.426	1.48
MediaSharing	-1.491	0.448	-3.33 **
AnimationAttitude	-1.373	0.448	-3.07 **

Table 8 summarizes the regression analysis for Subjective scores. In Model 1 (demographic variables only), using ASL at home had a significant and downward effect on a participant’s subjective impressions. Using ASL at home was also a significant factor in Model 2, which includes both Demographic and Technology variables. Moreover, AnimationAttitude and MediaSharing were other key components of Model 2. These results suggest that when considering the results of studies that collect subjective judgments about synthesized sign language animations, researchers can expect harsher judgments from participants who use ASL at home, are comfortable with media sharing or downloading, and whose general attitude about sign language animations and their usefulness is not positive.

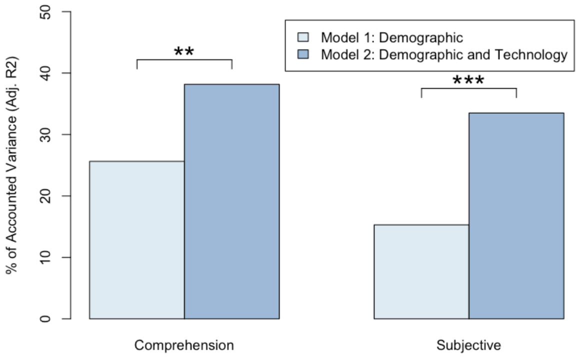


Figure 40: Regression model comparison summary. (Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01)

Figure 40 illustrates how Comprehension Model 2 accounts for significantly more variance than Comprehension Model 1, and the same is true for Subjective Model 2 and Subjective Model 1. An ANOVA was used to compare the models, and p-values are denoted in the graph by *** for p<0.001 or by ** for p<0.01. Model 2 represented a significant improvement in the amount of Comprehension accounted for between groups from 25.6% to 38.2%. Loosely speaking, this indicates that you can more accurately predict a signer’s success at answering comprehension questions by considering both their demographic characteristics and technology experience/attitudes, rather than relying on their demographic characteristics only. Similarly, there was a significant increase in accounted variance of participants’ subjective impressions of the animations from 15.3% to 33.5%.

7.3 Discussion of the Relative Importance of the Factors

Henceforth, our discussion will focus only on the best performing models: Comprehension Model 2 and Subjective Model 2, which contained both Demographic and Technology variables. In Section 7.2.2 we considered each variable’s coefficient (“Estimate” column in Table 7 and Table 8) to roughly identify those with large influence. However, coefficients are sensitive to the “order” in which the variables are considered in the model. For more meaningful interpretation, we calculated the relative importance of each of the variables in Comprehension Model 2 and Subjective Model 2, using the Linderman-Merenda-Gold (LMG) metric (Lindeman et al., 1980), calculated using the ‘relaimpo’ package (Grömping, 2006). This analysis assigns an R-squared percent contribution to each correlated variable obtained from all possible orderings of the variables in the regression model. Higher bars in Figure 41 indicate that the variable had greater importance in the model. We employed bootstrap to estimate the variability of the obtained relative importance value, to determine 95% confidence intervals (shown as whiskers in Figure 41). Importance values may be considered significant when a bar’s whiskers do not cross the zero line in the graph.

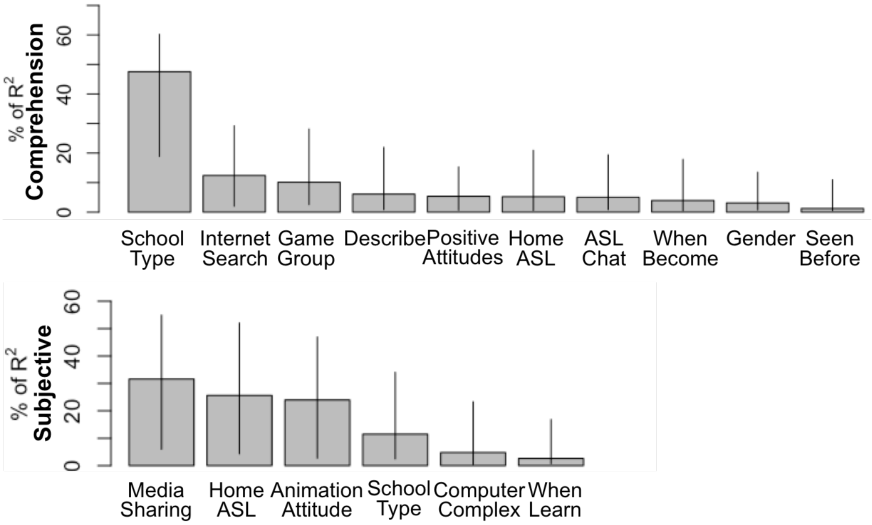


Figure 41: Relative importance (normalized to sum to 100%) of factors in Comprehension Model 2 and in Subjective Model 2, with 95% bootstrap confidence intervals.

For Comprehension Model 2, which contains variables that ‘leaps’ selected through an exhaustive search of all subsets of Demographic and Technology variables, we observe that the variables with highest and significant relative importance are SchoolType, InternetSearch, and GameGroup. Given the much higher relative importance of SchoolType, compared to the other variables, we focus on this variable in our discussion below:

Comprehension and SchoolType. As discussed in the analysis section, attending a residential school seems to have a significant positive relationship with a participant's comprehension-question scores for synthesized ASL animations. We therefore encourage sign language animation researchers to include this variable in their demographic questionnaire for each study and to report this characteristic of participants in publications. When evaluating the Comprehension scores for their animations, they should consider this factor when comparing their results to those for other studies (whose participant pools may have differed in this characteristic).

Comprehension and SeenBefore. Another aspect Figure 41 that may be of interest to sign language animation researchers is the low importance of the SeenBefore variable in this model. Prior exposure of a participant to signing avatars did not explain much variance in participants' Comprehension scores. For researchers who conduct user studies with deaf participants to frequently evaluate the progress of their animation software, this finding suggests that participants who have seen prior versions of their animation system may be re-recruited for future studies (with the caveat, of course, that the new study is showing different stimuli). Since there may be a relatively small local Deaf community nearby to some research groups, this is a useful finding. We note that in this study, we had a well-balanced sample of participants for the SeenBefore variable (yes=29, no=33).

For Subjective Model 2, containing variables that 'leaps' selected through an exhaustive search of all subsets of Demographic and Technology variables, we observe that the variables with the highest and significant relative importance are: MediaSharing, HomeASL, AnimationAttitude, and SchoolType. While the height of its bar in Figure 41 indicates each variable's importance, the direction of the relationship (positive/negative) is indicated by the sign of the coefficient in the "Estimate" column of Table 8.

Subjective and AnimationAttitude. A positive relationship exists between these two variables, which is not a surprising result: If a participant has an overall negative view of the usefulness or likeability of sign language animations in general (as measured by the AnimationAttitude scale, Section 7.1.2), then it is intuitive why they might have lower subjective scores for a specific animation.

Subjective and MediaSharing. Intuitively, we had expected that users with greater technology experience might have higher subjective scores due to their possible enthusiasm for technology. On the contrary: MediaSharing had a negative relationship to participants' subjective scores for animations. We can speculate that users with higher technology experience might have "higher standards" for the acceptable level of quality in an animation.

Subjective and HomeASL. A participant using ASL at home was also a factor with a negative relationship to their subjective score. We speculate that this might also be a case of "higher standards"; frequent ASL users may be harsher critics of animation quality.

Subjective and SchoolType. While SchoolType was important in both Comprehension Model 2 and in Subjective Model 2, the direction of the relationship is reversed. Attending a residential school had a positive relationship with Comprehension scores, but it had a negative relationship with Subjective scores. We note that it is reasonable that an independent variable may have opposite relationship with each of our dependent variables: Prior research has found low correlation between a participant's subjective score for an animation and his/her comprehension score for it (Huenerfauth et al., 2008).

7.4 Conclusions and Recommendations for Future Researchers

Since the overall goal of this dissertation is to improve the state of the art of software for automatically synthesizing animations of sign language from a simple script of the desired message, Part I of the dissertation has examined how to best conduct studies to evaluate the quality of such software. Thus, the findings of the study in this chapter will influence the selection of demographic and technology experience/attitude questions we ask participants in the study presented in Chapter 13.

Furthermore, a contribution of this work is a deeper understanding of the relationship between participant characteristics and evaluation scores in this field. Specifically, we found that the following variables were most important in explaining variance in comprehension and subjective scores of sign language animations:

- **SchoolType**: Assessed with a single multiple-choice question.
- **HomeASL**: Assessed with a yes/no question.
- **MediaSharing**: Assessed with four scalar response items indicating frequency of different activities, from (Rosen et al., 2013).

- **AnimationAttitude**: Assessed with six Likert agreement items.

While other variables were present in models presented in Section 7.2.2, the above four items correspond to the most important factors, and this abbreviated set may be useful for researchers who are interested in minimizing the amount of study time spent collecting demographic and technology experience/attitude data. Of course, we anticipate researchers will continue reporting other basic demographic data, e.g., age or gender, but based on the survey of prior work in Section 3.5.1, we know that few current sign language animation researchers regularly collect and report these four items.

To promote replicability and comparison of results across studies we have shared the questions and ASL videos that we used to measure the variables in this study, on our lab website: http://latlab.ist.rit.edu/assets2015/.

Through collection and publishing of these characteristics of study participants, we anticipate easier comparisons of research results across publications. We also believe that these factors would be useful for researchers to consider if they are balancing or matching participants across treatment conditions in a study.

Compared to prior non-online studies evaluating sign language animation, this study was relatively large (N=62). However, when conducting a regression analysis of factors, there is always an advantage in having even larger and more diverse participant sets. In this case, it would be useful to recruit more participants from the Deaf community in another geographic area, to ensure that the relationships observed in the current study are preserved.

Epilogue to Part I

In Part I of this thesis we conducted rigorous methodological research on how experiment design affects study outcomes when evaluating sign language animations with facial expressions. Our research topics involved: (i) stimuli design, (ii) effect of videos as upper baseline and for presenting comprehension questions, (iii) eye-tracking as an alternative to recording question- responses from participants, and (iv) participants' demographics and technology experience factors influencing responses.

Inventing stimuli that contain linguistic facial expressions and measure whether participants understand the intended information is challenging – but necessary for effectively evaluating sign language animations. To support this, we have engineered animation stimuli that can be interpreted (ambiguously) in different ways, depending on whether the participant correctly perceived a particular facial expression. We found that involvement of signers early in the stimuli design process is critical. To aid researchers, we have released our collection of stimuli for evaluating ASL facial expressions including Likert-scale and comprehension questions, and their answer choices.

Another methodological aspect we examined was whether the type (video of a human vs. a human-produced high-quality animation) of upper baseline presented for comparison purposes in a study (alongside the ASL animation being evaluated) affects the scores collected for comprehension and subjective questions for the animation being evaluated. We have also quantified whether changing the mode of presentation of comprehension questions in a study (video of a human vs. a human-produced high-quality animation) affects the responses recorded. Awareness of such effects is important so that future researchers can make informed choices when designing new studies and so that they can fairly compare their results to previously published studies, which may have made different methodological choices.

The third methodological research topic we investigated was how to measure users' reactions to animations without obtrusively probing their attention to the facial expressions by using eye-tracking technologies. We found that researchers can use eye tracking as a complementary or an alternative way of evaluating ASL animations, and we identified eye metrics that correlate with evaluation judgments from study participants.

The last methodological research topic we investigated was how demographic and experiential factors may influence acceptance of sign language animation by deaf users. We provided a deeper understanding of the relationship between participant characteristics and evaluation scores in this field and a concise set of questions that may be useful for researchers who are interested in minimizing the amount of study time spent collecting demographic and technology experience/attitude data.

Part I of this thesis answered the research questions:

***RQ1**: Can our stimuli and comprehension questions that contain linguistic facial expressions measure whether participants understand the indented facial expression effectively?* (The study presented in Chapter 4 supported this. In addition, Part II of this thesis shall discuss how these stimuli were used successfully in another study, thereby lending further support to this claim.)

***RQ2**: How does the modality (video of a human vs. a human-produced high-quality* *animation) of an upper baseline, presented for comparison purposes, affect the comprehension and subjective scores for the animation being evaluated?* (This question was examined in detail in Chapter 5, and various effects were identified in that chapter.)

RQ3*: Does the modality (video of a human vs. a human-produced high-quality animation) of instructions/comprehension questions in a study affect the comprehension and subjective scores for the animation being evaluated?* (This question was also examined in Chapter 5, and it was found that the modality did not make a difference in this case.)

RQ4: *Could eye-tracking be effectively used as a complementary or an alternative unobtrusive way of evaluating sign language animations with facial expressions?* (Chapter 6 examined this question, and it demonstrated that eye-tracking can be used successfully for this purpose.)

RQ5: *Which are the eye-tracking metrics that correlate with evaluation judgments from signers during the evaluation of sign language animations with facial expressions?* (This question was examined in Chapter 6, where several metrics were identified that correlated to participants' judgments of the quality of ASL animations.)

RQ6: *What demographic and technology-experience variables are predictive of participants' judgments during evaluation of sign language animations?* (Chapter 7 examined this question: several factors were identified that can influence participants' judgments of the quality of ASL animations.)

Part II: Data-driven Models for Syntactic Facial Expression Synthesis

Prologue to Part II

The survey of prior work in Chapter 9 will highlight limitations of current sign language animation systems; this survey will conclude that the future of ASL facial expression synthesis lies in combining linguistic knowledge of sign language and data resources from signers to achieve both intelligible and natural facial movements. In Part II of this thesis, we propose that an annotated sign language corpus, including both the manual and non- manual signs, can be used to model and generate linguistically meaningful facial expressions, if it is combined with facial feature extraction techniques, statistical machine learning, and an animation platform with detailed facial parameterization. To further improve sign language animation technology, we will assess the quality of the animation generated by our approach with ASL signers through the rigorous evaluation methodologies described in Part I.

We propose to collect human recordings of syntactic facial expressions from linguistically diverse sentences that cover subcases for each category of syntactic ASL facial expression considered in this thesis (Chapter 10). While there is still no consensus among animation researchers on how to best symbolically represent facial expressions in sign language we propose to adopt the MPEG-4 Facial Animation standard for the representation of facial movements in both the recordings of ASL signers, as extracted by modern computer vision techniques, and the animated sign language avatar.

After briefly discussing background information on key aspects of synthesis of facial expressions (Chapter 8), we will survey the literature on state-of-art facial expression synthesis for animations of sign language (Chapter 9), describe our facial-feature data collection (Chapter

10), and finally, explore each of the following four research questions in Part II of this thesis:

RQ7: *Is our MPEG4-enhanced animation platform sufficiently expressive such that it could produce facial expressions for ASL animations that are understandable and explicitly recognized by ASL signers?* (We will examine ***RQ7*** in Chapter 11.)

RQ8: *Is a Continuous Profile Model (CPM) able to produce a latent-trace curve that is representative of a set of ASL facial expressions, which had been provided as training data to the model?* (We will examine ***RQ8*** in Chapter 12.)

RQ9*: Can a Continuous Profile Model (CPM), which finds the underlying latent trace of extracted facial-feature data from multiple human recordings, identify feature curves that are more similar to human performances of novel sentences?* (We will examine **RQ9** in Section 13.1.)

RQ10: *Can a Continuous Profile Model (CPM), which finds the underlying latent trace of extracted facial-feature data from multiple human recordings, produce high-quality facial expressions for ASL animations, as judged by ASL signers in an experimental study?* (We will examine **RQ10** in Section 13.2.)

Chapter 8 Background on Synthesis of Facial Expression Animations

When synthesizing sign language animations, we often need to generate a novel animation by assembling a sequence of individual glosses (stem form of English one-word equivalents for each ASL sign) from a prebuilt dictionary. Each gloss usually has its own typical duration, which together with intervening pauses is used to determine a timeline for the full performance. It is important to emphasize that, for most expressions, adding facial expression performance to such animation is not a simplistic stretching or compressing of some recorded features of a human's face, such as motion-capture. Instead, careful synchronization of the two timelines is required. For example, there is additional intensity of the facial expression and head movements at the ending gloss 'WHAT' in the wh-word question 'YOUR NAME WHAT'. Many phrases with syntactic facial expressions begin, end, or include a word/gloss that has a special relationship to the facial expression being performed. This requires a good facial parameterization allowing detailed control of the face in a timeline as well as a good technique for extracting facial features from recordings of human signers. Moreover, researchers must incorporate a good evaluation methodology involving fluent signers for assessing their models.

This chapter will discuss key background topics such as facial parameterization and feature extraction, providing some nuance for most of the sections in Part II and specifically for the literature survey in Chapter 9. In that later chapter, the terminology and background concepts introduced here will be used to critique prior research on sign language facial expression animation.

8.1 Facial Parameterization

By the term "facial parameterization," we refer to a method of representing the configuration of a human face and the particular set of values that specify the configuration. Some parameterizations are more intuitive or elegant, and some use a more concise set of variables. Facial parameterization plays an important role for both the quality of data extracted from human recordings and the level of controls for the 3D avatar in the animations. A mapping between the two is required to directly drive animations from recordings of human signers.

This could be achieved manually by animators creating equivalent facial controls, similar to "rigging," in which a particular value is defined which has a pre-determined effect on a set of vertices that are on a region of the geometric mesh of the surface of the virtual human's face. The process of defining a subset of vertices affected by one of these rigging controls (and how they are affected) is often referred to as "creating a blendshape" for an animated face. Skilled 3D animation artists are able to define these relationships between a rigging control and a region of a virtual human's face such that the resulting movements are useful for animation and appear natural.

Alternatively, a parameterization could be determined automatically, by learning the weights of the blendshapes for each of the extracted features and their corresponding deformations. A characteristic example in the area of sign language is the work of Gibet et al. (2011), which is discussed in greater detail in Section 9.4. It is important that the level of parameterization should be detailed enough to allow for modeling of co-occurring facial expressions and allow synchronization with the head, torso, and manual movements.

8.1.1 Facial Action Coding System (FACS)

FACS, adopted by Ekman and Friesen (1978), describes the muscle activity in a human face using a list of fundamental actions units (AUs). AUs refer to the visible movements affecting facial appearance and are mapped to underlying muscles in a many-to-many relation. For example, AU12 is named Lip Corner Puller and corresponds to movements of the zygomaticus major facial muscle as shown in Figure 42. FACS has been adopted by many facial animation systems (e.g. Weise et al., 2011). However, in computer vision, automatic detection of AUs receives low scores in the range of 63% (Valstar et al., 2012). Some inherent challenges of FACS are a) some AUs affect the face in opposite direction thus conflicting with each other, and b) some AUs hide the visual presence of others.

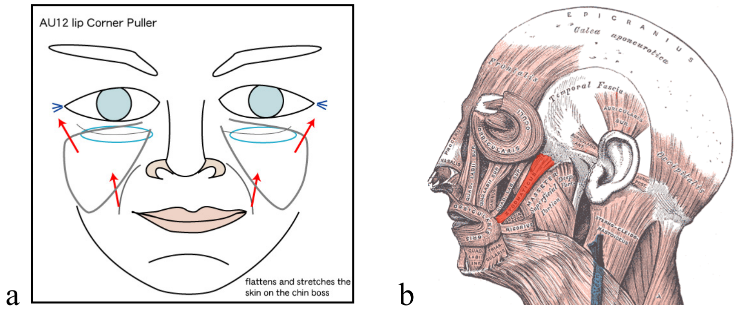


Figure 42: A FACS action unit example: (a) AU12 Lip Corner Puller (source: ISHIZAKI Lab) mapped to (b) zygomaticus major facial muscle (source: Wikipedia).

8.1.2 MPEG-4 Facial Animation (MPEG-4 FAPS)

The MPEG-4 compression standard (ISO/IEC 14496-2, 1999) includes a 3D model-based coding for face animation specified by 68 Facial Animation Parameters (FAPs) for head motion, eyebrow, nose, mouth, and tongue controls that can be combined for representing natural facial expressions. The values of these parameters are defined as the amount of displacement of characteristic points in the face from their neutral position (Figure 43a) normalized by scale factors (Figure 43b), which are based on the proportions of a particular human (or virtual human) face. Thus, the use of these scale factors allows for interpretation of the FAPs for any facial model in a consistent way. For example, FAP 30 is named "raise_l_i_eyebrow" and is defined as the vertical displacement of the feature point 4.1 (visible in Figure 43a) normalized by the scale factor ENS0 (visible in Figure 43b). MPEG-4 facial animation allows for an integrated solution for performance-driven animations, where facial features are extracted from the recording of multiple humans and applied across multiple MPEG-4 compatible avatars.

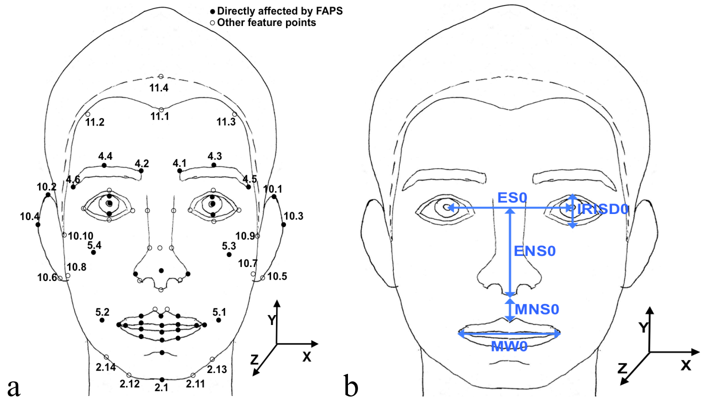


Figure 43: MPEG-4 facial animation (a) feature points and (b) scale factors.

The research community in facial expression synthesis has often adopted the MPEG-4 FAP standard. Their research focuses on non-sign-language facial expressions such as expressive embodied agents (Mlakar and Rojc, 2011), emotional facial expressions during speech in synthetic talking heads (Mana and Pianesi, 2006.), and dynamic emotional expressions (Zhang et al., 2008).

8.1.3 Proprietary Facial Parameterization

Driven by particular research questions and phenomena to be investigated, researchers have often adopted proprietary parameters to describe facial movements. For example, some sign language linguists have used normalized eyebrow height without distinguishing between the left and right eyebrow or between the inner, middle, and outer points of the eyebrow (Grossman and Kegl, 2006). Computer science researchers investigating the facial movement during lexical facial expressions have adopted similar approaches. Schmidt et al. (2013) tracked higher level of facial features such as mouth vertical and horizontal openness, left and right eyebrow states. A limitation was the sensitivity of the features to the facial anatomy of the recorded person.

8.2 Facial Feature Extraction

The discussion above focused on parameterization of the face from the perspective of synthesizing an animation; however, for much research on sign language facial expressions, it is also important to consider techniques for analyzing the performance of human signers and automatically extracting facial movement features. To analyze facial expressions in videos of human signers, researchers have adopted marker and marker-free techniques that allow extraction of facial features and their dynamic changes in time.

Motion capture is one representative approach for obtaining facial features in 3D space using markers, which are typically small reflective, or specially colored, dots affixed to key locations on the performer’s face. Gibet et al. (2011) have favored this approach to drive animations with facial expressions in French Sign Language. One drawback of this approach is that it prevents reuse of videos of signers because the video recordings that are collected require that “dots” are affixed to the human performer’s face, thereby making these videos ill-suited to some types of research uses. For example, the presence of the dots makes such videos infeasible for use as an upper baseline or for extracting facial features using other approaches such as computer vision. Given the scarcity of video corpora for sign language, it is desirable that collected video corpora be suitable for many types of research.

There are some marker-free approaches for tracking the face of a human performer in a video to extract key movement features. At the most manual end of the spectrum, many linguists have watched videos of human signers and manually recorded the changes in the performers’ face e.g. Grossman and Kegl (2006), and Weast (2008). Advances in the field of computer vision have made it possible to automate this process using statistical modeling and machine learning techniques. Schmidt et al. (2013) used active appearance models to obtain mouth patterns for lexical facial expressions in German Sign Language. Some limitations of this approach are that it requires training for each person, the results are not signer independent, and it does not compensate for obstacles in front of the face, a frequent phenomenon in sign language with the hands performing in the face area. Quick recovery from temporary loss of track due to occlusion has been investigated for ASL (Neidle et al., 2014) and MPEG-4 compliant face tracking systems are often adopted for face proportion independent data (e.g. Visage Technologies, 2014) though not yet adopted in sign language research.

Chapter 9 Literature Survey of Facial Expression in Sign Language Animation[13]

This survey reviews the literature on facial expression synthesis for animations of sign languages from different countries and regions, with the goal of understanding state of the art methods presented in the selected papers. To the author's knowledge, currently, there are five notable sign language animation generation projects that incorporate facial expression; these projects have been based in the United Kingdom, United States, France, Germany, and Austria. In the following discussions, this chapter will compare and critique representative papers from these projects. To help manage the paper discussion, the papers will be grouped and assigned a nickname (in bold font below) based on their project's name or a prominent feature of their approach. They will be discussed in the following order:

- **HamNoSys-based**: Elliott et al. (2004), Jennings et al. (2010), and Ebling and Glauert (2013) focus on automatically synthesizing sign language animations from a high-level description of the signs and facial expressions in terms of HamNoSys transcription system.
- **VCOM3D**: DeWitt et al. (2003) patent is about a character animation system with support for facial expressions that allows a human to efficiently "word process" a set of ASL sentences.
- **DePaul**: Wolfe et al. (2011) and Schnepp et al. (2012) used linguistic findings to drive eyebrow movement in animations of syntactic facial expressions with or without co-

[13] An earlier version of this survey is available as a technical report (Kacorri, 2015).

occurrence of affect;

- **SignCom**: Gibet et al. (2011) used machine-learning methods to map facial motion-capture data from sign language performance to animation blend-shapes;
- **ClustLexical**: Schmidt et al. (2013) used clustering techniques to automatically enhance a gloss-based sign language corpus with lexical facial expressions for animation synthesis.

Table 9: Comparison of state-of-the art approaches on sign language animation with facial expressions.

Project	**HamNoSys-based**	**VComD**	**DePaul**	**SignCom**	**ClustLexical**
Which sign language?	Sign Language	American Sign Language	American Sign Language	French Sign Language	German Sign Language
Which are the supported facial expressions?	Lexical, Modifiers, Syntactic, & Paralinguistic	Modifiers, Syntactic, & Paralinguistic	Syntactic and Paralinguistic (Affective)	Lexical, Modifiers, Syntactic, & Paralinguistic	Lexical
What portion of the animation pipeline process is this project focusing on?	Generates facial expressions given a higher level of description for their shape and dynamics.	Generates static or repetitive facial expressions as a parallel track to the manual signs.	Generates facial expression based on linguistic-driven models.	Maps motion capture data to facial blendshapes.	Selects a representative video of a human signer to drive the animation.
What corpus was used?	N/A	N/A	No corpus (pre-existing linguistic analysis of videos)	SignCom corpus (3 long dialogues with narrow vocabulary)	RWTH-Phoenix-Weather corpus (glosses and translations)
What is the input for the modeling of facial expressions?	Manual user selection from a set of available facial controls.	We believe 'empirical' rules and selected videos.	Linguistic findings	Motion capture data from one signers	Extracted facial features from videos (Active Appearance Models)
What type of facial parameterization was used?	Mix of proprietary and SAMPA. (E.g. both eyebrows raised, frown,)	Proprietary High level of facial expressions	Proprietary (E.g. eyebrow height, wrinkling)	Facial morphs FACS-like	Facial morphs MPEG-4-like
How is the facial expression time-adjusted to the manual signs?	User defined	Static key-frame or repetitive with proprietary transition rules.	Linear time warping is indicated.	No time warping (puppetry).	N/A facial expressions were not animated
How were deaf people involved in the design process or their team?	Signers are involved in corpus annotation and conduct of user studies.	Their linguist is a signer.	One of the authors is a signer.	Composition scenarios design by a sign language linguist.	Signers were involved in corpus annotation.
How was the approach evaluated?	There were 2 independent user studies including comprehensibility	User Study (Details about experiment setup, stimuli, evaluation metrics, etc. N/A)	User study	User study	Similarity Based Score without user feedback.

Table 9 highlights the similarities and differences of the projects along 9 axes: support for specific sign language, categories of facial expressions investigated, the portion of the animation generation process studied, use of annotated corpora, input data or hypothesis for their approach, details on face parameterization, synchronization of facial expressions to the manual movements, level of involvement of deaf people in the projects, and assessment with and without users.

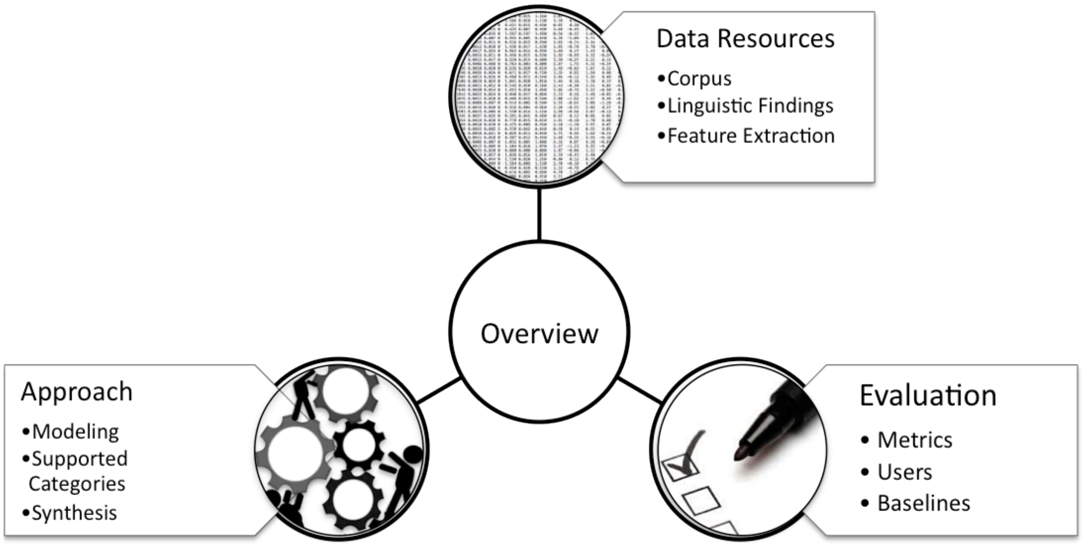


Figure 44: Organization of the critique on the selected sign language animation projects.

Since facial expression synthesis is not the primary research goal for all the projects, we merged related information and details of their work from multiple sources such as scientific papers, patents, tutorials, and presentations. For ease of comparison we review their contributions and critique their work based on the data sources they used, their approach to support facial expression synthesis, and the evaluation methodology and results (Figure 44).

9.1 HamNoSys-based

To represent their data resources, many European projects in sign language synthesis such as ViSiCast (Elliott et al., 2004), eSign (Elliott et al., 2007), DictaSign (Efthimiou et al., 2010) have adopted the Hamburg Notation System (HamNoSys) (DGS-KORPUS, 2014), a notation system for sign language transcription at a "phonetic" level. The stream of symbols is then parsed into an XML representation, Signing Gesture Markup Language (SiGML) (Elliott et al., 2000) that allows further processing of the signs to be synthesized in 3D animated avatars. The incorporation of the facial expressions into this scheme was not part of the original design and begun in the last version (4.0) of HamNoSys. Researchers are working on how to best represent the non-manual channel in SiGML and the animation software that supports it, e.g. JASigning (Ebling and Glauert, 2013; Jennings et al., 2010). Despite their progress, their work always assumes an input describing the sequence of facial expressions and changes over one or more signs; so far they have not focused on automatic synthesis through inference from multiple data/instances.

9.1.1 Approach

Gestural SiGML is an XML representation of the linear HamNoSys notation, with a structure similar to that of abstract syntax trees, containing additional information about the speed or duration of signing. In HamNoSys, a sign is transcribed linearly with iconic symbols, extending from 5 to 25 symbols from 200 "sub-lexical" units that are not language specific, describing its hand shape, orientation, location in the 3D space, and a number of actions as illustrated in Figure 45.

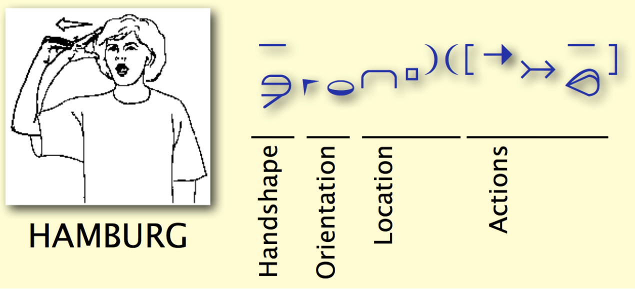
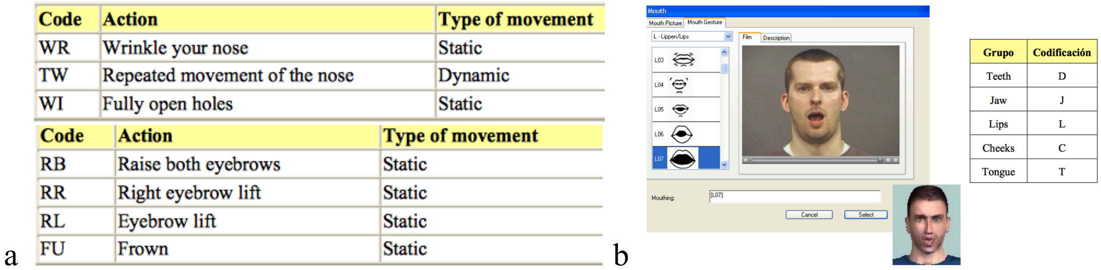


Figure 45: An example of the HAMBURG sign transcribed in HamNoSys. (source: Hanke, 2010)

In HamNoSys 4.0, non-manual information is supported in additional tiers synchronized to the manual movements and separated by: Shoulders, Body, Head, Gaze, Facial Expression, and Mouth. SiGML follows a similar structure for the representation of non-manuals. Specifically, focusing on the face only, the Facial Expression category includes information about the eyebrows, eyelids, and nose (Figure 46a), and the mouth category includes a set of static mouth pictures based on the Speech Methods Phonetic Alphabet (SAMPA) (Wells, 1997) to be used for lexical facial expressions, and a second set of mouth gestures irrelevant for speech, e.g. pursed lips, puffing cheeks, that could be used for adverbial facial expressions, emotions, and other linguistic and paralinguistic expressions involving the mouth (Figure 46b). The grouping of the eyebrows, eyelids, and nose in the same tier could complicate the modeling of facial expressions where these parts of the phase are not moving in parallel and could pose restrictions in co-occurring facial expressions that share some of the parts of the face. For example both a happy face and a wh-question involve movements of the eyebrow defined by different models.

The facial controls available in JASigning (Figure 46) are manually mapped to a set of detailed facial deformations of the avatar (e.g. Figure 47), called morphs or blendshapes, with information about the morphs involved, the amount they should be applied compared to their maximum, and timing characteristics such as onset, hold, and release times (Jennings, 2010). The onset and release of the facial movements can be defined as normal, fast, slow, sudden stop, or tense in HamNoSys and similarly in SiGML. In JASigning these timing profiles are mapped to different parameterized interpolations (Kennaway, 2007).



Code	Action	Type of movement
WR	Wrinkle your nose	Static
TW	Repeated movement of the nose	Dynamic
WI	Fully open holes	Static

Code	Action	Type of movement
RB	Raise both eyebrows	Static
RR	Right eyebrow lift	Static
RL	Eyebrow lift	Static
FU	Frown	Static

Mouthing: [L07]

Grupo	Codificación
Teeth	D
Jaw	J
Lips	L
Cheeks	C
Tongue	T

Figure 46: Face movement options in JASigning organized by alphanumeric tags for: (a) eyebrows and nose, and (b) mouth gestures. (source: San-Segundo Hernández, 2010)

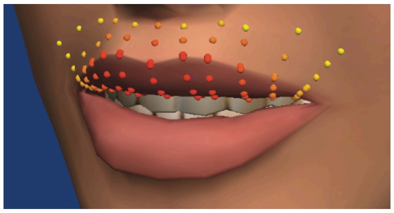


Figure 47: Blendshape controlling the upper lip movements in JASigning. (source: Jennings, 2010)

It seems that the capabilities of SiGML are not fully implemented. For example, JASigning does not allow facial expressions to be applied over multiple signs and does not automatically time-warp mouthing to the manual activity of the sign. In their work Ebling and Glauert (2013) suggested manual synchronization of the lower level face controls (morphs) separately for each of the signs to overcome the multiple signs coverage issue. For mouth time-warping problem they had to manually speed-up the mouthing over the manual actions of a sign and stretch out the duration of the sign.

9.1.2 Data Resources

The preexistence of the sign language corpus transcribed in HamNoSys drove SiGML and the sign language animation tools developed around it. The collection of videos in these corpora is often domain oriented, e.g. train announcements (Ebling and Glauert, 2013). In the last project, Dicta-Sign, corpus tasks evolve from transportation route description, description of places and activities, to more interactive content such as story telling, discussion, and negotiation (Dicta-Sign, 2014) that have a higher presence of facial expressions. However, as of today, these projects have not focused on modeling and automatic synthesis of facial expressions; instead, they have focused on synthesizing animations based on human description of each particular facial expressions and manual synchronization with the hand movements. For example, if you have an ASL sentence and you want it to be a Yes/No Question, then you should have the linguistic knowledge on how to select detailed movements of the different tiers of the nonmanuals (e.g., the precise tilt of the head and changes on the face) and how to synchronize them with the manual signs. It seems that a corpus, where the facial expressions are solely annotated in HamNoSys, cannot be used directly to train models of facial movements given that a more fine-grained description of these movements is required over the time axis. For example in HamNoSys, the eyebrow height in a wh-question is described as raised, lowered, or neutral over each of the signs in the sentence without specifying dynamic changes within the duration of a sign. To model data-driven natural movements of eyebrows in a wh-question, a greater precision to the dynamic changes in eyebrow vertical position at a fine-grained time scale is needed. The authors could benefit from computer vision techniques (see Section 8.2) that would allow them to extract this information from the training examples in their corpus.

9.1.3 Evaluation

Few researchers have conducted user studies to evaluate the results of their SiGML animations with facial expressions. Ebling and Glauert (2013) collected feedback from 7 signers in a focus-group setting. The participants watched 9 train announcements that included facial expressions such as rhetorical questions; however, the comprehensibility of the animations was not evaluated (i.e., no comprehension questions were asked after displaying the animations). The study indicated time synchronization issues between the facial expressions and the hand movements. In other work, Smith and Nolan (2013a) conducted a user study (15 participants) to evaluate emotional facial expressions in Irish Sign Language. They extended JASigning with 7 high level facial morphs corresponding to the 7 emotions to be evaluated in their study such as happiness, sadness, anger, disgust, contempt, fear and surprise. The participants watched 5 story segments, originally transcribed in HamNoSys, and performed by two avatars with and without facial expressions and answered comprehension questions. They found that the enhancement of the avatars with facial expressions did not increase comprehensibility (a small decrease was observed instead though not significant). This could be due to the facial movements being inaccurate or due to a lack of other linguistic phenomena such as syntactic facial expressions, etc. The results of their study would have been more useful for future researchers if they had included an upper baseline (e.g. the original video of a human signer performing the same stimuli) or if they had included some side-by-side comparison. The addition of such elements would have made it easier for future researchers to compare their results to the animation quality of these authors.

9.2 VCOM3D

VCom3D has designed a commercially available American Sign Language authoring tool, Sign Smith Studio (DeWitt et al., 2003; Hurdich, 2008), which allows users to produce animated ASL sentences by arranging a timeline of animated signs from a vocabulary (scripted). The tool offers a list of facial expressions that users can arrange in a track parallel to the manual signs. While this list covers adverbial, syntactic, and emotional categories of facial expressions, none of them is flexible in intensity nor can they be combined or co-occur, thus limiting the types of fluent ASL sentences that can be produced with this software.

9.2.1 Approach

Sign Smith Studio includes a library of ASL facial expressions that can be applied over a single sign or multiple manual signs one at a time. Their facial expressions usually consist of one key frame (Figure 48b) or a repetitive movement of few key-frames that are interpolated during the animation (Figure 48c). The time warping of the single key frame facial expression is simply a static face throughout the performance of the manual signs. The facial expressions with two or more key frames are either looped or performed once and the last key-frame is held static for the rest of the duration. The authors have a slightly sophisticated approach for the transitions. The software sentence-scripting interface graphically suggests that the start and end time points of a facial expression are aligned with start and end points of manual sign(s). Given that this is not in agreement with ASL, their software applies internal transitions rules. E.g. we notice that syntactic facial expressions start a bit before the first manual sign and repetitive facial expressions continue to loop on holds of the manual signs (Figure 49). However, the details of this rules and their inference method are not mentioned in peer-reviewed publications.

Hurdich's (2008) paper mentions the modeling of 60 facial expressions from VCom3D that vary in intensity and that can be combined to form a bigger set of ASL facial expressions. However, details to this implementation and version of the software are not published. Originally in their patent, the mesh model for the face of the avatar is really sparse in vertices, as shown in Figure 48a, and this can result in poor 3D animations. For example, in the MPEG4 standard 500 is the minimum number of vertices required to support facial expressions (Pandzic and Forchheimer, 2002).

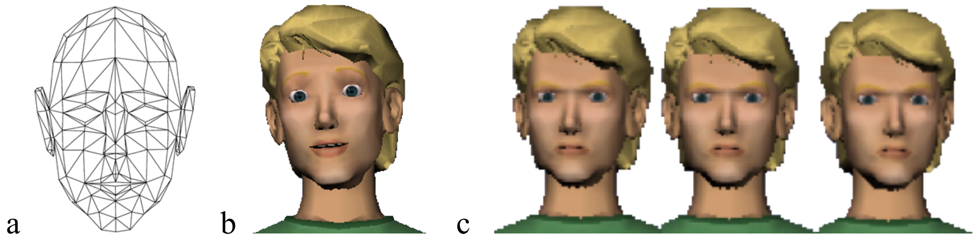


Figure 48: VCom3D (a) deformable mesh model of the avatar's head (source: DeWitt et al., 2003), (b) a 'one key-frame pose' topic facial expression, and (c) a 'few key-frames in a looping mode' negative facial expression.

The main limitation of VCom3D's approach to facial expressions is the lack of sufficient expressive control for the facial expressions in their sentence-scripting interface. The system does not allow for overlapping or co-occurring facial expressions. For example, an animated ASL sentence where a yes/no question facial expression applied throughout the sentence cannot convey emotion at the same time, include adverbial facial expressions, or a negative facial expression. Such combinations are necessary in fluent ASL sentences.

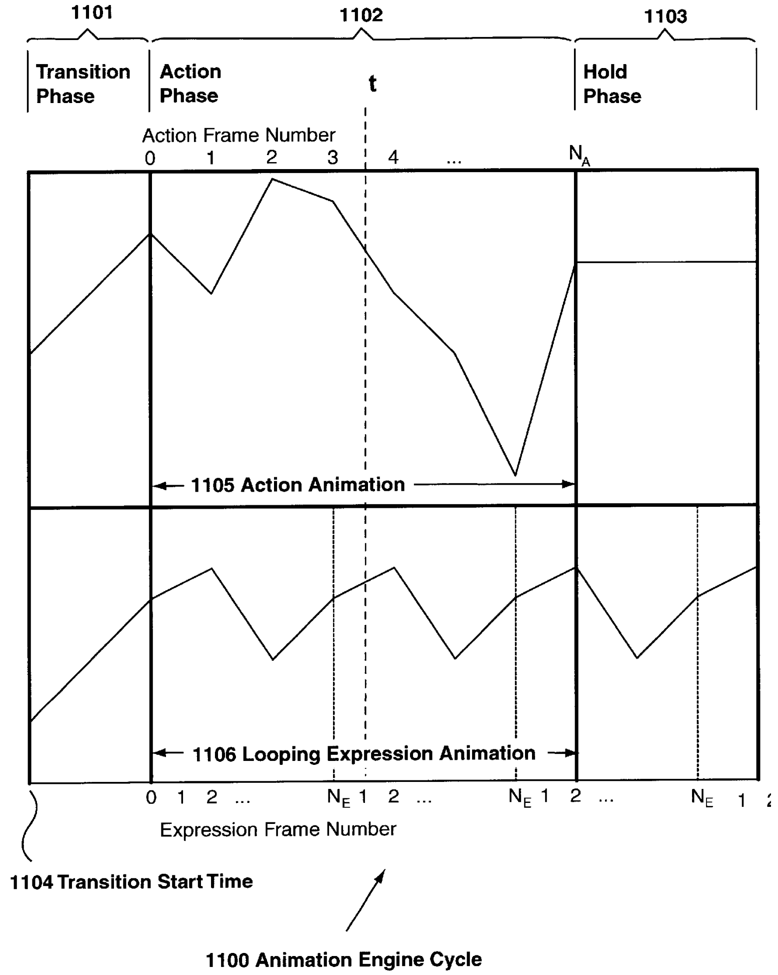


Figure 49: Timing diagram example that shows how a '3 key-frame' facial expression is applied to the hand movements (interpolation, looping, transitions, and holds). (source: DeWitt et al., 2003)

9.2.2 Data Resources

The authors do not explicitly mention the data sources they used to create the facial expressions, though it is likely that it was an animator who created them with the guidance of ASL videos where these facial expressions were performed and the support of an ASL signer in the VCOM3D team.

9.2.3 Evaluation

Evaluation details are sparsely mentioned in published work from the authors. From indirect allusions to their evaluation that were suggested in several of their publications, it was possible to conclude that they tested ASL animations generated by their system in a classroom setting with children of age 5-6 at the Florida School for the Deaf and Blind (Sims, 2000). They mention that English comprehension among young deaf learners was shown to improve from 17% to 67% (Sims and Silverglate, 2002). However, the details of this user study are not available in their published work. In order to obtain theses details about their evaluation of the system, this author needed to consult multiple non-peer reviewed papers (Sims, 2000; Sims and Silverglate, 2002; Hurdich, 2008) and read between the lines to infer what type of evaluations efforts have occurred on this project. It would be beneficial for the field of sign language animation if more of these details were in an easily available, peer-reviewed publication.

9.3 DePaul

Wolfe et al. (2011) and Schnepp et al. (2010, 2012) focus on American Sign Language facial expression synthesis using linguistic-based rules. An important aspect of their work is that they build models for the dynamic eyebrow movements of the facial expressions to be investigated. However, they make no use of raw motion-capture data of facial expressions from recordings of signers; nor do they use statistical analysis or machine learning techniques to drive their animation models and algorithms. Instead, the authors use previous linguistic findings to produce a natural exemplar to describe the eyebrow motion in those facial expressions. Specifically, they study brow height in two categories of syntactic facial expressions (yes/no and wh-word questions) with and without the co-occurrence of affective facial expressions such as happiness and anger.

9.3.1 Approach

Given two animation controls for the avatar's eyebrow movement, 'brows up' and 'brows down' with min and max values (0, 1), an artist created animations that follow exemplar curves that describe the motion of the eyebrows separately for wh-question, yes/no-question, happy affect, and angry affect based on the researcher's reading of prior ASL linguistic studies. Figure 50 illustrates an example of an ASL sentence, where a wh-question follows a topic facial expression. In the case of co-occurrence of the syntactic facial expressions with the affective ones, both curves contribute to the final eyebrow movement with 25% compression for the values of the syntactic facial expressions. The authors mention that these curves are not constrained as to length, which indicates that linear stretching may apply when the animation has different time duration. However, it seems that the authors have not considered all the cases for the syntactic facial expressions. For example, a wh-question may spread over one or multiple glosses, in the beginning or the end of an ASL sentence of varying length, and could be affected by the preceding or succeeding facial expression, as well as a co-occurring one (Watson, 2010).

A limitation of their approach to facial expressions is that the authors assume that the eyebrow movements are completely symmetrical for both left and right eyebrows. Further, the authors' controls are not detailed enough, e.g., they do not have separate control for inner, middle, and outer eyebrow points, which is needed to produce the full variety of eyebrow movements in ASL. Also, the authors do not mention horizontal movements of eyebrows, which are important e.g. for eyebrow furrowing in wh-questions and affection. To compensate for these limitations, the authors used artistic facial wrinkling, which can reinforce the signal being produced by the eyebrows, as illustrated in Figure 50. In addition, the authors only discuss their work on eyebrow height controls: beyond eyebrow movements, syntactic facial expressions also involve other facial and head parameters, such as head position, head orientation, and eye aperture. For example, while topic and yes/no-questions share similar eyebrow movements, they can be differentiated based on the head movements. However, the authors have not extended their animations models for these controls.

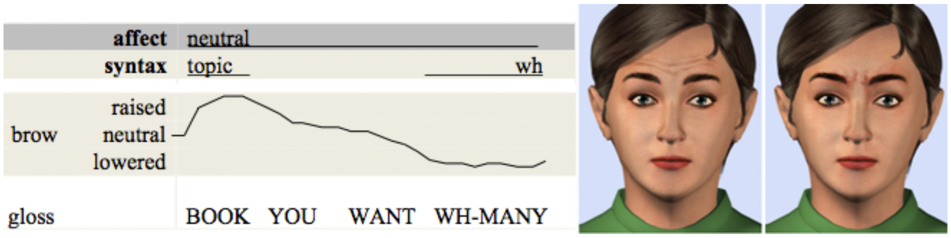


Figure 50: The ASL sentence “How many books do you want?” (source: Wolfe et al., 2011)

9.3.2 Data Resources

In addition to discussing earlier ASL linguistic research (e.g. Boster, 1996; Wilbur, 2003; and Crasborn et al. 2006) that had investigated the contribution of eyebrow movements and their intensity in syntactic and affective facial expressions, the authors consolidated the work of Grossman and Kegl (2006) and Weast (2008) that provide a greater precision to the dynamic changes in eyebrow vertical position (an example is shown in Figure 51).

Grossman and Kegl (2006) recorded 2 signers performing 20 ASL sentences in 6 different ways based on the facial expression category to be investigated such as neutral, angry, surprise, quizzical, y/n question, and wh-question. Then they averaged the common eyebrow vertical movements (among other features) for each of the facial expression category. One limitation of Grossman and Kegl's approach (as used by the DePaul researchers), is that Grossman and Kegl could have benefited from applying time warping techniques before averaging, since their sentences under consideration had different time durations.

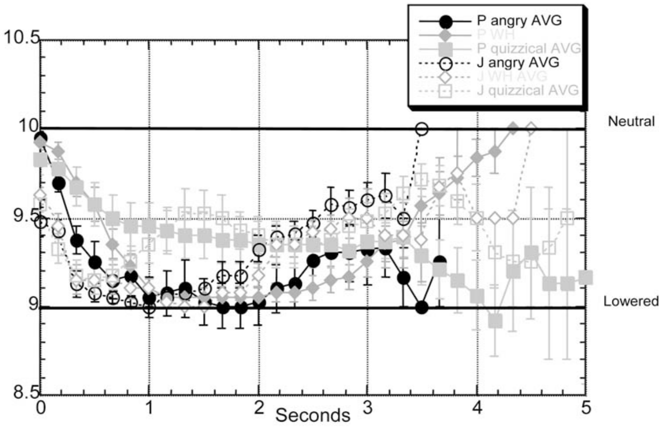
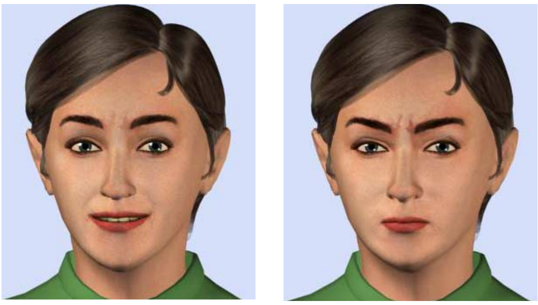


Figure 51: Eyebrow height on wh-questions, angry, and quizzical expressions. (source: Grossman and Kegl, 2006)

Wolfe et al. (2011) also based their animation algorithm for handling co-occurrence of syntactic and emotional facial expression on the findings of Weast (2008), who found that in the presence of some types of emotional affect, the eyebrow height range for the yes/no-questions and wh-questions is compressed. However, it seems that the selection of the numerical compression factor in Wolfe et al. (2011) animation algorithm was arbitrary.

9.3.3 Evaluation

To test the feasibility of their approach, the authors conducted a user study (Schnepp et al. 2010, 2012) with an ASL sentence (shown in Figure 50) where the wh-question co-occurred with a positive emotion, such as happiness, and a negative emotion, such as anger. Participants were asked to repeat the sentence and assign a graphical Likert-scale score for the emotional state and a 1-5 Likert-scale score for its clarity. Both studies were limited to the same, single short stimulus. A bigger cardinality and diversity in the stimuli set (different length sentences, different location of the wh-word, etc.) would be a requirement for a statistical analysis. The authors could also have benefited by including in their study a lower baseline to compare with their animations, e.g. a wh-question with neutral emotion state, or by including videos of a signer as an upper baseline for comparison. These enhancements to the study design would have made their results more comparable with future work. Another methodological concern is that it appears that in Schnepp et al. (2012), the facial expressions of the two stimuli (happy vs. angry) did not differ only in their eyebrow position and wrinkling. They also differed in the mouth shapes that conveyed the emotion (Figure 52). This would make it rather difficult to conclude that the participants perceived the intended affect in animations solely due to the quality of the author's co-occurrence algorithm for the eyebrow movements. The mouth, a point on the face where deaf people tend to focus during signing (Emmorey et al., 2009), could have driven the results instead.



WH-question, happy

WH-question, angry

Figure 52: Co-occurrence of wh-question with emotions (source: Schnepp et al., 2012).

9.4 SignCom

The SignCom project seeks to build an animation system that combines decomposed motion capture data from human signers in French Sign Language. The system architecture, proposed by Gibet et al. (2011), incorporates a multichannel framework that allows for on-line retrieval from a motion-capture database of independent information for each of the different parts of the body (e.g., hands, torso, arms, head, and facial features) that can be merged to compose novel utterances in French Sign Language. Their focus on synthesis of facial expressions lies at the level of mapping facial mocap markers to values of animation controls in the avatar's face (these puppetry controls for the face are sometimes referred to as "blendshapes"), which are designed by an animator to configure the facial geometry of the avatar, e.g. vertical-mouth-opening.

9.4.1 Approach

The authors recorded the facial movement of a signer with 43 facial motion capture markers (Figure 53a) resulting in 123 features when considering the marker's values in the 3D space. The values of the markers (calculated in a common frame) were normalized based on their relative distance to the upper nose sensors considered by the authors to remain unchanged during most of the face deformations. To map these features to the values of 50 blendshapes in the geometrical model of their avatar the authors considered probabilistic inference and used Gaussian Process Regression to learn the corresponding blendshape weights from the mocap values.

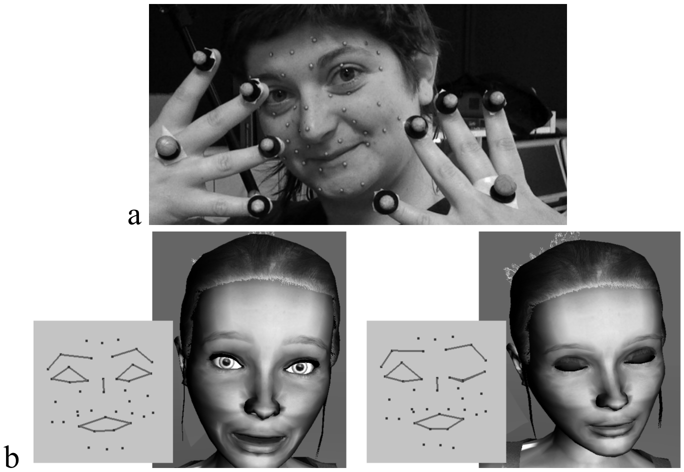


Figure 53: (a) Motion-capture sensors on a signer's face and (b) blended faces of the avatar driven by the values of the facial markers. (source: Gibet et al., 2011)

As discussed in Section 8.2, blendshape-weight learning wouldn't be necessary had the facial features been extracted in an MPEG-4 format and used to drive an MPEG-4 compatible avatar. It is also unclear whether this approach would require motion data recorded from different signers (different face geometry, signing style, etc.) to be treated separately. Also, the use of motion capture sensors is a time consuming approach for recording a big corpus of facial expressions when compared to the alternative of applying computer-vision software to pre-existing video recordings of signers.

In our understanding, the authors are not capturing or rendering any tongue movements that are important for the understandability of lexical facial expressions that involve mouthing. Even though the authors used 12 motion capture cameras, marker-occlusion may occur and thus this is a limitation to the quality of the retrieved data. The focus of their work in facial expressions is limited to playing pre-recorded face motion in their avatar (puppetry) and does not include any synthesis aspect or statistical modeling of the recorded facial expressions.

9.4.2 Data Resources

The facial expressions were obtained from the SignCom corpus (Duarte and Gibet, 2010), a collection of three scripted elicitation sessions, each including about 150-200 signs with a narrow vocabulary (~50 unique signs). The narratives were designed to obtain multiple occurrences of particular signs. However, the authors do not refer to details of the facial expressions covered, e.g. types of facial expressions and their co-occurrence. Another limitation of this corpus is that only one signer is recorded performing all the sentences. Thus, the facial performances that are obtained from the dataset could be peculiar to this signer's facial movement style.

9.4.3 Evaluation

The authors conducted a web-based user study with 25 participants to evaluate the facial expressions that were directly driven by the motion capture data using Gaussian Process Regression. They compared their animation to manually synthesized facial expressions (as an upper baseline) and to a neutral static face without facial expressions (as a lower baseline). Given three FSL passages, participants were shown three pairs of animations and were asked to select the one they preferred and to explain their choice. When comparing their approach to the upper baseline, the authors did not notice an important difference. However, the authors do not mention which categories of facial expressions were included in the stimuli. For example, the lack of syntactic meaningful facial expressions in the stimuli could explain why the scores for animations with facial expressions were not that different from the one with a static face. A limitation of this study is the small number of stimuli used. Given a greater cardinality and diversity in the set of stimuli used in the study, the results obtained could be considered more generalizable to sentences beyond those specifically appearing in the study. Last but not least, their user study did not include comprehension questions, an evaluation approach that is often adopted in user studies of sign language animation synthesis (discussed in Part I)

9.5 ClustLexical

Schmidt et al. (2013) is the most sophisticated study among the five selected projects in this survey from a facial expression synthesis perspective. One important characteristic of this study is that it uses machine-learning techniques, such as clustering, to obtain a representative video of a signer from which features can be extracted for avatar animation. This paper is restricted only to lexical facial expressions in German Sign Language (GSL), which are facial expressions that specially appear during particular signs. In GSL, these facial expressions involve mouth patterns that often derive from spoken German, and they can distinguish signs with identical manual movements (details in Section 2.2.1). Compared to syntactic facial expressions, lexical facial expressions pose fewer difficulties because:

i) they can be investigated without manually annotating a corpus of full sentences,

ii) the underlying variation of signs corresponding to the lexical facial expression is limited (usually one),

iii) in German Sign Language, the mouth movements of lexical facial expressions are closely tied to sounds of the spoken German language and thus well investigated mouth formations in the literature, and

iv) the mouthing does not co-occur in other linguistically meaningful facial expressions.

9.5.1 Approach

Starting from a set of German Sign Language videos that each contained only a single sign, the authors performed automatic cluster techniques based on the similarity of the signer's facial features in the videos. Specifically, the authors focused on German Sign Language signs that are disambiguated by the lexical mouthing that occurs during the sign. Each video was labeled with an ID-gloss (a label for the sign based on its manual component only) and a German translation of the sign based on the context (which is suggested by the facial expression and mouthing that occurs). Thus, the sign for "mountain" and the sign for "Alps" would be labeled with identical ID-glosses but different translations; these signs consist of identical manual movements but different facial expressions and mouthing. The video that is found to be the central element of the biggest cluster for a particular (ID-gloss, translation) pair is then selected as the representative video from which the mouth movements can be extracted for avatar animation. Their clustering approach uses Hidden Markov Models (HMMs) to estimate the similarity between the pairs, and the Adaptive Medoid-Shift[14] approach (Asghar and Rao, 2008) for the selection of the representative video. To define the distance between two videos, they train a HMM in the facial features of one video and calculate the Viterbi path on the facial features of the second video. Given that the speed of the same sign may vary across sentences and across signers, it is likely that the authors would have had better results if they had first applied some time warping in the time series of facial features for each of the videos. For

[14] Adaptive medoid-shift modifies the medoid-shift clustering algorithm (Sheikh et al., 2007). Both are nonparametric and perform mode-seeking to determine clusters. Data points are assigned to clusters by following local gradients to the modes. The number of clusters is computed automatically without any initialization, and it can work on non-linearly separable data.

example (Oates et al., 1999) suggests that Dynamic Time Warping (DTW) can help for the initialization for HMM clustering. Also it would have been interesting if the authors had investigated whether an averaging approach would have worked better than the selection of a centroid, since averaging could have eliminate some noise. E.g., the authors could apply DTW between all the other videos in the cluster and the centroid to get the same timing and then average (e.g., using weights based on the distance).

9.5.2 Data Resources

The short videos and the (gloss, translation) pairs were obtained from the RWTH-Phoenix-Weather corpus, a gloss-annotated video corpus with weather forecasts in German sign language consisting of 2711 sentences with a rather limited vocabulary of 463 ID-glosses collected over 7 signers. The corpus was annotated with the ID-glosses, their time boundaries in the video, corresponding translation in the spoken language, and time boundaries of the translated sentences in the video. The authors used the open-source toolkit GIZA++ (Och and Ney, 2003) to automatically obtain the pairs (ID-gloss, translation) from the aligned German Sign Language and spoken German sentences in the corpus as shown in Figure 54.

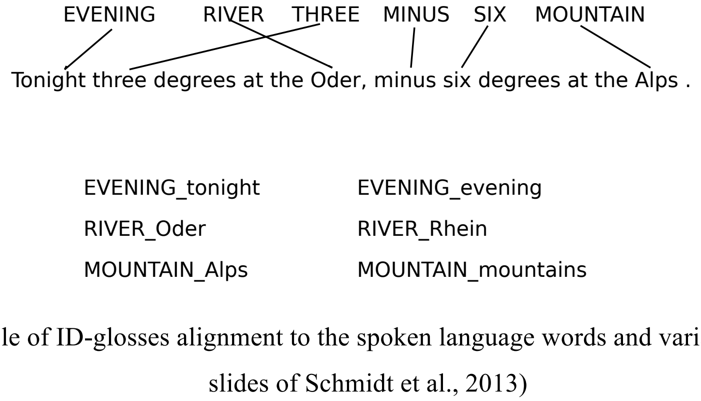

EVENING RIVER THREE MINUS SIX MOUNTAIN
Tonight three degrees at the Oder, minus six degrees at the Alps .
EVENING_tonight
EVENING_evening
RIVER_Oder
RIVER_Rhein
MOUNTAIN_Alps
MOUNTAIN_mountains

Figure 54: An example of ID-glosses alignment to the spoken language words and variants extraction. (source: slides of Schmidt et al., 2013)

The authors used Active Appearance Models (described in Section 8.2) to extract high-level facial features used in the clustering approach such as vertical and horizontal openness of the mouth, distances of the lower lip to chin and upper lip to nose, states of the left and right eyebrow and, the gap between eyebrows, as shown in Figure 55. However, they did not include any features describing tongue positions, which are important for lexical facial expressions in German Sign Language that involve mouthing. Also, as mentioned in Section 8.2, appearance based modeling does not generalize well across different human signers in videos and would provide poor results when the hands are in front of the face, which is a common occurrence during signing. The authors might have had better results if they had employed a state-of-art solution for MPEG-4 feature extraction (e.g. Visage Technologies, 2014), which is a face-tracking technique that is less specific to the appearance of a single human, or other approaches that allow for quick recovery from temporary loss of track due to occlusion (e.g. Yu et al., 2013).

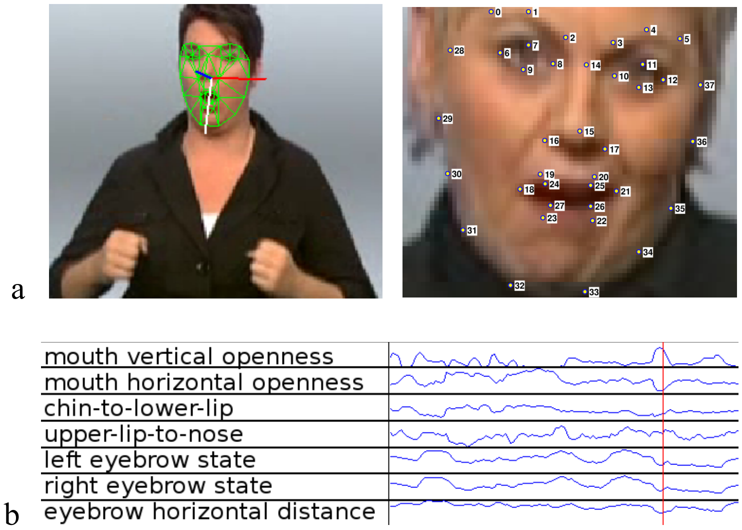


Figure 55: High-level feature extraction with Active Appearance Models. (source: Schmidt et al., 2013)

9.5.3 Evaluation

While Schmidt et al. (2013) described an approach for selecting representative videos of signers for the purpose of driving animations of sign language requiring lexical facial expressions, the authors never actually produced any such animations. Further, the authors did not evaluate their results in a user study with signers. Instead, they evaluated their system against a collection of manually labeled pairs of (ID-gloss, translation). Specifically, they performed the following calculations:

- **Cluster evaluation**: They evaluated the quality of the clustering algorithm using a subset of manually labeled pairs of glosses by calculating precision, recall, and F-measure between the clusters provided by the algorithm and the manually labeled mouthings for each (gloss, translation) pair. They found that about two thirds of the pairs were correctly classified, on average.

- **Medoid evaluation**: They also evaluated the quality of the representative video, selected as the medoid of the cluster, compared to the other glosses in the same cluster. They defined accuracy to be the fraction of the labeled videos that have the same label as the medoid of the cluster they belong to. They found an average accuracy of 78.4%.

It would have been preferable if the authors had been able to demonstrate that a good representative video can actually be used to drive natural and understandable animations of sign language as perceived by signers (especially since driving animations of facial expression was listed as a motivation for their work).

9.6 Conclusions

This chapter surveyed modern techniques in the field of facial expression generation in sign language animations over the past fifteen years with a detailed critique of five representative projects and their related papers. Strengths and drawbacks of these projects were identified in terms of: their support for a specific sign language, the categories of facial expressions investigated, the portion of the animation generation process studied, use of annotated corpora, input data or hypothesis for each approach, and other factors. This conclusions section summarizes the main observations, across all of the papers in this literature survey, in regard to these factors.

Most of the projects focus on a specific **sign language** such as American, French, and German Sign Languages, and only one of them is a sign-language-independent approach based on a “phonetic” level notation. The **categories of facial expressions** supported in these projects extend over linguistic facial expressions such as lexical, modifiers, and syntactic, to paralinguistic ones such as affection.

One of the main differences between these approaches was the portion of the **facial expression animation pipeline** they focus on. In three out of five projects, the animation of the facial expressions was driven by a description of their shape and dynamics, either manually animated as a single static frame or few repetitive frames, or driven by a signer wearing motion capture equipment. The other two projects present the most interesting approaches in this perspective. The first introduces linguistically driven modeling of eyebrow movements for wh-question facial expressions. The second uses machine learning approaches, such as clustering, for the selection of a representative video, whose facial features could drive the animations for lexical facial expressions.

A limitation of the surveyed work is that, despite attempts to collect and annotate data from recordings of signers, there is still need for additional **corpora** including video recordings from multiple signers and of multiple performances of facial expressions of multiple linguistic categories. To support future research on sign language facial expression animation, these corpora would also need: linguistic annotation and detailed and signer-independent feature extraction. Therefore, given these current resource limitations, it is still difficult to make use of statistical machine learning techniques to synthesize animations of all the types of facial expressions, combined transitions, or co-occurrence (because too few examples of these combinations appear in the corpora).

The **input type** for the modeling of facial expressions also varied among the surveyed projects. Researchers obtained facial features from signers either *manually*, where linguists observe videos and note down changes in the signers' face e.g. using HamNoSys, or *automatically* using motion capture or computer vision techniques.

The **facial parameterization** for both the extracted facial features and the facial controls of their avatars were often proprietary and specific to the project tasks. In some cases, similarities to the FACS or MPEG-4 Facial Animation standards were observed. An interesting approach was the incorporation of artistic artifacts such as wrinkling to reinforce the signal being produced by the facial movements.

Overall, the **time-adjustment** of the generated facial expressions to the manual movements of the avatar, a critical quality for sign language animations, has not been thoroughly addressed by the sign language animation research community.

Chapter 10 Extracted Facial-Feature Data for ASL Animation Synthesis

This chapter describes the dataset and features that will be used to drive our ASL syntactic facial expression animations. Collecting and linguistically annotating a video corpus of ASL sentences that are performed by human signers and that include a variety of syntactic facial expressions is a time-consuming process. To drive our final models of facial expressions for ASL animation synthesis, we plan to use a set of stimuli videos that have been recorded and annotated in our lab as initial pilot data and to also use a more linguistically diverse video corpus that is being collected by collaborators at Boston University.

10.1 Extracted MPEG4 Head/Face Features

To extract the facial features and head pose of the ASL human signers in the video recordings, we use Visage Face Tracker, an automatic face tracking software (Pejsa and Pandzic, 2009) that provides MPEG-4 compatible output. While automatic facial tracking is still in development, Visage is able to track about 42 out of 68 MPEG4 facial features (described in Section 8.1.2) and 3 additional features with head displacement in the 3D plane. The system attempts to automatically fit a 3D mask to the detected neutral face in the plane and allows for human intervention for further adjustments to the fit, as shown in Figure 56. This initial step, called profiling, is crucial and is required only once given a similar setting for all the videos that are analyzed. It saves information on the texture of the human's face and calculates the facial proportions that are later used to internally normalize the facial features and export

MPEG4 parameters. A bad profiling could result in low quality feature extraction. The system is able to compensate during brief periods of time when the hand occludes the face of the signer in the video, which occurs frequently during ASL signing since some ASL manual signs are performed near the face.

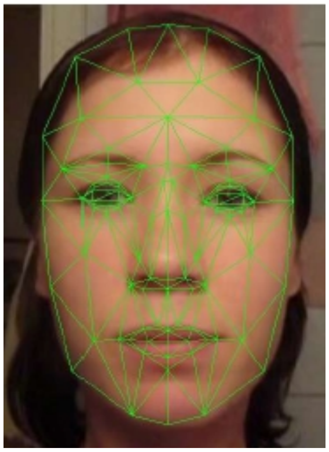


Figure 56: Fitted 3D mask in Visage. (Source: visagetechnologies.com)

Given that the categories of ASL facial expressions we are investigating mostly involve head and upper face movements, as discussed in Section 2.1, for this thesis we are interested only in a subset (a total of 18) of the features extracted by Visage, which includes:

- **Head orientation** (FAP48-FAP50)**:** orientation parameters given in Euler angles (10^{-5}rad) defined as pitch, yaw, and roll.
- **Head displacement** (Head x, Head y, Head z): 3 parameters describing head location in 3D space.
- **Vertical displacements of eyebrows** (FAP31-FAP36)**:** 6 parameters describing vertical movements of the inner, middle, and outer points of the left and right eyebrow. Their values are in the range (-360, 360).

- **Horizontal displacements of eyebrows** (FAP37-FAP38)**:** 2 parameters describing the horizontal movements of the inner points of the eyebrows. Their values are in the range (-300, 300).
- **Eye aperture** (FAP19-FAP22): 4 parameters describing the upper and lower eyelid position. Currently, Visage only produces Boolean values for these parameters.

10.2 Initial Pilot Data[15]

While investigating our initial approaches for a data-driven facial expression synthesis, we have been using a small set of 48 videos as pilot data that were designed as evaluation stimuli for ASL syntactic facial expressions, described in detail in Section 4.2. We used Visage (version 7.1) to analyze the video recordings and to produce files that contain information about the head pose and facial features of the human signer for each frame of the video. The tracking results, part of the collection, are shared with the research community as comma-separated values (CSV) files.

Head pose data is given as translation from the camera in the 3 dimensions (x, y, z) and as head rotation (pitch, yaw, roll). The obtained facial features included all 68 MPEG-4 facial action parameters for each frame of the video where only 42 included non-zero values. This information could also be used by future researchers to animate the face of a virtual human character (Pandzic and Forchheimer, 2003) performing these stimuli passages. Such a character could be displayed as a baseline for comparison in an experimental evaluation study.

[15] This section describes joint work with Professor Matt Huenerfauth (Huenerfauth and Kacorri, 2014).

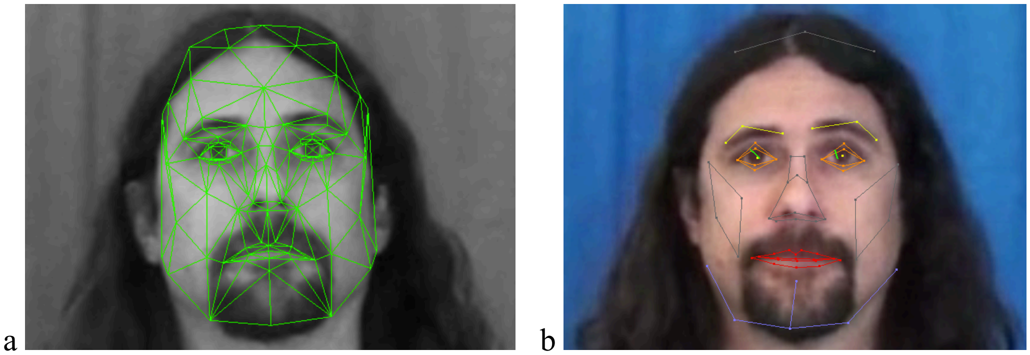
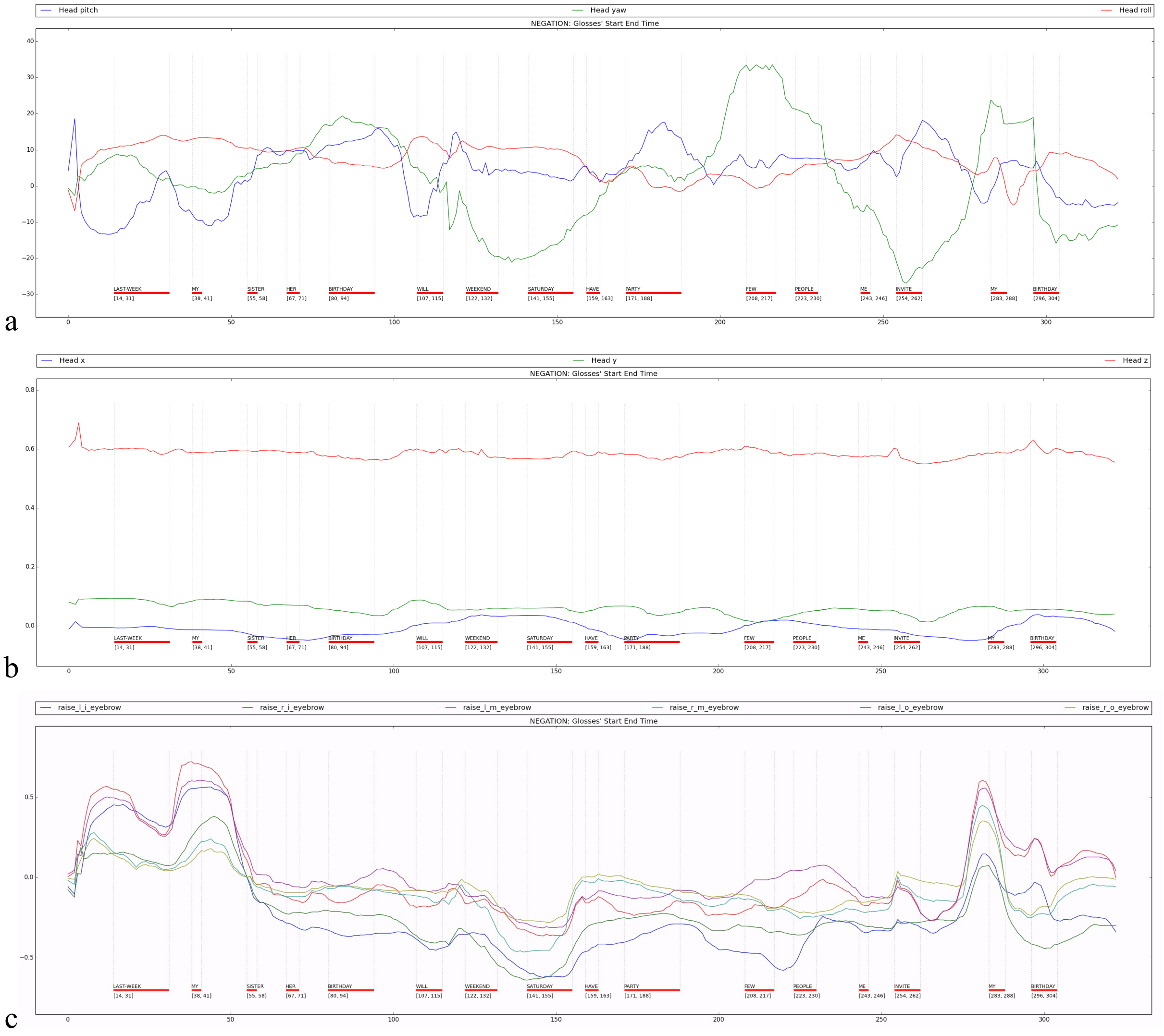


Figure 57: ASL signer’s a) fitted face shape mask and b) tracking screenshot in Visage software.

For optimal results, the Visage software was used in offline mode. The quality of the results is bounded by the performance of the software on the video recordings and the initial manual process of mask fitting to the face as shown in Figure 57a. For example, the tracker (a screenshot is shown in Figure 57b) may lose the face if the head movement is too fast or if large parts of the face are covered for a prolonged time, e.g., by the hands. We observed that this is happening for 0%-7.6% (avg. 1.6%) of the story duration in our stimuli collection. In this case, the lost frames are indicated with a tracking status other than “OK” in the CSV file, and all the extracted head and facial features would normally have the value 0. We processed the data and filled in the values of the lost frames using spline interpolation (smoothing degree 1) while maintaining the tracking status information. Although interpolation may work well for the facial feature values, it can sometimes be problematic for head rotation, because it is currently represented in the form of Euler angles (pitch, yaw, roll). We advise future researchers to consider first converting the head rotation into another representation (e.g., quaternions) before applying interpolation techniques to fill in the rotation values for the lost frames.

In addition to feature extraction, we annotated the videos with the glosses (labels for the manual signs), start and end frames of each of the glosses, and start and end frame for the facial expression, which often precedes the first gloss and exceeds the last gloss in the phrase. Figure 58 illustrates the extracted features and glosses in a timeline for the ASL passage “LAST WEEK MY SISTER HER BIRTHDAY. WILL WEEKEND SATURDAY HAVE PARTY. FEW PEOPLE ME INVITE. MY BIRTHDAY.” A Negative facial expression occurs during the phrase “MY BIRTHDAY”.



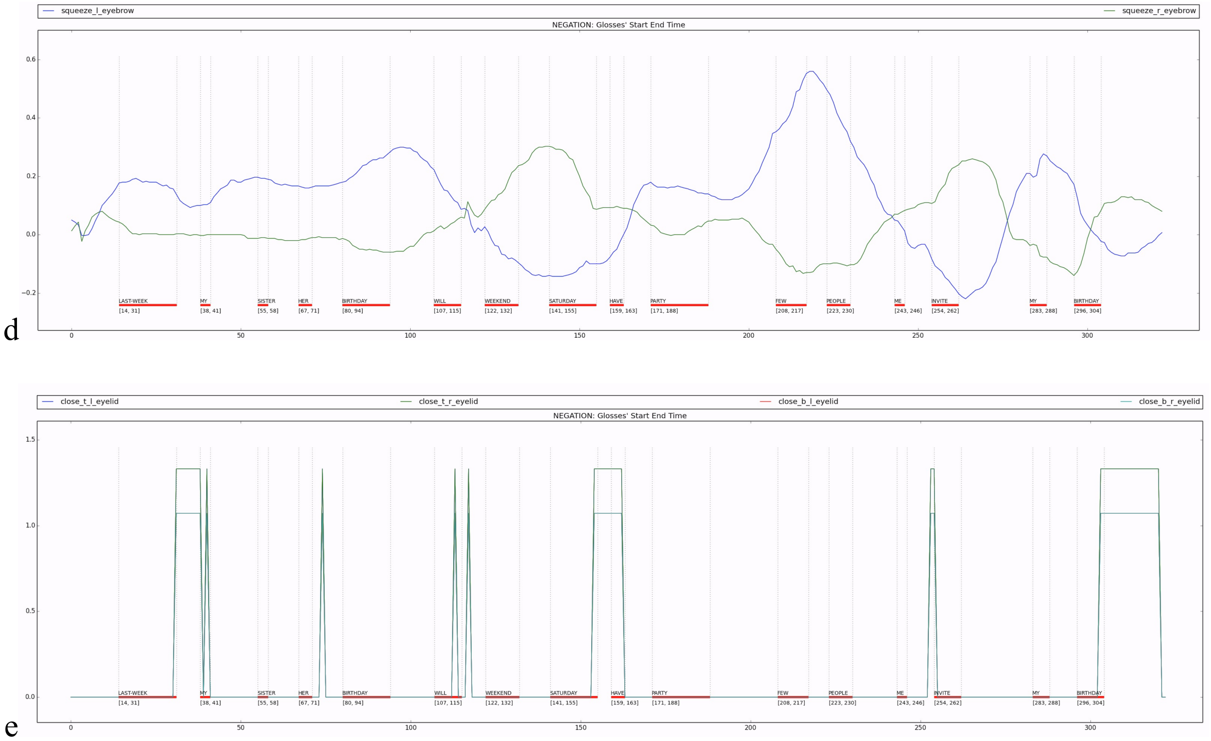


Figure 58: Extracted features and gloss annotation for an ASL passage in the pilot data that includes facial expressions such as topic and negative. The features displayed in each graph include: (a) pitch, yaw, roll, (b) head x, y, z, (c) raise_l_i_eyebrow, raise_r_i_eyebrow, raise_l_m_eyebrow, raise_r_m_eyebrow, raise_l_o_eyebrow, raise_r_o_eyebrow, (d) squeeze_l_eyebrow, squeeze_r_eyebrow, and (e) close_t_l_eyelid, close_t_r_eyelid, close_b_l_eyelid, close_b_r_eyelid.

10.3 Linguistically Diverse Data

Our pilot data set described above was obtained during our prior work (Chapter 4), in which we collected ASL videos for the purpose of producing a set of stimuli for evaluation experiments. To support the modeling work in Chapter 12, we have collected an even larger data set of videos. Specifically, to produce this larger data set, we extract facial features from ASL videos created by linguists and computer scientists (Neidle and Sclaroff at BU; Athitsos, now at U. Texas, Arlington; and Metaxas at Rutgers) as part of the "Corpora Through the American Sign

Language Linguistic Research Project" (ASLLRP) and the NSF-funded collaborative project "Generating Accurate Understandable Sign Language Animations Based on Analysis of Human Signing."

While the video collection and linguistic-annotation aspects of these projects are still in progress, we obtained 174 annotated video recordings from 1 ASL signer (female, Figure 59). Since the MPEG4 standard ensures normalized features by the signer's face proportions, we are able to use this data to drive the models of facial expressions that will be animated in our avatar and evaluated by comparison against a second signer (from our pilot data in Section 10.2).

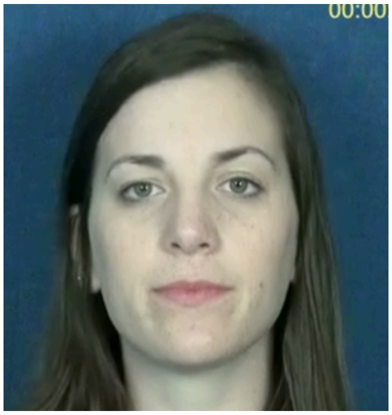


Figure 59: Neutral facial poses to be used for the signer's profile in Visage.

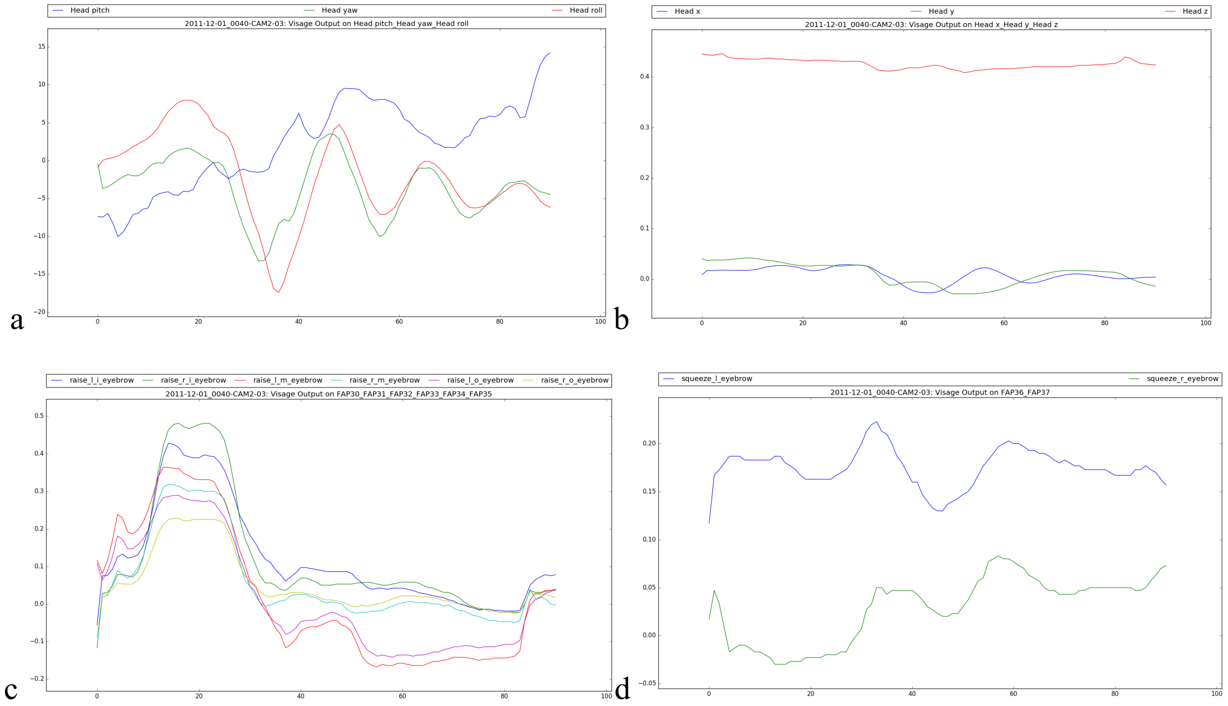
Based on linguistic insights, such as distance from adjacent facial expressions and single versus multiple glosses, we group the dataset into categories and sub-categories, as shown in Table 10. For each subcategory, we will train a model and use it to synthesize the corresponding facial expression for a novel animated sentence.

Table 10: Grouping of syntactic ASL facial expressions into subcategories.

	Subcategories	#Examples	#Glosses
Yes/No-Question	**A** – Immediately Preceded by Eyebrow Raise: A facial expression, e.g. topic that requires rising of the eyebrows, immediately precedes a yes/no question.	9	(2, 5) μ: 3.33
	B – Not adjacent to Eyebrow Raising: The Yes/No facial expression is not immediately preceded by an eyebrow raise facial expression.	10	(2, 6) μ: 3.8
WH-Question	**A** – During a Single Gloss: Performed during a single word, namely the wh-word (such as what, where, and when).	4	1
	B – During a Multi-Gloss Phrase: Performed during a phrase consisting of multiple words.	8	(2, 5) μ: 2.88
Rhetorical Question	**A** – During a Single Gloss: Performed during a single word, namely the wh-word (such as what, where, and when) that is performed at the end of the sentence.	2	1
	B – During a Multi-Gloss Phrase: Performed during a phrase consisting of multiple words.	8	(3, 4) μ: 3.5
Topic	**A** – During a Single Gloss: Performed during a single world.	29	1
	B – During a Multi-Gloss Phrase: Performed during a phrase consisting of multiple words.	15	(2, 4) μ: 2.27
Negative	**A** – Immediately Preceded by Eyebrow Raise: A facial expression, e.g. topic, conditional/when or rhetorical that requires rising of the eyebrows, immediately precedes a negative facial expression.	16	(2, 5) μ: 3.06
	B – Not Adjacent to Eyebrow Raising: Not immediately preceded by eyebrow raise facial expression.	25	(2,7) μ:3,88

The grouping in Table 10 is based on the available dataset of videos that we have received from our collaborators at Boston and Rutgers Universities. However, there could be more granular modeling of ASL facial expressions, if a larger dataset were available with more examples of various combinations or special cases of facial expressions.

We used Visage[16] (version 7.1) to analyze the video recordings and to produce files that contain information about the head pose and facial features of the human signer for each frame of the video. Figure 60 illustrates the extracted features in a video-frame timeline for the ASL passage "PEOPLE REFUSE GO-OUT". A Topic occurs during "PEOPLE" and a Negative facial expression during "REFUSE GO-OUT".



[16] Visage Technologies – http://www.visagetechnologies.com/.

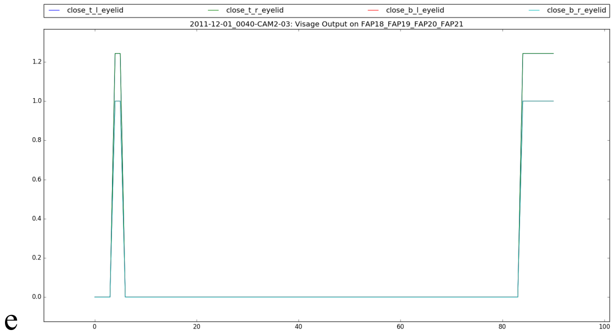


Figure 60: Extracted features and gloss annotation for an ASL passage in the linguistically diverse data that includes facial expressions such as topic and negative. The features displayed in each graph include: (a) pitch, yaw, roll, (b) head x, y, z, (c) raise_l_i_eyebrow, raise_r_i_eyebrow, raise_l_m_eyebrow, raise_r_m_eyebrow, raise_l_o_eyebrow, raise_r_o_eyebrow, (d) squeeze_l_eyebrow, squeeze_r_eyebrow, and (e) close_t_l_eyelid, close_t_r_eyelid, close_b_l_eyelid, close_b_r_eyelid.

Based on the gloss annotation and non-manual annotations we isolate the portion of the extracted features from the video, where the facial expression of interest occurs.

Chapter 11 Virtual Human Platform[17]

Our new animation platform is based on the open source animation engine EMBR (Embodied Agents Behavior Realizer, Heloir and Kipp, 2009), which has been previously used for creating sign language animations. EMBR produces real time multimodal animations specified in a high-level control language named the EMBRScript. Since the system originally supported German sign language, which is distinct from ASL, we extended their 3D avatar and GUI for creation of new signs and utterances with ASL handshapes. EMBR supports nonmanual behaviors that are important in sign language such as detailed torso, shoulders, neck, and head movements. A limitation of this animation platform was that the facial controls of their animated character were designed as a subset of the FACS systems (Section 8.1.1). To be able to use recordings from multiple human signers to drive the facial expressions of the avatar, we extended EMBR with detailed upper-face controls (eyes, eyebrows, and nose) supporting the MPEG-4 Facial Animation standard (Section 8.1.2).

11.1 Animation Platform Extended with ASL Handshapes

In the process of adopting the EMBR animation engine for ASL we created 71 new ASL handshapes and mapped only a subset of 16 German Sign Language handshapes to ASL. Both new and mapped handshapes were named based on the ASLLVD[18] handshape annotator (Thangali et al., 2011) illustrated in Figure 61.

[17] This section describes joint work with Professor Matt Huenerfauth (Kacorri and Huenerfauth, 2014; Kacorri and Huenerfauth, 2015a; Huenerfauth and Kacorri, 2015b).

[18] ASLLVD is the American Sign Language Lexicon Video Dataset from Boston Universtiy.

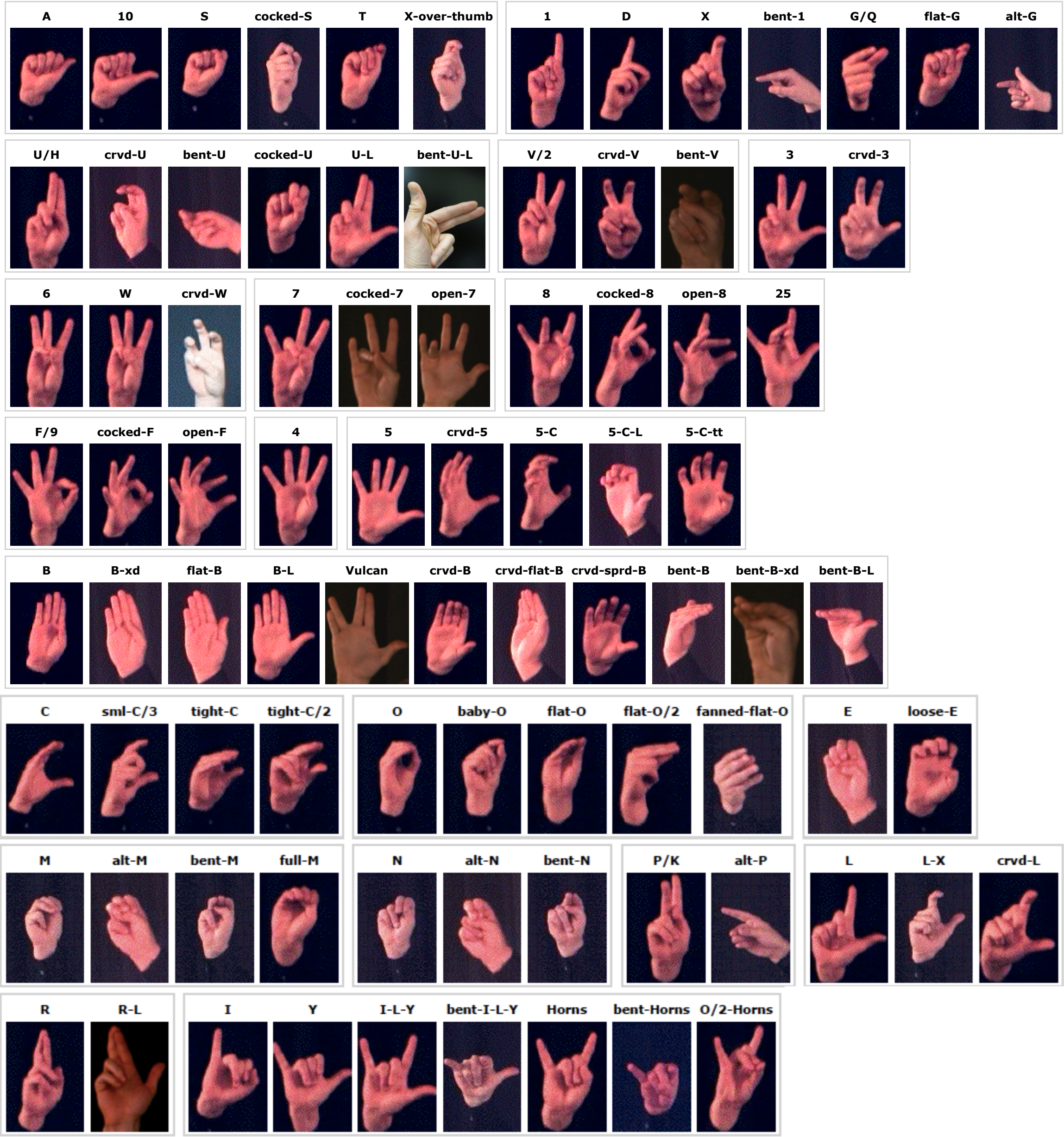


Figure 61: Naming convention for the new ASL handshapes added to EMBR. (Source: ASLLVD)

A 3D-animation artist was instructed on how to create the handshapes in Blender 2.49b, an open source 3D graphics and animation software, and to export them using ChickenExportR91 in a format compatible with the EMBR animation engine (a binary .bam file compatible with Panda3D animation engine). Figure 62 demonstrates a screenshot of the

ASL handshape “W” in Blender. Where the 2D visualization of the ASLLVD collection (Figure 61) was not clear to build a 3D model of the handshape, we provided the artist with video recordings of those handshapes in the 3D space. We then modified the EMBR GUI that is used to create new signs and sentences so that the new handshapes are supported.

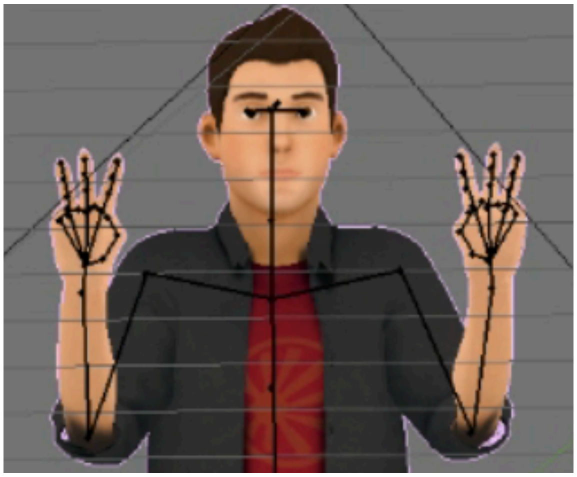


Figure 62: A skeleton view of the avatar in Blender performing the ASL handshape “W”.

11.2 Animation Platform Extended with MPEG4 Facial Controls

To support this research, we had to parameterize the face of our virtual human character so that we can control it by specifying a vector of numbers. Then, a full performance is a stream of such vectors. We needed a parameterization with some properties:

- Values should be invariant across signers with different face proportions who are performing an identical facial expression so we could use recordings from multiple humans in our work.

- The parameterization must be sufficient for controlling the face of a character and should be invariant across animated characters with different facial proportions. This property would allow us to use a variety of characters in our work.
- The parameterization should be a well-documented, standard method of producing and analyzing facial movements. This property would enable our research to be useful for other researchers, using other animation platforms.

The MPEG-4 standard, described in Section 8.1.2, has all the above properties. In short, a face is controlled by setting values for 68 Facial Action Parameters (FAPs), which are displacements of points shown in Figure 63a with the displacements normalized according to scaling factors based on the proportions of the character's face (Figure 63b). This normalization allows for a set of 68 FAPs to produce equivalent facial expression on faces of different sizes or proportions.

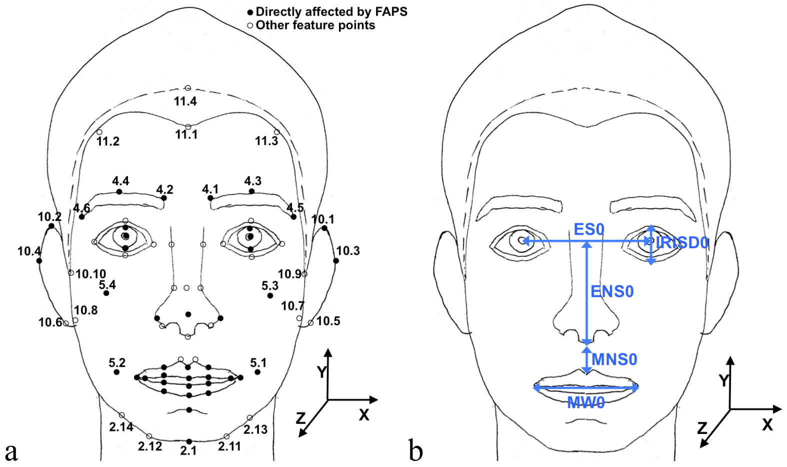


Figure 63: (a) Some feature points and (b) FAPU scale factors in MPEG-4 standard.

11.2.1 Supported MPEG4 Parameters

Our lab with the help of a professional 3D artist extended the character named Max (Heloir et al., 2011) from the open source animation platform EMBR with additional vertices and MPEG-4 FAPs for the upper face controlling the eyes, eyebrows, and nose. EMBR allows for head and torso movements, enables blinking as a background behavior, and has been used for creating sign language animations. As a compliant MPEG4 3D face model (ISO/IEC 14496-2, 1999), the upgraded Max is specified by:

Vertices in 3D face model: MPEG4 requires a minimum of 500 vertices for a pleasant and reasonable face models. To support detailed control of the eyes, eyebrows, and nose, and future controls of the cheeks, mouth, tongue, etc., we added more vertices to Max's face. Figure 64 shows wireframe screenshots of a) original Max and b) enhanced Max.

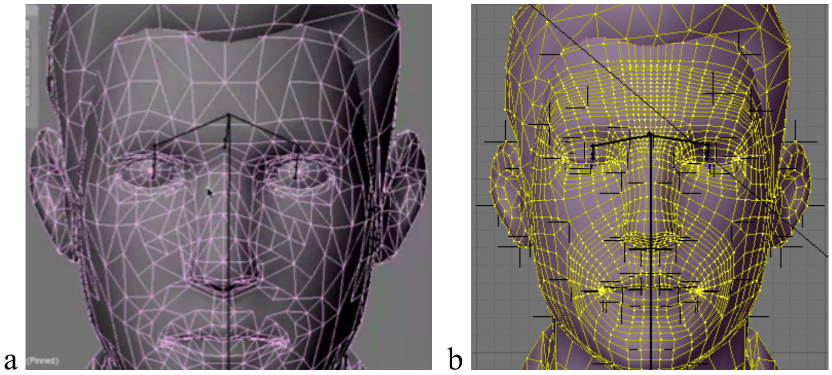


Figure 64: Wireframe Max in (a) the initial EMBR platform and (b) in the enhanced platform (MPEG4 feature points are illustrated with black crosses).

Face Definition Parameters (FDPs): We have defined and visualized in the 3D model the feature points that are directly affected by the FAPs. These are specifically specified by the MPEG4 standard as a subset out of 84 characteristic points in the face that provide spatial reference for defining the FAPs (Figure 63a).

Face Animation Parameter Units (FAPUs): To be able to calculate the maximum displacements for each of the FAPs, we have measured, in Blender Units (animation software distance units), the 5 scale factors defined as fractions of distances between key feature points (Figure 63b). The existence of these FAPUs allows interpretation of the FAPs consistently.

Facial Animation Parameters (FAPs): Out of 68 FAPs defined in MPEG4, we have currently implemented 16 and have mapped an additional 3 to Max's existing head controls resulting in a total of 19 controls (Table 11) for the upper face and head movements involved in the syntactic ASL facial expressions in this thesis. Additional FAPs will be implemented in our lab, not as part of this thesis, to support future research in ASL animation synthesis.

Table 11: Supported MPEG4 controls in the enhanced EMBR platform.

FAP	Name and Description
FAP19-FAP22	Vertical displacements of eyelids. E.g. close_t_l_eyelid: raise or lower from a neutral position the top left eyelid with control values in the (-1, 1) range.
FAP31-FAP36	Vertical displacements of eyebrows. E.g. raise_l_i_eyebrow: raise or lower form a neutral position the left inner eyebrow with control values in the (-1,1) range.
FAP37-FAP38	Horizontal displacements of eyebrows. E.g. squeeze_l_eyebrow: squeeze or move further apart left eyebrow with control values in the (-1,1) range.
FAP48-FAP50	Head orientation. E.g. head_yaw: head yaw angle form the top of the spine. These 3 FAPs (pitch, yaw, roll) were mapped to 6 head controls already defined in the EMBR platform (x, y, and z direction for the tip of the nose and for the top of the head).
FAP61-FAP64	Nose movements (e.g. bend_nose: nose tip displacement to the left or right with control values in the (-1, 1) range.

As part of our enhancements to EMBR, the professional artist modified the surface mesh and constraints to cause the skin on the face to wrinkle automatically as the face controls are modified. The artist also assisted in the design of a lighting scheme for the character to highlight these wrinkles, which are essential to perception of ASL facial movements (Wolfe et al., 2011).

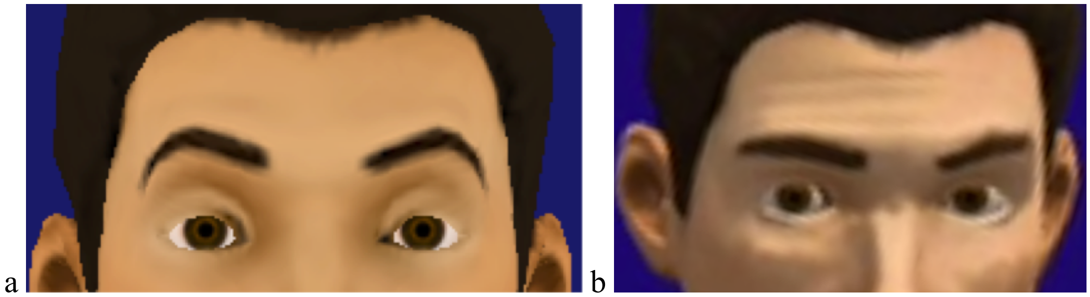


Fig 65: (a) forehead with eyebrows raised before the addition of MPEG-4 controls, facial mesh with wrinkling, and lighting enhancements, (b) eyebrows raised in our current system.

11.2.2 MPEG4 Features-to-Animation Pipeline

A full pipeline from MPEG4 feature extraction to an EMBR facial expression animation generation is required before we investigate more sophisticated approaches for data-driven modeling and synthesis of ASL facial expressions in this thesis. We implemented an intermediate component that converts MPEG-4 data, extracted with the Visage software (Section 10.1), to EMBRscript, the script language supported by the EMBR platform.

Although converting the extracted values of the newly implemented FAPs into the script was a straightforward process, our script generation component dealt with more complicated issues such as:

- **Mapping head orientation (FAP48-50) to EMBR:** Beside their differences in the coordinate systems, Visage and EMBR have entirely different definitions for the head orientation controls, whose details are not all publicly available. In Visage the orientation is given by three values: pitch, yaw, roll. We assumed the rotation order first Pitch, second Yaw, and last Roll, which corresponds to left-handed X-axis rotation, right-handed Y-axis rotation, and a left-handed rotation around Z-axis. In EMBR head orientation is defined by 6 values. The first 3 (x, y, z) describe the direction of the nose tip and the other 3 (x, y, z) describe the direction of the top of the head. Mapping between the two was performed through Euler rotations and coordinate system transformations.
- **Mapping head displacement detected in Visage to EMBR torso controls:** We estimated the torso movements of the avatar (defined as spine orientation in EMBR) based on the Visage head location values (x, y, z) that correspond to displacement of the head from a neutral position in the 3D space.
- **Adjusting eye-gazing to the viewer**: While eye-gazing in sign language is very important and often governed by linguistic rules, we are not investigating this complex aspect in the current thesis. We assume a simplistic case where the signer's gaze is directed to the viewer across the sentence. Since the avatar's head orientation is changing during the performance of a signed sentence we compensate for these movements such that a viewer-directed eye-gaze is approximately maintained. We are currently not considering the eyelid movements that are extracted from Visage. Instead the auto-blinking behavior incorporated in the original EMBR is enabled.

- **Time-adjustment of the extracted facial movements to the manual movements:** Synchronization of face and head movements to the manual signs was implemented with a simplistic stretching or compression (resampling with cubic interpolation) of the frame duration (originally 33msec) of the recorded movements to match the duration of the animated signs where the facial expression occurs.

11.3 Evaluation of the Extended Animation Platform

To evaluate whether the extended EMBR animation platform has sufficient expressivity to convey ASL sentences and facial expressions, as judged by ASL signers, we conducted a user study. We hypothesize:

H1: Our animation platform is sufficiently expressive such that it could produce understandable facial expressions for ASL animations.

H2: Our animation platform is sufficiently expressive such that it could produce explicitly recognized facial expressions for ASL animations.

To investigate our hypotheses, a total of 14 ASL signers, viewed animations of short stories and then answered comprehension questions and scalar-response questions as to whether they noticed the correct facial expression. The 18 stimuli stories were selected from our publicly released ASL stimuli collection, details appear in Section 4.2. They included three types of syntactic facial expressions: Yes/No Question, WH-word Question, and Negative (with 6 stimuli stories per type) with codenames Y2-Y7, W1-W6, and N1-N6, respectively.

As in Part I, responses for “Notice” questions were in 1-to-10 ranges from “yes” to “no” in relation to how much participants noticed an emotional, negative, questions, and topic facial expression during the story. As described in Section 4.2, four comprehension questions were engineered for each of the stories in such a way so that the correct answer depends on understanding the facial expression. Responses for the comprehension questions were given on a 7-point scale from "definitely no" to "definitely yes." Participants could choose "I'm not sure" instead of answering.

In a between-subjects design, we compared two types of animations with identical hand movements but differing in their face, head, and torso movements: (a) ***driven*** by a recording of a human performing that type of facial expression or (b) face, head, and torso movements are static and ***neutral*** throughout the story. The type "b" animations therefore did not reveal any of the capabilities of the new MPEG-4 controls or skin-wrinkling of the EMBR character. Face and head movements for the ***driven*** animations were created using the pilot data described in Section 10.2 and our MPEG4 features-to-animation pipeline. Figure 66 illustrates the 2 versions of a Yes-No Question story (codename Y3). The video size, resolution, and frame-rate for all stimuli were identical. The hand movements in both versions were identical and were created by ASL signers. In the recording-driven animations, facial movements were added during the portion of the story where the facial expression of interest should occur; the rest of the story had a static neutral face. Example stimuli animations from our study are available at: http://latlab.ist.rit.edu/2014assets.

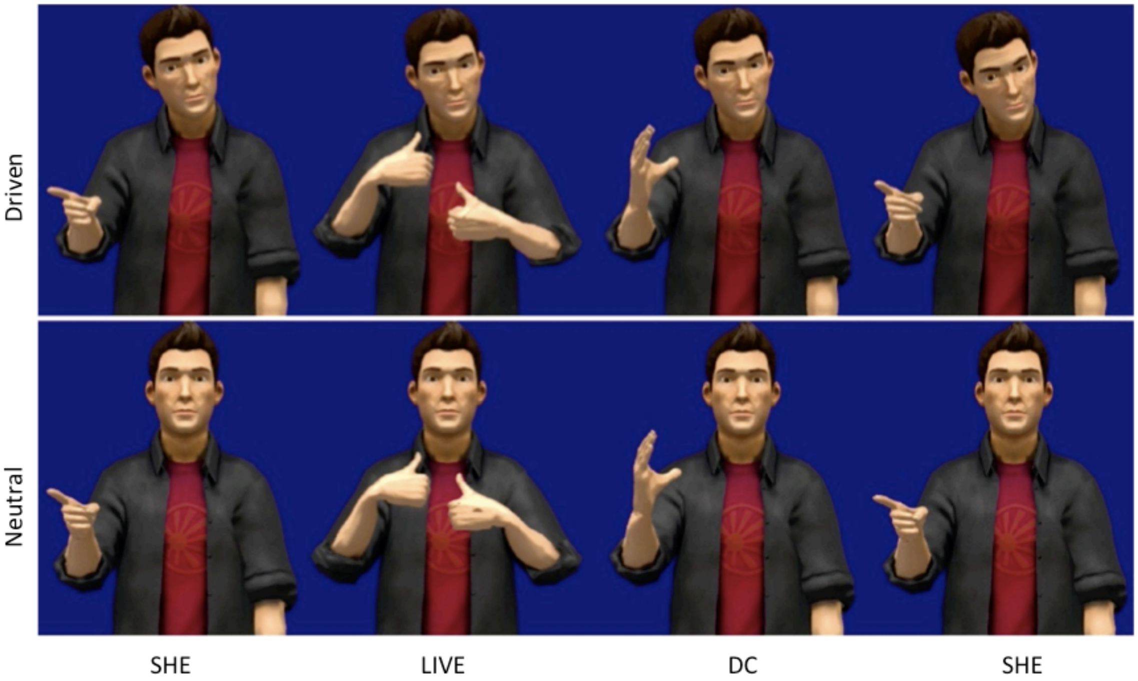


Figure 66: Screenshots from a human-recording-driven and neutral Yes/No-question stimulus in the study.

At the beginning of the study, participants viewed a sample animation, to familiarize them with the experiment. An ASL signer conducted all of the experiments in ASL. In Part I, we discussed the importance of participants being ASL signers and the study environment being ASL-focused with little English influence; we used the questions developed in (Huenerfauth et al., 2008) to screen for fluent ASL signers. For this study, ads were posted on New York City Deaf community websites asking potential participants if they had grown up using ASL at home or had attended an ASL- based school as a young child. Of the 14 participants recruited for the study, 12 participants learned ASL prior to age 9. The remaining 2 participants had been using ASL for over 11 years, learned ASL as adolescents, attended a university with classroom instruction in ASL, and used ASL daily to communicate with a significant other or family member. There were 7 men and 7 women of ages 23-42 (avg. 30.2).

In support of our hypotheses, Figure 67 displays the scores of the comprehension questions and the question that asked if participants noticed the correct facial expression. Medians are shown above each boxplot. There was a significant difference in the Notice scores (Mann-Whitney test used since the data was not normally distributed, p<0.00014). There was also a significant difference in the comprehension question scores (t-test, p<0.000001). Note that comprehension scores depend on the difficulty of the questions asked; so, such scores are meaningful only for comparison within a single study.

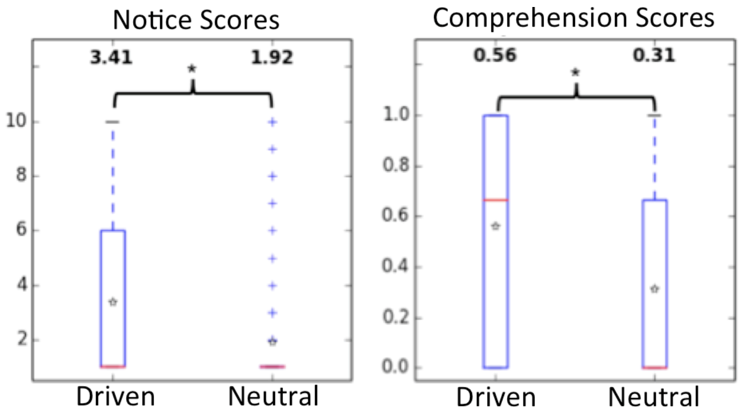


Figure 67: Notice and Comprehension scores for animations with Driven and Neutral facial expressions.

These results indicate that our animation system, the extended EMBR platform, is a useful platform for evaluating our on-going research on designing new methods for automatically synthesizing facial expressions of ASL.

This finding is significant because it allows for research on ASL facial expression to take advantage of prior tools and research on facial animation with MPEG-4. In order to evaluate the expressivity of our character, we used human recordings in this study; however, in future sections, we will be investigating learning-based models for *automatic* synthesis of ASL facial expressions.

These results, presented above, support ***RQ7***:

RQ7: *Is our MPEG4-enhanced animation platform sufficiently expressive such that it could produce facial expressions for ASL animations that are understandable and explicitly recognized by ASL signers?*

11.4 Comparison of the Extended EMBR Animation Platform to Our Prior ASL Animation Platform

Since we had conducted several years of research in our laboratory using a previous ASL animation platform (and the methodological studies in Part I of this dissertation), we compare the *new* extended-EMBR avatar to the *old* VCOM3D avatar, in regard to their understandability and naturalness. This section presents the results of experiments with deaf participants evaluating animations from both of these platforms. This comparison will enable future researchers to compare our published results before and after this platform change, and it will allow us to better evaluate whether our new avatar is sufficiently understandable to support our facial expression modeling work.

As discussed in Part I, our prior animation platform was based on a commercially available ASL authoring tool, VCOM3D Sign Smith Studio, which allows users to produce ASL sentences by arranging a timeline of animated signs from a prebuilt or user-defined vocabulary. The software includes a library of facial expressions that can be applied over a single sign or multiple manual signs, as shown in Figure 68. While this finite repertoire covers adverbial, syntactic, and emotional categories of facial expressions, the user cannot modify the intensity of the expressions over time, nor can multiple facial expressions be combined to co-occur.

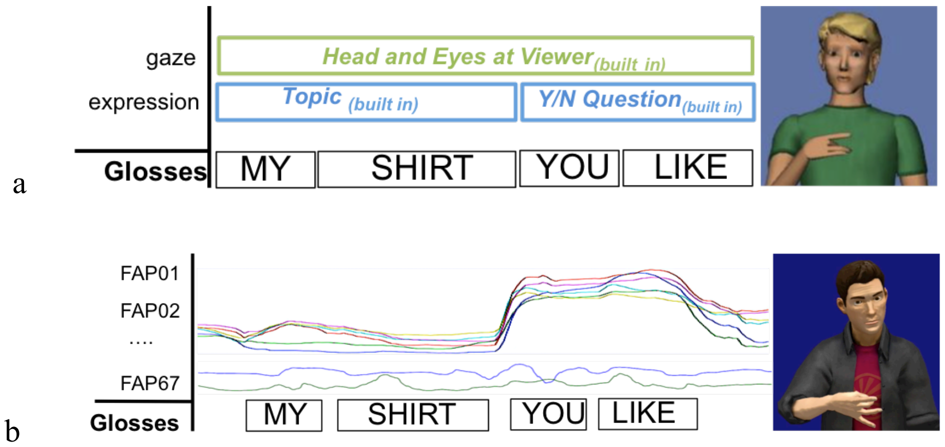


Figure 68: This graphic depicts the same timeline of an ASL sentence consisting of four signs (shown in the "Glosses" row) with co-occurring facial expressions in (a) VCOM3D and (b) extended EMBR.

Figure 68 demonstrates the timeline for an ASL sentence in the two animation systems. In Figure 68a, the facial expressions that are available from the software's built-in repertoire, are specified by the user and are shown in the "expression" row which is in parallel to the hand movements ('Glosses' row). In particular, the creator of this timeline has specified that a "Topic" facial expression should occur during the first two words and a "Yes/No Question" facial expression during the final two. In Figure 68b, the facial expressions are illustrated as curves plotted above the 'Glosses' row, each of which depicts the changing values of a single MPEG-4 parameter that governs the movements of the face/head. For instance, one parameter may govern the height of the inner portion of the signer's left eyebrow.

11.4.1 Comparison of New versus Old Avatar Platform

To compare the naturalness and understandability of the ASL facial expressions synthesized by the two animation platforms, we analyzed data from two user studies in which signers evaluated animations from each. The multi-sentence stimuli shown and the questions asked in the studies were identical: specifically, the hand movements for both avatars in those sentences are nearly identical (differences in avatar body proportion contributes to some hand movement differences). Further, for each platform, the stimuli were shown in two versions: with facial expressions ("Expr.") and without facial expressions ("Non"). Thus, there were a total of four varieties of animations shown to participants. Participants were asked to report whether they noticed a particular facial expressions being performed by the avatar and to answer comprehension questions about the stimuli.

In addition to comparing results across the two platforms ("old" vs. "new"), we are also interested in our ability to see a difference between the animations with facial expressions ("Expr.") and those without facial expressions ("Non") in each platform. (If we can't see any difference in "notice" or "comprehension" scores when we animate the face of the character, this suggests that the platform is not producing clear sign language facial expressions.) In particular, we hypothesize:

H1: When comparing our "new" and "old" animation platforms, we expect the "notice" scores will be statistically *equivalent* to the corresponding scores ("Expr." or "Non") between both platforms.

H2: When comparing "Expr." animations with facial expressions and "Non" animations without facial expressions, we expect that our new platform will reveal differences in "notice" scores *at least as well as* our earlier platform.

To explain our reasoning for H1 and H2: For the "Non" case, there is no reason to think that the change in virtual human platform should affect the scores since the face does not move during these animations. For the “Expr.” case, while the new platform may have more detailed movements, there is no reason to think that people would notice face movements more in our new character, even if they were more detailed.

We also hypothesize the following, in regard to the “comprehension” scores:

H3: When comparing our “old” and “new” animation platforms, comprehension scores assigned to “Non” animations without facial expressions will be statistically *equivalent* between both platforms.

H4: When comparing “old” and “new” platforms, comprehension scores assigned to “Expr.” animations with facial expressions in our new platform will be statistically *higher* than those for the old platform.

H5: When comparing “Expr.” animations with facial expressions and “Non” animations without facial expressions, we expect our new platform to reveal differences in comprehension scores *at least as well as* our old platform.

To explain our reasoning: When comparing the “Non” versions (no face movements), we expect the comprehension scores to be similar between both platforms because the hand movements are similar. However, we expect the animations with facial expressions created in the new platform to be more comprehensible, given that the new platform should be able to reproduce subtle movements from human signers.

11.4.2 Experiment Setup and Results

While two different animation platforms were used to generate the animations shown to participants, the “script” of words in the stimuli for both studies was identical, adopted from the stimuli collection in described in Chapter 4. In particular there were nine multi-sentence stimuli including three categories of ASL facial expressions: yes/no questions, negative, and wh-word questions with codenames N1, N2, N3, W2, W3, W5, Y3, Y4, and Y5 (details on Appendix A).

In both studies, a fully-factorial design was used such that: (1) no participant saw the same story twice, (2) order of presentation was randomized, and (3) each participant saw half of the animations in each version: i) without facial expressions (“Non”) or ii) with facial expressions (“Expr”). All of the instructions and interactions were conducted in ASL by a deaf signer, who is a professional interpreter. Part of the introduction, included in the beginning of the experiment, and the comprehension questions of both studies were presented by a video recording of the interpreter.

Animations generated using our old animation platform were shown to 16 ASL signers (in Chapter 5). Of 16 participants, 10 learned ASL prior to age 5, and 6 attended residential schools using ASL since early childhood. The remaining 10 participants had been using ASL for over 9 years, learned ASL as adolescents, attended a university with classroom instruction in ASL, and used ASL daily to communicate with a significant other or family member. There were 11 men and 5 women of ages 20-41 (average age 31.2). Similarly, animations generated using our new animation platform were shown to 18 ASL signers (as discussed in Section 11.3), with the following characteristics: 15 participants learned ASL prior to age 9, the remaining 3 participants learned ASL as adolescents, attended a university with classroom instruction in ASL, and used ASL daily to communicate with a significant other or family member. There were 10 men and 8 women of ages 22-42 (average age 29.8).

After viewing each animation stimulus one time, the participant answered a 1-to-10 scale question as to whether they noticed a facial expression during the animation; next, they answered four comprehension questions about the information content in the animation (using a 1-to-7 scale from "Definitely Yes" to "Definitely No").

Figure 69Figure 70 display the distribution of the Notice and Comprehension scores for the "Expr." and "Non" types of stimuli in the studies. (Box indicates quartiles, center-line indicates median, star indicates mean, whiskers indicate 1.5 inter-quartile ranges, crosses indicate outliers, and asterisks indicate statistical significance. To aid the comparison, mean values are added as labels at the top of each boxplot.) Labels with the subscript "(OLD)" indicate animations produced using our prior animation platform VCOM3D, and labels with the subscript "(NEW)" indicate animations produced using our new animation platform EMBR.

Since hypotheses H1 and H3 require us to determine if pairs of values are statistically equivalent, we performed "equivalence testing" using the two one-sided test (TOST) procedure (Schuirmann, D.J. 1987), which consists of: (1) selecting an equivalence margin theta, (2) calculating appropriate confidence intervals from the observed data, and (3) determining whether the entire confidence interval falls within the interval (-theta, +theta). If it falls within this interval, then the two values are deemed equivalent. We selected equivalence margin intervals for the "notice" and comprehension scores based on their scale unit as the minimum meaningful difference. This results intervals of (-0.1, +0.1) for the 1-to-10 scale "notice" scores and (-0.14, +0.14) for the 1-to-7 scale comprehension scores. Having selected an alpha-value of 0.05, confidence intervals for TOST were evaluated using Mann-Whitney U-tests for scalar "notice" responses and t-tests for comprehension responses. (Non-parametric tests were used for the scalar responses because they were not normally distributed.)

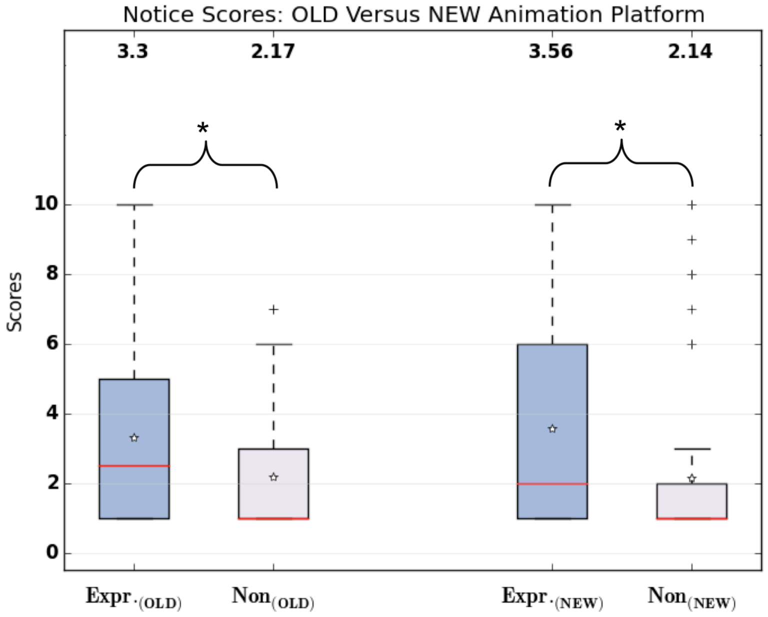


Figure 69: Notice Scores for OLD and NEW Animation Platform.

Hypothesis H1 would predict that the "notice" scores for both "Non" and "Expr." stimuli would be unaffected by changing our animation platform. The following confidence intervals were calculated for TOST equivalence testing: (-0.00002, +0.00003) for Non(OLD) vs. Non(NEW) and (-0.000008, 0.00006) for Expr.(OLD) vs. Expr.(NEW). Given that these intervals are entirely within our equivalence margin interval of (-0.1, +0.1), we determine that the pairs are equivalent. Thus, hypothesis H1 is supported.

Hypothesis H2 would predict that evaluations conducted with our new animation platform are able to reveal with-vs.-without facial expressions differences in "notice" scores at least as well as our old animation platform. Thus, the statistical test is as follows: If there is a pairwise significant difference between Expr.(OLD)-Non(OLD), then there must be a statistically significant difference between Expr.(NEW)-Non(NEW). In support of H2, Figure 69 illustrates significant difference between both pairs on the basis of Kruskal-Wallis and post hoc tests (p<0.05). We also observed that the magnitude of this difference is bigger in our new platform (d: 33, p-value: 0.001) than it is in our prior animation platform (d: 24, p-value: 0.02). Thus, hypothesis H2 is supported.

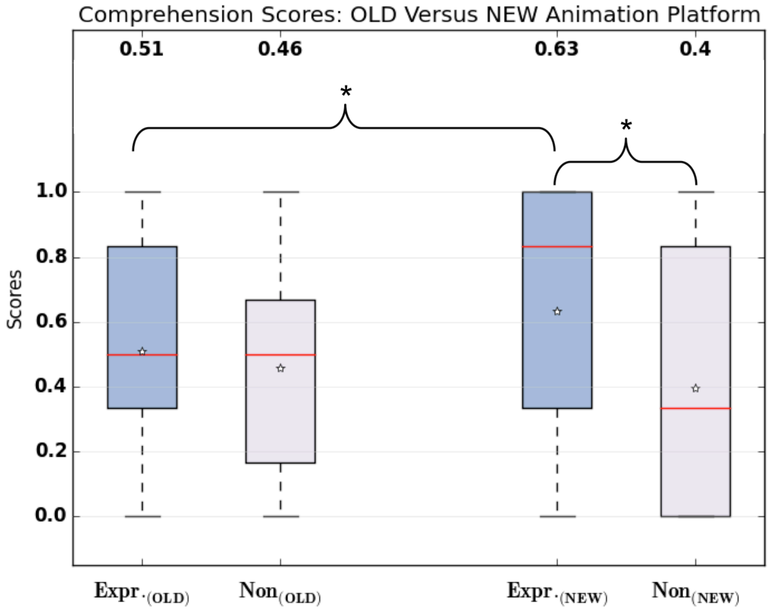


Figure 70: Comprehension Scores for OLD and NEW Animation Platform.

Hypothesis H3 would predict that the comprehension scores for "Non" stimuli would be unaffected by changing our animation platform. The following confidence intervals were calculated for TOST equivalence testing: (+0.002, +0.119) for Non(OLD) vs. Non(NEW). Given that these intervals are within our equivalence margin interval of (-0.14, +0.14), we determine that the pairs are equivalent. Thus, H3 is supported.

Hypothesis H4 predicted that when considering the comprehension questions in evaluations conducted with our new animation platform the “Expr.” stimuli would receive higher scores than the “Expr” scores for our old platform. As illustrated in Figure 70, we observed a significant difference ($p<0.05$) between Expr(OLD)-Expr(NEW) comprehension scores by performing one-way ANOVA. Thus, H4 is supported.

Hypothesis H5 predicted that evaluations conducted with our new animation platform would reveal with-vs.-without facial expressions differences in comprehension scores at least as well as our old animation platform. Figure 70 illustrates a significant difference when comparing Expr.(NEW)-vs.-Non(NEW) comprehension scores for our new animation platform but not for the prior platform. Significance testing was based on one-way ANOVA and post hoc tests ($p<0.05$). Thus, H5 is supported.

Overall in this section we found that our new EMBR platform was able to produce animations that achieve similar scores to our old VCOM3D platform (when no facial expressions are included) or higher scores (when facial expressions are included). We also found that our new platform was able to produce animations with facial expressions that achieved significantly higher scores than animations without facial expressions. The experimental user study in Chapter 13 will make use of this new EMBR platform and the analysis in this section can allow readers to understand how the results and benchmark baselines published in prior work would compare to results that are published using the new platform.

Chapter 12 ASL Facial Expression Synthesis with Continuous Profile Models[19]

A novel aspect of our proposed facial-expression animation synthesis approach is the use of data from multiple recordings of an ASL signer to drive our models. The advantage of considering data from multiple recordings is that we can identify the common elements of a facial expression performance that are essential, while generalizing across multiple performances so as to avoid idiosyncratic aspects of a single performance. This would thereby allow us to produce a more generalizable model of facial expression performance for ASL.

In this chapter we use machine-learning approaches, such as probabilistic generative models, that can automatically uncover the underlying trace from multiple recordings of the same type of ASL facial expression. We train our models for each of the granular categorization of the linguistically diverse dataset, described in Section 10.3 and demonstrate that they are identifying a curve that seems to be a good representative of the training data set. To further assess our approach we conduct both a metric evaluation and a user study where signers evaluate animations with facial expressions produced by our enhanced EMBR platform with MPEG-4 facial controls (Chapter 13).

Multiple recordings of an ASL signer performing different sentences with the same type of facial expression share an underlying representation of the observable face/head movements. However, the time axis for each of these observed movements could differ from the underlying

[19] This chapter describes joint work with Professor Matt Huenerfauth and Ali Raza Syed, a Ph.D. student at CUNY.

"stereotypical" performance of the facial expression. Such differences may arise from characteristics of the sentence itself or factors related to a particular recording of a sentence. Such factors might have led to the time axis of the performance being shifted (e.g., compressed, expanded, or distorted in other complex non-linear ways), depending on the manual signs' timing, overall length of the sentence, the context, etc. In some cases, systematic variability in the scale of the extracted movements (i.e., the magnitude of the movements) is observed even when sentences replicated by the same signer have the same hand movements occurring with the facial expression. To model the underlying characteristic facial and head movements that correspond to a particular syntactic ASL facial expression while accounting for both variation in the timing and amplitude, we investigate a class of probabilistic generative models called Continuous Profile Model (CPM, Listgarten, 2007).

12.1 Continuous Profile Model

Continuous Profile Model (CPM), previously evaluated on biological and speech signals (Listgarten et al., 2004), can align a set of related time series data while simultaneously accounting for changes in the amplitude of the time series. With the assumption that a noisy, stochastic process generates the observed time series data, the approach automatically infers the underlying noiseless representation of the data, the so-called "latent trace." Figure 71 illustrates an example of multiple speech time series in the unaligned and aligned space, where the latent trace is uncovered using CPM.

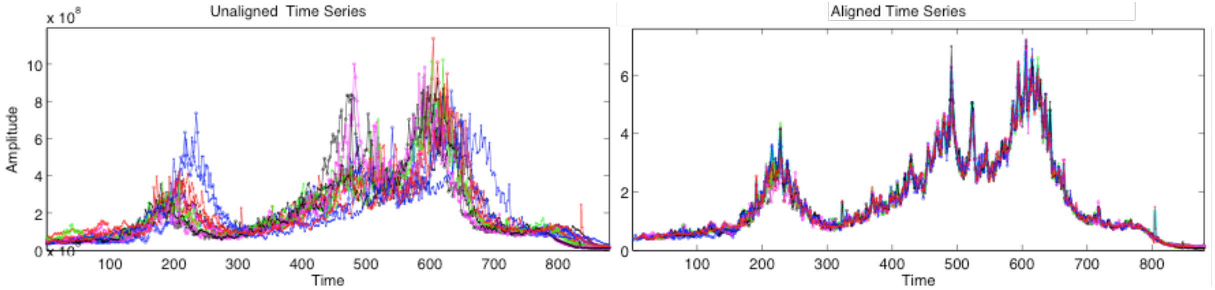


Figure 71: An example of unaligned and aligned (proteomic) time series using CPM. In the lower image, the latent trace is illustrated in cyan, mostly obscured by foreground lines. (Source: Listgarten et al., 2004)

Continuous Profile Models (CPMs) build on Hidden Markov Models (HMMs, Poritz, 1988) and share similarities with Profile HMMs which augment HMMs by two constrained-transition states: 'Insert' and 'Delete' (emitting no observations). Similar to the Profile HMM, the CPM has strict left-to-right transition rules, constrained to only move forward along a sequence. Figure 72 includes a visualization we created, which illustrates the graphical model of a CPM.

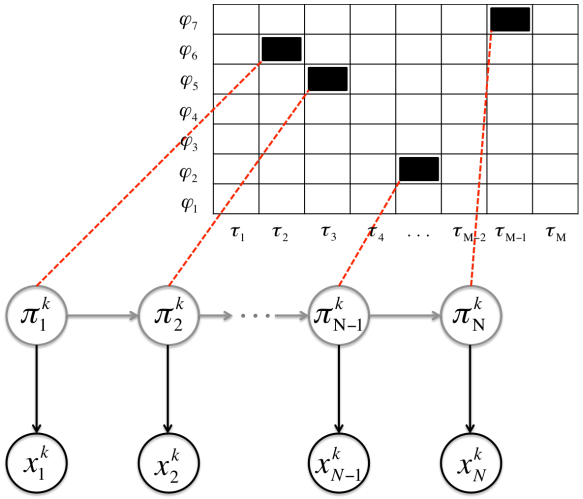


Figure 72: Graphical model depiction of a CPM for a particular series x^k. Top nodes, are the hidden state variables underlying each observation, x_i^k. The table illustrates the state-space: time-state/scale-state pairs mapped to the hidden variables, where time states belong to the integer set (1… M) and scale states belong to an ordered set with 7 evenly spaced scales in logarithmic space (as in prior CPM experiments in Listgarten et al., 2004).

Given a set K of observed time series, $\vec{x}^k = \left(x_1^k,\ x_2^k, \dots, x_N^k\ \right)$, CPM assumes there is a latent trace $\vec{z} = (z_1, z_2, \dots, z_M)$. While not a requirement of the model, the length of the time series data is assumed to be the same (N) and the length of the latent trace used in practice is $M = (2 + \varepsilon)N$, where the ideal M would be very large relative to N or infinite to allow precise mapping between observed data and an underlying point on the latent trace. The higher temporal resolution of the latent trace also accommodates flexible alignments by allowing an observational series to advance along the latent trace in small or large jumps (CPM, Listgarten, 2007).

For each observed time series $\vec{x}^k$ the state sequence $\vec{\pi}^{\kappa}$ determines the subsampling and local scaling of the latent trace to generate this observation. A hidden state π_i^{κ} maps to a state pair: time state and scale state (τ_i, φ_i), as illustrated in Figure 72. The time states have transition probabilities $p^k(\tau_i|\tau_{i-1})$ and the scale states have transition probabilities $p(\varphi_i|\varphi_{i-1})$. Therefore the state transition probability for a hidden variable in CPM is given as $p(\pi_i|\pi_{i-1}) = p(\varphi_i|\varphi_{i-1})p^k(\tau_i|\tau_{i-1})$. Each x_i^k in Figure 72 is assumed to be emitted by a Gaussian distribution with mean μ_i^k and standard deviation σ: $x_i^k \sim \mathcal{N}\left(\mu_i^k, \sigma\right)$. The mean is estimated by $\mu_i^k = z_{\tau^k}\varphi_i^k u^k$, where u^k is a real-valued scale parameter specific to a time series k to allow for a global scaling between time series k and the latent trace.

While CPM covers a class of generative models, in this dissertation when referring to CPM we actually refer to a specific type of CPM, the single-class EM-CPM where the training is performed on data from the same class using the expectation-maximization (EM) algorithm (Dempster et al., 1977).

12.2 Obtaining the CPM Latent Trace on ASL Facial Expressions

We applied the CPM model to time align and coherently integrate time series data from multiple ASL facial expression performances that belong to the same subcategory of facial expression described in Section 10.3 (a detailed list of the set of videos used to train each CPM is available in Appendix C). We hypothesize that the inferred 'latent traces' can be used to drive ASL animations with facial expressions in those subcategories. In this section we describe our experiments to train the CPM and to obtain the latent traces using the freely available implementation of the CPM[20] in MATLAB, Version 8.5.0.197613 (R2015a).

Table 12: Training data and the obtained latent traces for each of the CPM models on ASL facial expression subcategories. The length of time series (N) corresponds to the duration in video frames of the longest example in the data set. Both the obtained latent trace and each time series in the training set has 14 dimensions corresponding to the extracted features from Visage: 'Head x', 'Head y', 'Head z', 'Head pitch', 'Head yaw', 'Head roll', 'raise_l_i_eyebrow', 'raise_r_i_eyebrow', 'raise_l_m_eyebrow', 'raise_r_m_eyebrow', 'raise_l_o_eyebrow', 'raise_r_o_eyebrow', 'squeeze_l_eyebrow', 'squeeze_r_eyebrow'. The latent trace has a time axis of length M, which is approximately double the temporal resolution of the original training examples.

CPM Models	*Training Data* *(#Examples x N x #Features)*	*Latent Trace* *(M x #Features) where* $M = (2 + \varepsilon)N$
YesNo_A	9 x 51 x 14	105 x 14
YesNo_B	10 x 78 x 14	160 x 14
WhQuestion_A	4 x 24 x 14	50 x 14
WhQuestion_B	8 x 41 x 14	84 x 14
Rhetorical_A	2 x 16 x 14	33 x 14
Rhetorical_B	8 x 55 x 14	113 x 14
Topic_A	29 x 29 x 14	60 x 14
Topic_B	15 x 45 x 14	93 x 14
Negative_A	16 x 67 x 14	138 x 14
Negative_B	25 x 76 x 14	156 x 14

[20] CPM MATLAB Toolbox: http://www.cs.toronto.edu/~jenn/CPM/.

In our experiments, each set of time series data corresponds to one of the subcategories in Section 10.3 as shown in Table 12. For each time series set, all the time series training examples are stretched (resampled using cubic interpolation) to meet the length of the longest example in the set. All experimental results reported in this thesis are based on raw time series, data extracted by Visage, as inputs for the CPM.

One of the alternatives investigated in Listgarten (2007) for regularizing the latent trace is a smoothing parameter (λ), with reasonable values for λ being dataset-dependent. To estimate a good smoothing parameter for our latent traces we experimented with five out of ten subcategories: YesNo_B, WhQuestion_B, Rhetorical_B, Topic_B, and Negative_B. For each subcategory, we used half of the time series data as a hold out set. We found that $\lambda = 4$ and number of iteration 3 result in a latent trace curve that captures the shape of the data well. We kept the other parameters of the CPM unchanged from their defaults[21].

To demonstrate our experiments, we present Figure 73Figure 74 in this chapter that focus on one of the subcategories, Rhetorical_B. In Figure 73, we illustrate the training set, before and after the alignment and amplitude normalization with the CPM, and the obtained latent trace for this subcategory. Figure 73a and Figure 73b illustrate each of the 8 training examples with a subplot extending from $[0, N]$ in the x-axis, which is the observed time axis in video frames. Figure 73c illustrates the learned latent trace with a subplot extending from $[0, M]$ in

[21] *USE_SPLINE=0*: if set to 1, uses spline scaling rather than HMM scale states
oneScaleOnly=0: no HMM scale states (only a single global scaling factor is applied to each time series.)
extraPercent (ε) = 0.05: some extra slack on the length of the latent trace M, where $M = (2 + \varepsilon)N$.
learnStateTransitions=0: whether to learn the HMM state transition probabilities
learnGlobalScaleFactor=1: learn single global scale factor for each time series

the x-axis, which is the latent time axis. While the training set for this subcategory is very small and has high variability, upon visual inspection of Figure 73a, Figure 73b, and Figure 73c, we can observe that the learned latent trace shares similarities with most of the time series in the training set without being identical to any of them.

Given that head location, head orientation, and eyebrow movements are on different scales, to have a better understanding what happens at a feature level, we plot the examples together per feature/dimension before (Figure 74a) and after the CPM (Figure 74b). Figure 74c shows the latent trace learned per feature. Based on our description of the syntactic ASL facial expressions in Section 2.1, we expect that during Rhetorical questions signer's eyebrows to be raised and the head to be tilted backwards and to the side. We observe this to be true for the latent trace. For example, the eyebrows are raised (Figure 74c, plots 7-12), where left inner, middle, and outer eyebrow points (Figure 74c, plots 7, 9, 11) share similarities with right inner, middle, and outer eyebrow points (Figure 74c, plots 8, 10, 12). (Note how the height of the lines in those plots is raised, which indicates increased eyebrow height.) Also for the horizontal displacement of the eyebrows the learned latent trace maintains the mirroring relationship for left and right (Figure 74c, plots 13-14). (Note the general tendency for the lines in plot 13 to increase in height as the lines in plot 14 decrease in height, and vice versa.)

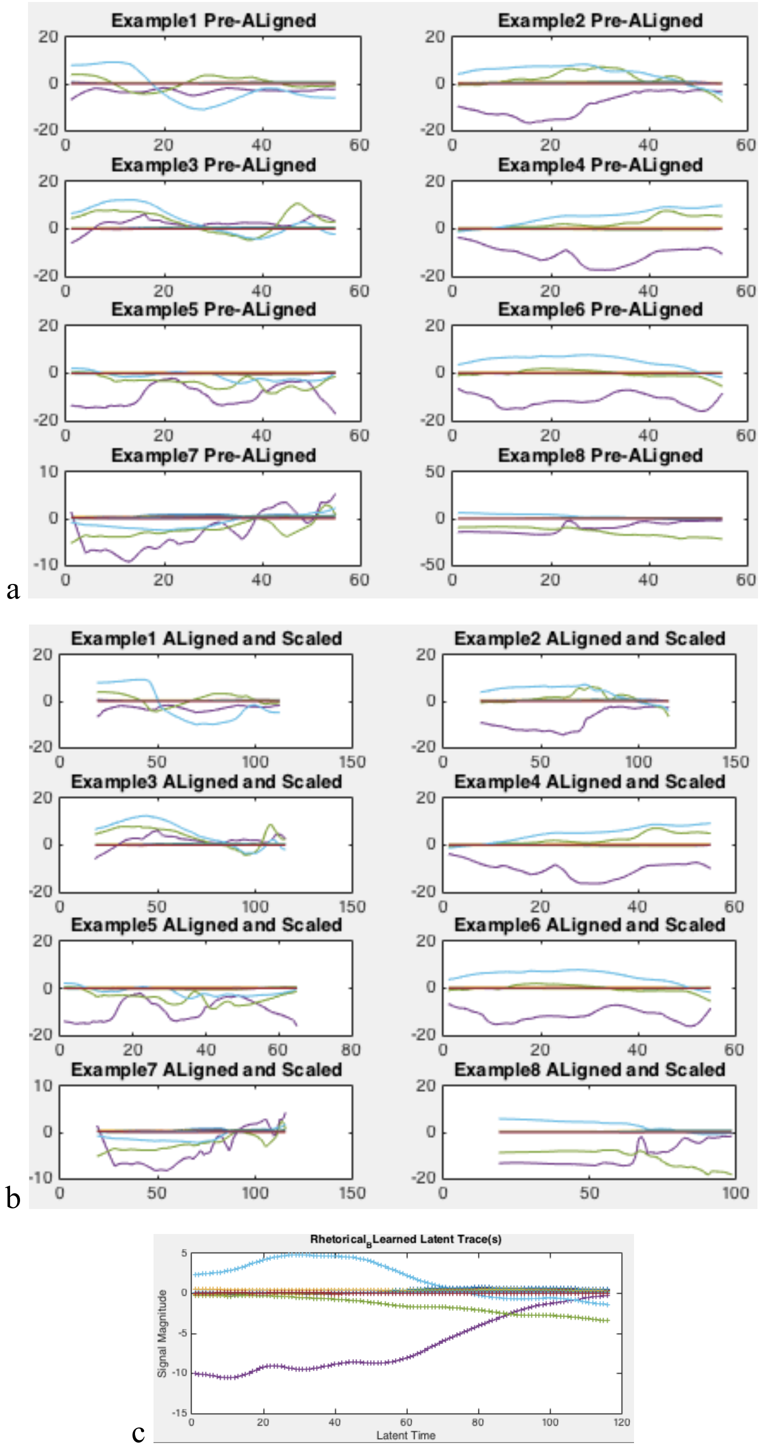


Figure 73: Rhetorical_B (a) training examples before CPM (lines represent values of the 14 face features over time), (b) training examples after CPM (time alignment and rescaling), and (c) the obtained latent trace (based upon all eight examples). Note that the y-axis limits differ across the subplots.

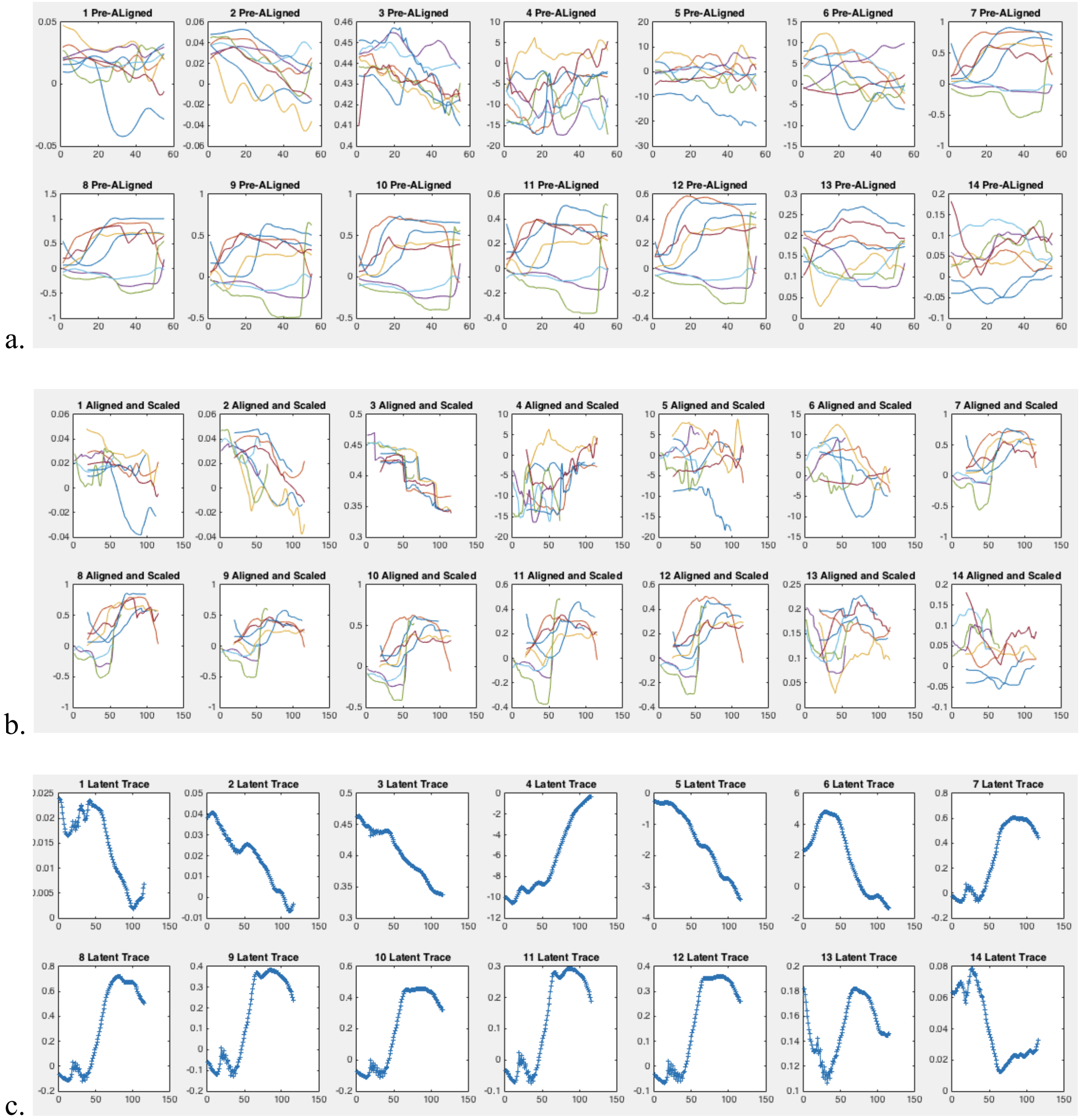


Figure 74: Rhetorical_B per head/face-feature time series in the a) training examples before CPM, and b) training examples after CPM, and c) obtained latent trace.

12.3 Comparison of CPM Latent Trace and Training Data with DTW

In this thesis we propose to use the latent trace, obtained from training a CPM model on multiple recordings of an ASL facial expression, as a representative of that facial expression.

Before proceeding with this CPM model as the basis for generating ASL facial expressions, we are interested in answering Research Question ***RQ8:*** *Is a Continuous Profile Model (CPM) able to produce a latent-trace curve that is representative of a set of ASL facial expressions, which had been provided as training data to the model?* To investigate this question, we compare the similarity of the latent trace to all the examples in the training set, using Dynamic Time Warping (DTW) as a measure of similarity. In this way, we can explore whether the latent trace is near the "centroid" of the original set of data examples upon which it was based.

12.3.1 Dynamic Time Warping (DTW)

DTW arose in the field of speech recognition (Sakoe and Chiba, 1978; Velichko and Zagoruyko, 1970) as a generalization of algorithms for comparing series of values with each other. DTW sums the distance between the individual aligned elements of two time series, which are locally stretched or compressed, to maximize their resemblance. Unlike the Euclidean distance, it can serve as a measure of similarity even for time series of different length. An advantage of DTW over other cross-correlation similarity measures is that it allows for non-linear warping of the time axis.

Figure 75 illustrates a DTW example where the similarity of eyebrow movements between two performances of a Negative ASL facial expression is calculated as the normalized distance between the time series of the MPEG-4 feature "raise_l_i_eyebrow".

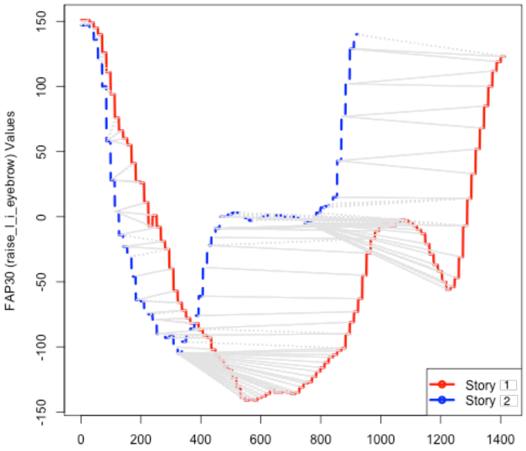


Figure 75: *Example of DTW alignment between the "raise_l_i_eyebrow" values detected in human recordings of two ASL stories containing a Negative facial expression. The duration of the facial expression in Story 1 and Story 2 is 1414 and 924 frames, respectively and their calculated normalized distance was found to be 8.76.*

12.3.2 DTW Distance of the Latent Trace from the Training Data

Given that DTW has been used as a similarity scoring technique for facial animation, e. g., for the retrieval of facial animation based on a key-pose query (Ouhyoung et al., 2012) and spatio-temporal alignment between face movements recorded from different humans (Zhou and Torre, 2009), we adopt this approach in our comparison of the latent trace in our CRM models to the training time series data for each of the ASL facial expression subcategories in Table 12.

We want to see whether the obtained latent trace for a particular subcategory of ASL facial expression appears to be a good representative of the recorded examples included in the training set. To do so, we construct a new set that we call the 'comparison set', which includes union of the time series from the training set and the obtained latent trace. We then use DTW to obtain the distances between all the pairs in the comparison set. We define the "centroid" of the time series as the series with the minimum sum distance from all the other time series in the set.

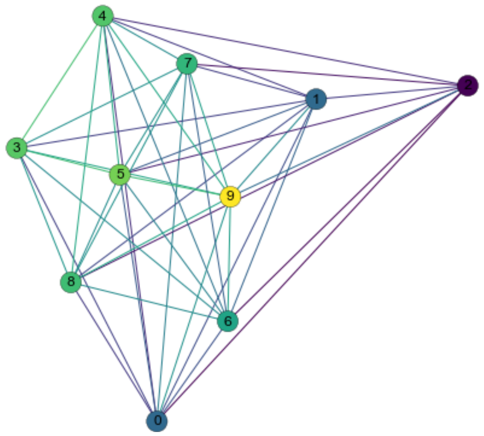
To calculate the distance between two time series in the comparison set we used the implementation of multivariate DTW in R (Giorgino, 2009). It computes a global alignment with minimum distance normalized for path length using Euclidean as a local distance. Computing global alignments means that the time series' heads and tails are constrained to match each other. While the MPEG4 features describing eyebrow movements are all in the [-1, 1] scale, the head location and orientation are on different scales. We scale them in the [-1,1] range by diving with the maximum values observed in the comparison set.

In our experiments, we found the latent trace to be the centroid for all 10 subcategories of ASL facial expressions. We visualize our findings in Figure 76, where each node in a graph represents a time series in the comparison set and each edge the DTW distance between the nodes. Nodes are numbered based on their listing in the set with the last node (the circle with the largest number) being the latent trace. Lighter colors for both nodes and edges denote smaller distances. To produce the graphs we used the Python package NetworkX (Hagberg et al., 2008.) with the fruchterman_reingold layout[22] and the Viridis colormap[23]. Looking at Figure 76, the reader should note that in all cases, the node with the highest number (the latent trace) appears with the brightest yellow color and at the most central location of the graph image – visually indicating that the latent trace was the "centroid" of each group.

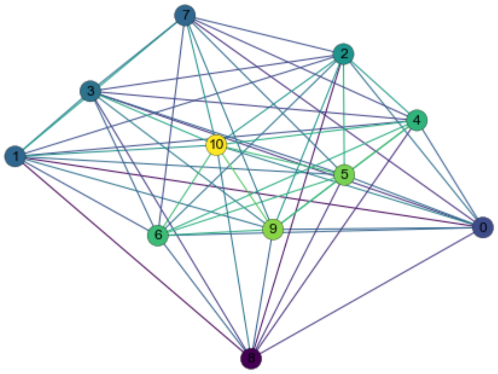
[22] Since the software default layout locates the nodes with the highest degree in the center, the input for the algorithm was the DTW distance matrix for a comparison set where each of the DTW distances is replaced with its absolute difference with the max distance. Thus the node in the center is the centroid with smallest total DTW distance to its neighbors.

[23] Created by Stéfan van der Walt and Nathaniel Smith. This color map is designed in such a way that it will analytically be perfectly perceptually-uniform, both in regular form and also when converted to black-and-white. It is also designed to be perceived by readers with the most common form of color blindness (Garnier et al., 2015).

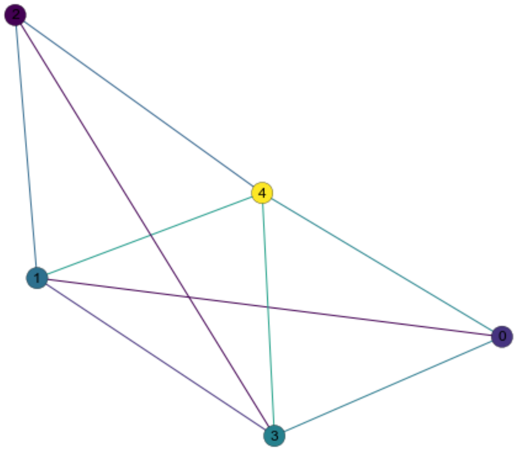
YesNo_A



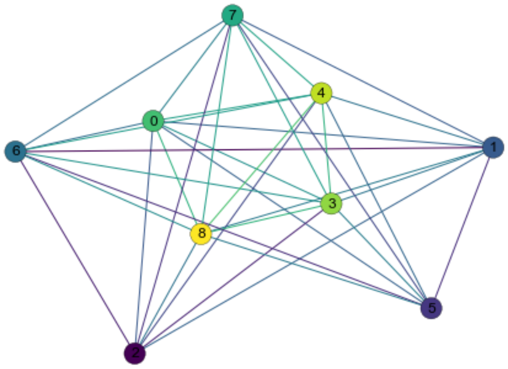
YesNo_B



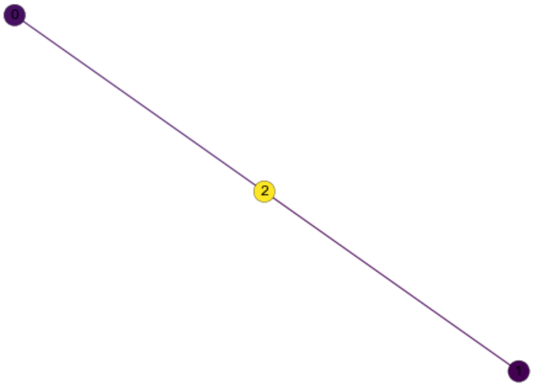
WhQuestion_A



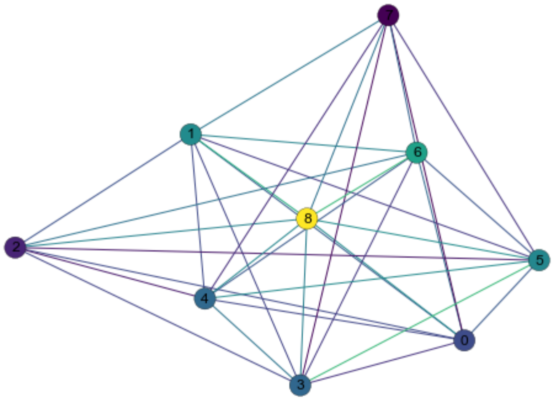
WhQuestion_B



Rhetorical_A



Rhetorical_B



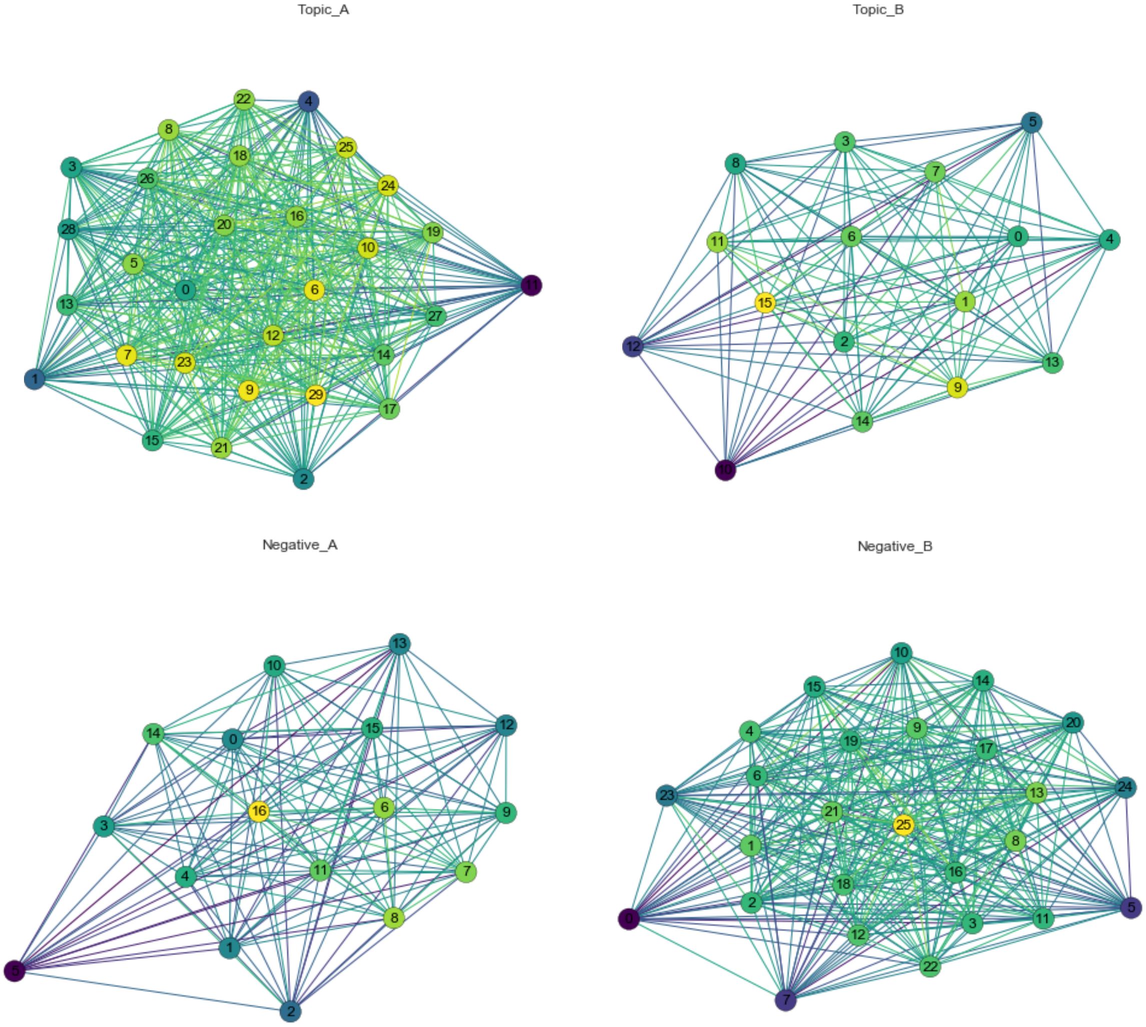


Figure 76: Visualization of DTW distances and the centroid in the comparison set for each of the ASL facial expression subcategories. The centroid is represented as the node with the smallest sum distance; hence the lightest color in the graph. For all graphs, the latent trace was found to be the centroid corresponding to the following node numbers: YesNo_A – 9, YesNo_B – 10, WhQuestion_A – 4, WhQuestion_B – 8, Rhetorical_A – 2, Rhetorical_B – 8, Topic_A – 29, Topic_B – 15, Negative_A – 16, and Negative_B – 25.

Chapter 13 Evaluation of the Data-Driven ASL Facial Expression Synthesis with Continuous Profile Models

This chapter will present two forms of evaluation of the CPM latent trace model for ASL facial expression synthesis. In Section 13.1, the CPM model will be compared to a "gold standard" (fluent human ASL performance) using a distance-metric-based evaluation, and in Section 13.2, the results of a user-study will be presented, in which ASL signers evaluated animations of ASL based upon the CPM model.

To provide a basis of comparison, in this chapter, we evaluate the CPM approach in comparison to an alternative approach that we call 'Centroid', which is inspired by previous work adopting cluster medoid on lexical facial expression for German Sign Language (Section 9.5). Similar to Section 12.3, for the Centroid approach we use multivariate DTW to select one of the time series in the training set as a representative performance of the facial expression. (Note that in Section 12.3, the DTW distance analysis was performed on a dataset that included the union of all of the training examples and the latent trace; in this chapter, the "Centroid" used as a basis for comparison is selected by a DTW analysis of the set of training examples only.)

As mentioned above, in Section 13.1, we will compare the output of the CPM model to a "gold standard" performance of each sub-category of ASL facial expression. We have selected a set of video recordings of a male ASL signer from our initial pilot dataset described in Section 10.2 to serve as the "gold standard" fluent human performance. These ASL video recordings were previously "unseen" during the creation of the CPM model, that is, they were not used in the training data set during the creation of the CPM model.

Table 13 shows the names of the videos that were selected as centroids (Appendix C) and the codenames of the gold standard performance (Appendix A). The length (bold numbers in Table 13) is the number of video frames; the training set was performed by a female signer and the gold standard was performed by a male signer, as described in Chapter 10.

Table 13: CPM latent trace, centroid, and gold standard for each ASL facial expression subcategory.

Subcategory	***Latent Trace*** *#Frames$_{double\ resolution}$*	***Training Set Centroid*** *#Frames (Video codename)*	***Gold Standard*** *#Frames (Codename)*
YesNo_A	**105**	**45** (2011-12-01_0037-cam2-05)	**45** (Y4)
YesNo_B	**160**	**56** (2011-12-01_0037-cam2-09)	**73** (Y3)
WhQuestion_A	**50**	**21** (2011-12-01_0038-cam2-05)	**35** (W1)
WhQuestion_B	**84**	**33** (2011-12-01_0038-cam2-07)	**55** (W2)
Rhetorical_A	**33**	**13** (2011-12-01_0041-cam2-04)	**30** (R3)
Rhetorical_B	**113**	**49** (2011-12-01_0041-cam2-02)	**101** (R9)
Topic_A	**60**	**19** (2012-01-27_0050-cam2-05)	**24** (T4)
Topic_B	**93**	**45** (2012-01-27_0051-cam2-09)	**55** (T3)
Negative_A	**138**	**20** (2012-01-27_0051-cam2-03)	**67** (N2)
Negative_B	**156**	**38** (2012-01-27_0051-cam2-30)	**33** (N5)

13.1 Metric Evaluation

For each of the gold standard videos, we have both the video recordings of a human signer performing the utterances, and the EBMRscripts corresponding to the animated hand movements for each stimulus. Thus, the extracted facial expressions from the human recording can serve as a gold standard for how the animated character's face and head should move, since they share the same glosses and thus similar hand movements. In this section, we compare: (a) the distance of the CPM latent trace from the gold standard to (b) the distance of the centroid form the gold standard.

As we observed in Table 13 there is a high variability in the length of all: latent trace, centroid, and gold standard videos. For a fairer comparison, we first resample these time series, using cubic interpolation, to match the duration (in milliseconds) of the animation they would be applied to and then we use multivariate DTW to estimate their distance as described in Section 12.3. In prior work (Kacorri and Huenerfauth, 2015b) we have shown that a scoring algorithm based on DTW had some moderate (though significant) correlation with comprehension and notice scores that participants assigned to ASL animation with facial expressions.

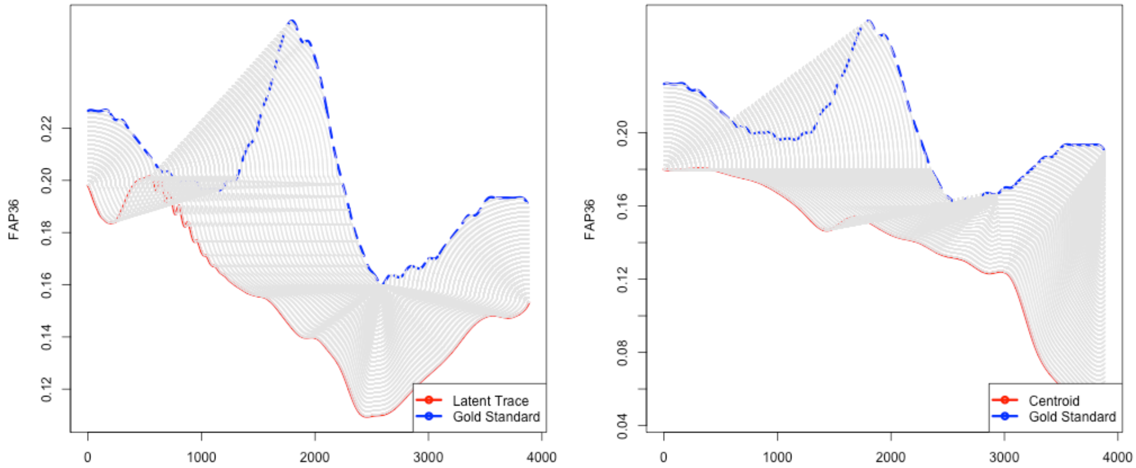


Figure 77: DTW distances on the FAP36 feature, squeeze_l_eyebrow, during a Negative_A ASL facial expression: (left) between the CPM latent trace and the gold standard and (right) between the centroid and the gold standard. The timeline for both graphs is given in milliseconds.

Figure 77 illustrates an example with the similarity of the latent trace and the centroid to the gold standard for left eyebrow squeezing during a Negative_A ASL facial expression. Given that the centroid and the training data for the latent trace are driven by recordings of a (female) signer other than the (male) signer in the gold standard, there are inherent differences between these facial expressions due to idiosyncratic aspects of individual signers. Thus this metric evaluation is especially challenging because it is an inter-signer evaluation.

13.1.1 Results

Figure 78 illustrates the overall calculated DTW distances, including a graph with the results broken down per subcategory of ASL facial expression. The results indicate that the CPM latent trace is closer to the gold standard than the centroid is. Note that the distance values are not zero since the latent trace and the centroid are being compared to a recording from a different signer on novel, previously unseen, ASL sentences. The results in these graphs suggest that the latent trace model out-performed the centroid approach; thus, our research question **RQ9** is supported.

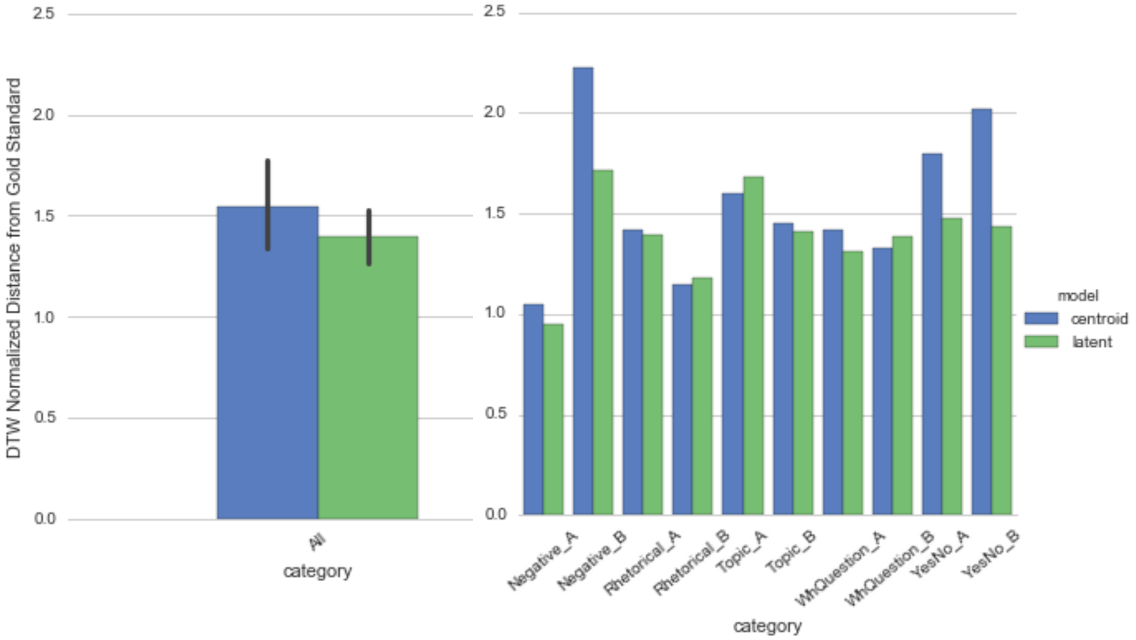


Figure 78: Grouped barplots of overall normalized DTW distances for latent trace and centroid (left) and per each subcategory of ASL facial expression (right).

13.2 User Evaluation

To assess our ASL data driven synthesis approach, we conduct a user study where ASL signers watch short ASL sentences of three types of stimuli: a) animations with a static neutral face (as a lower baseline), b) animations with facial expressions driven by the centroid, and c) animations with facial expressions driven by the CPM latent trace. Figure 79 illustrates screenshots of each stimulus type for YesNo_A facial expressions.

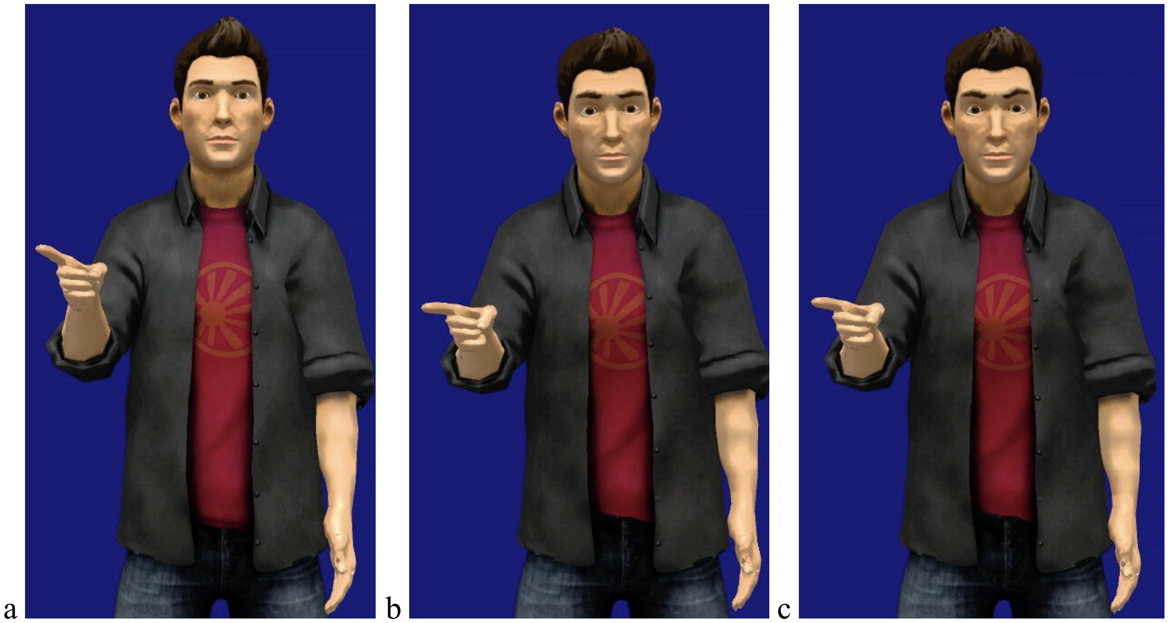


Figure 79: Screenshots of YesNo_A stimuli of three types: a) neutral, b) centroid, and c) latent trace.

In Part I of this thesis, we investigated key methodological considerations in conducting a study to measure comprehension of sign language animations with deaf users, including the use of appropriate baselines for comparison, the appropriate method for presenting comprehension questions and instructions, demographic and technology experience factors influencing acceptance of signing avatars, and other factors that we have considered in the design of this current study.

During our study, after participants answered the demographic and technology-experience questions that were established as screening criteria in Chapter 7, they viewed a sample animation, to become familiar with the experiment setup and the questions they would be asked about each animation. (This sample animation used a different stimulus than the other ten animations shown in the study.) Next, after viewing each of the ten main animations, an onscreen video of an ASL signer asked participants four fact-based comprehension questions about the information conveyed in the animation. Participants responded to each question on a 7-point scale from "definitely no" to "definitely yes." As described in Section 7.2, a single "Comprehension" score for each animation can be calculated by averaging the scores of the four questions.

Next, the participants were asked to respond to a set of questions that measured their subjective impression of the animation, using a 1-to-10 scalar response. Each question was conveyed using ASL through an onscreen video, and the following English question text was shown on the questionnaire:

(a) Good ASL grammar? (10=Perfect, 1=Bad)

(b) Easy to understand? (10=Clear, 1=Confusing)

(c) Natural? (10=Moves like person, 1=Like robot)

Questions (a-c) have been used in many of our prior experimental studies, e.g. Part I of this thesis, and were included in the collection of standard stimuli and questions that we released to the research community (Chapter 4 and Appendix A). To calculate a single "Subjective" score for each animation, the scalar response scores for the three questions were averaged.

13.2.1 Animation Stimuli

All three types of stimuli (neutral, centroid and latent trace), share identical EMBR scripts specifying the hand and arm movements; these scripts were hand-crafted by ASL signers in a pose-by-pose manner. For the neutral animations, we do not specify any torso, head, nor face movements; rather, we leave them in their neutral pose throughout the sentences. As for the centroid and latent trace animations, we apply the head and face movements (as specified by the centroid model or by the latent trace model) only to the portion of the animation where the facial expression of interest occurs, leaving the head and face for the rest of the animation to a neutral pose. For instance, during a stimulus that contains a Wh-question, the face and head are animated only during the Wh-question, but they are left in a neutral pose for the rest of the stimulus (which may include other sentences).

In order to identify face animation poses from the centroid model or from the latent trace model, we resample these time series, using cubic interpolation, to match the duration (in milliseconds) of the animation they would be applied to. While, in earlier work, we explored alternative time adjustment techniques such as piece-wise time warping based on defined linguistic boundaries using timeline milestones (Kacorri and Huenerfauth, 2015b) and Cardinal Splines (Catmull and Rom, 1974), we found that resampling with cubic interpolation yields similar animation results, with less complexity.

To convert the centroid and latent trace time series into EMBR script, we used the MPEG4 features-to-animation pipeline described in Section 11.2.2. Specifically, the generated EMBR script for these facial expressions was an EMBR PoseSequence with a pose defined every 133 milliseconds. Prior to selecting this 133-msec time interval for specifying face and head poses, we also experimented with alternatives such as using 33-msec intervals, 66-msec intervals, or by sampling face and head poses in a manner that was time aligned with the key frames (key poses) of hand movements; however, we found that the 133-msec interval produced smoother animations with the EMBR animation system. For example, Figure 80 illustrates how face and head poses defined at a 133-msec rate still maintained the overall shape of the curves for these movements.

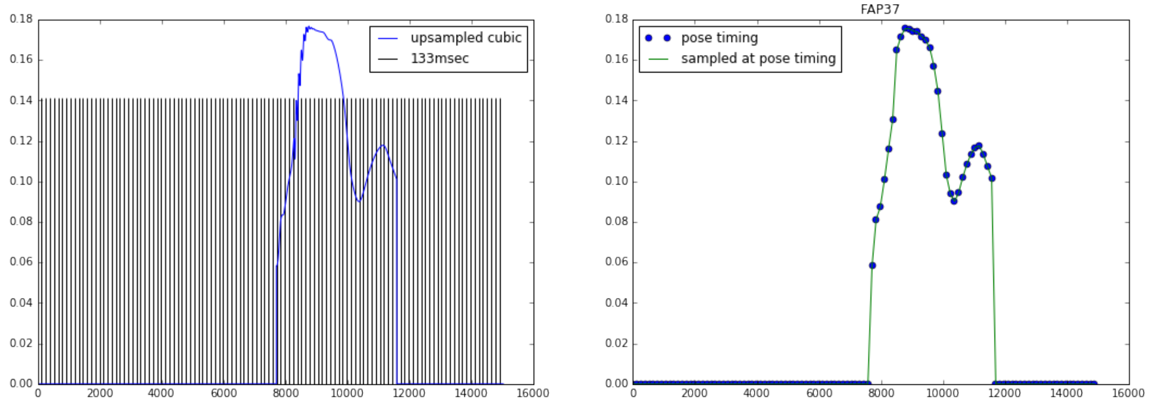


Figure 80: Before and after pose sampling of the FAP37 feature, squeeze_r_eyebrow, during a Negative_A ASL facial expression driven by the latent trace.

13.2.2 Results

In this study, a Deaf researcher and a Deaf undergraduate students (all ASL signers) recruited and collected data from participants, during meetings conducted in ASL. Initial advertisements were sent to local email distribution lists and Facebook groups. The advertisement included contact information for a Deaf researcher, including an email address, videophone, and text messaging (mobile phone).

The difference in appearance between our animation stimuli was subtle: The only portion of the animations that differed between the three conditions (neutral, centroid, and latent-trace) was the face and the head movements during the span of time when the syntactic facial expression should occur (e.g., during the Wh-question). Given this subtlety, we used the demographics in Chapter 7 to identify participants in the study who would be 'harsher critics.' Specifically, we screened for participants that identified themselves as "deaf/Deaf" or "hard-of-hearing," who had grown up using ASL at home or had attended an ASL-based school as a young child, such as a residential or daytime school.

A total of 17 participants met the above criteria, where 14 participants self-identified as deaf/Deaf and 3 as hard-of-hearing. 10 participants had attended a residential school, and 7, a daytime school for deaf students. 14 participants had learned ASL prior to age 5, and 3 had been using ASL for over 7 years. There were 8 men and 9 women of ages 19-29 (average age 22.8). Participant's responses on the MediaSharing scale varied between 3 and 6 (mean score of 4.3), and on AnimationAttitude scale, they varied from 2 to 6 (mean score of 3.8). We report these technology experience factors since they were found to be important in explaining variance in comprehension and subjective scores of sign language animations (Chapter 7).

In a between-subjects design, we compared the comprehension and subjective scores assigned to three types of animations with identical hand movements but differing in their face, head, and torso movements: (a) static and '*neutral'* throughout the story, (b) driven by a recording of a human performing a *'centroid'* of that type of facial expression, and (c) driven by the *'latent trace'* from multiple recordings of a human performing that type of facial expression.

Figure 81 and Figure 82 show distributions of comprehension question scores and subjective scores, respectively, as boxplots with a 1.5 interquartile range (IQR). For comparison, means are denoted with a star and their values are labeled above each boxplot.

We found no significant difference (ANOVA) when comparing the comprehension scores between centroid, latent trace, and neutral animation in Figure 81. We speculate that these highly skilled ASL signers recruited for this study may have been able to infer the meaning of the animations regardless of lack of, or minor errors in, the facial movements.

However, when comparing the subjective scores that participants assigned to the animations in Figure 82, we found a significant difference (Kruskal-Wallis test used since the data was not normally distributed) between the latent trace and centroid $(p < 0.005)$ and between the latent trace and neutral $(p < 0.05)$; thus, our research question **RQ10** is supported.

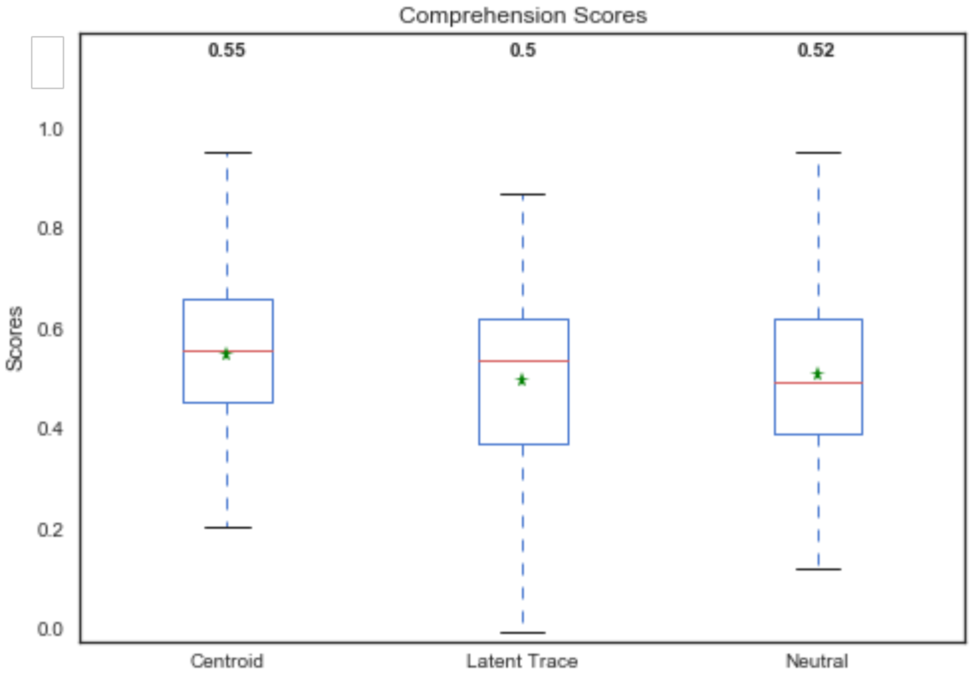


Figure 81: Comprehension scores for animations with facial expressions driven by centroid, latent trace, and with

a neutral head/face pose.

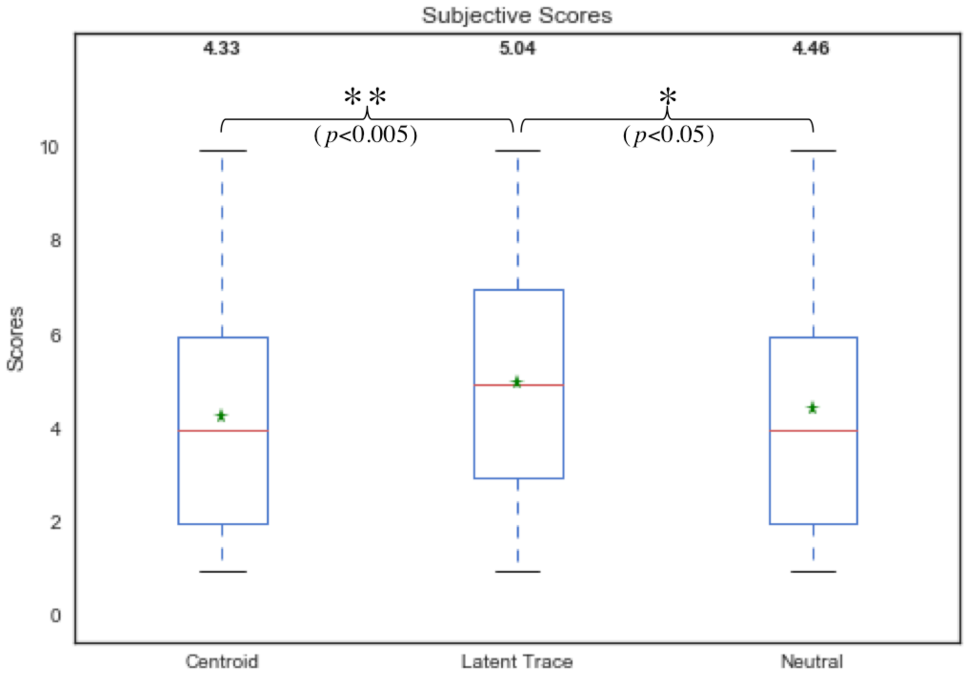


Figure 82: Subjective scores for animations with facial expressions driven by centroid, latent trace, and with a neutral head/face pose.

Epilogue to Part II

Part II in this thesis described our work on syntactic facial expression synthesis for ASL animations based on facial-feature data from multiple recordings of ASL signers.

We have collected and shared with the research community a set of stimuli videos that have been recorded and annotated in our lab as initial pilot data. In partnership with collaborators, we are producing MPEG-4 facial feature data from a video corpus of ASL with linguistically diverse occurrences of syntactic facial expressions.

To support the generation of ASL animations with facial expressions, we have enhanced a virtual human character from the open source animation platform EMBR with face controls following the MPEG-4 Facial Animation standard and ASL handshapes. In a user-study, we determined that these controls were sufficient for conveying understandable animations of ASL facial expressions and we compared it to our previous animation platform.

To synthesize a syntactic ASL facial expression, we use probabilistic generative models such as Continuous Profile Model that can automatically uncover the underlying trace from multiple recordings of the same type of facial expression. We have trained our models for each of the granular categorization of the linguistically diverse dataset and we have demonstrated that they are identifying a curve that seems to be a good representative of the training data set.

Lastly, to further assess our ASL facial expression modeling approach we have conducted both a metric evaluation and a user study where signers evaluate animations with facial expressions produced by our enhanced EMBR platform with MPEG-4 facial controls. Our synthesis approach is compared against Centroid, an alternative approach for choosing a representative performance of ASL facial expressions. Our results suggested that the Continuous Profile Model outperforms Centroid and it is a promising approach for the synthesis of ASL facial expressions.

Part II of this thesis has addressed the following research question:

RQ7: *Is our MPEG4-enhanced animation platform sufficiently expressive such that it could produce facial expressions for ASL animations that are understandable and explicitly recognized by ASL signers?* (This was supported by the results presented in Chapter 11.)

RQ8: *Is a Continuous Profile Model (CPM) able to produce a latent-trace curve that is representative of a set of ASL facial expressions, which had been provided as training data to the model?* (This was supported by results presented in Chapter 12)

RQ9: *Can a Continuous Profile Model (CPM), which finds the underlying latent trace of extracted facial-feature data from multiple human recordings, identify feature curves that are more similar to human performances of novel sentences?* (This was supported by the results presented in Section 13.1.)

RQ10: *Can a Continuous Profile Model (CPM), which finds the underlying latent trace of extracted facial-feature data from multiple human recordings, produce high-quality facial expressions for ASL animations, as judged by ASL signers in an experimental study?* (This was supported by the results presented in Section 13.2)

Chapter 14 Conclusions and Contributions

This thesis describes ASL syntactic facial expression synthesis that is driven by extracted facial data from recordings of ASL signers. Its research focus is divided into two "Parts." Part I investigates methodological aspects when evaluating facial expression synthesis in a user study. Part II examines data-driven modeling and generation of novel ASL sentences with syntactic facial expressions.

While the production of linguistically meaningful facial expressions in sign language animations is crucial for the interpretation of signed sentences, little attention in prior work has been focused on their automatic generation and evaluation methodology. To assess their animations, prior researchers have typically conducted user studies with signers. However, there has been a lack of methodological rigor and standardization in the conduct of these evaluations, which must overcome the challenges specifically tied to facial expression assessment. Another limitation on this prior work has been the lack of linguistically diverse corpora enriched with extracted facial features; such data resources are necessary to support data-driven modeling of syntactic facial expressions and their synchronization to the manual movements. Therefore, the goal of this dissertation research was two-fold:

- Conduct rigorous methodological research in Part I on how experiment design affects study outcomes when evaluating facial expressions in sign language animations; with ASL syntactic facial expressions being the case study.

- Demonstrate in Part II that an annotated sign language corpus can be used to model and generate linguistically meaningful facial expressions, if it is combined with facial feature extraction techniques, statistical machine learning, and an animation platform with detailed facial parameterization.

To achieve those two research goals, we first engineered and shared with the research community a collection of stimuli and comprehension questions that contain syntactic ASL facial expressions and can measure whether ASL signers understand the intended facial expression effectively. We further used these stimuli in user studies that investigated the effect of videos as upper baseline and for presenting comprehension questions, explored eye-tracking as an alternative to recording question-responses from participants, and identified demographic and technology-experience variables predictive of participants' responses. Although no effects were identified in participants' response scores when videos vs. animations used for presenting comprehension questions, there were various effects on participants' responses when the upper baseline in the study was a video vs. an animation. Thus, the choice of upper baseline must be considered when comparing results across studies. Further, the potential effect of the choice of upper-baseline must be considered when a researcher is designing a new study. Another finding in Part I was in regard to eye-tracking: In our user study that investigated eye-tracking as an alternative evaluation approach, we found that deaf and hard-of-hearing signers' eye movements correlate with their subjective judgments about ASL videos or animations with facial expressions. Last, we found a set of demographic and technology experience/attitude factors important in explaining variance in comprehension and subjective scores of sign language animations and thus useful for researchers to consider when reporting or balancing characteristics of their participants in user studies.

To achieve our second research goal in Part II, we collected extracted facial movement data from human recordings of syntactic facial expressions from linguistically diverse sentences. While there is still no consensus among animation researchers as how to best symbolically represent facial expressions in sign language, we have adopted the MPEG-4 Facial Animation standard for the representation of the extracted facial movements. We have extended the EMBR animation platform with facial controls that support the same standard, thus allowing recordings from multiple human signers to drive the facial expressions of the avatar. In a user study, we found that both our animation platform and the adopted MPEG-4 parameterization are sufficient to generate ASL animations with syntactic facial expressions. When directly compared to our prior platform with finite-repertoire facial expressions, the extended EMBR platform achieved better understandability and expressiveness of animations with facial expressions. To avoid idiosyncratic aspects of a single performance, we have modeled a facial expression based on the underlying trace of the movement trained on multiple recordings of different sentences where this type of facial expression occurs. We obtain the latent trace with Continuous Profile Model (CPM), a probabilistic generative model that relies on Hidden Markov Models. We assessed our modeling approach through comparison to an alternative centroid approach, where a single performance was selected as a representative. Through both a metric evaluation and an experimental user study, we found that the facial expressions driven by our CPM models produce high-quality facial expressions that are more similar to human performance of novel sentences.

To summarize, we list our research questions again below. Each of these research questions have been examined (or partially examined) in this thesis:

RQ1: *Can our stimuli and comprehension questions that contain linguistic facial expressions measure whether participants understand the indented facial expression effectively?* (**RQ1** was supported by a user study in Chapter 4 and further supported by another user study in Chapter 11.)

RQ2: *How does the modality (video of a human vs. a human-produced high-quality animation) of an upper baseline, presented for comparison purposes, affect the comprehension and subjective scores for the animation being evaluated?* (**RQ2** was examined in Chapter 5, and various effects were identified.)

RQ3: *Does the modality (video of a human vs. a human-produced high-quality animation) of instructions/comprehension questions in a study affect the comprehension and subjective scores for the animation being evaluated?* (**RQ3** was also examined in Chapter 5, and no effects were identified.)

RQ4: *Could eye-tracking be effectively used as a complementary or an alternative unobtrusive way of evaluating sign language animations with facial expressions?* (**RQ4** was supported by a user study in Chapter 6.)

RQ5: *Which are the eye-tracking metrics that correlate with evaluation judgments from signers during the evaluation of sign language animations with facial expressions?* (**RQ5** was examined in Chapter 6, and several metrics were identified.)

RQ6: *What demographic and technology-experience variables are predictive of signers' judgments during evaluation of sign language animations?* (Chapter 7 examined this question: several factors were identified that can influence participants' judgments of the quality of ASL animations.)

RQ7: *Is our MPEG4-enhanced animation platform sufficiently expressive such that it could produce facial expressions for ASL animations that are understandable and explicitly recognized by ASL signers?* (**RQ6** was supported by a user study in Chapter 11.)

RQ8: *Is a Continuous Profile Model (CPM) able to produce a latent-trace curve that is representative of a set of ASL facial expressions, which had been provided as training data to the model?* (This was supported by the results presented in Chapter 12.)

RQ9: *Can a Continuous Profile Model (CPM), which finds the underlying latent trace of extracted facial-feature data from multiple human recordings, identify feature curves that are more similar to human performances of novel sentences?* (This was supported by the results presented in Section 13.1.)

RQ10: *Can a Continuous Profile Model (CPM), which finds the underlying latent trace of extracted facial-feature data from multiple human recordings, produce high-quality facial expressions for ASL animations, as judged by ASL signers in an experimental study?* (This was supported by the results presented in Section 13.2.)

14.1 Contributions

As discussed in earlier chapters, limitations in the grammatical-correctness and naturalness of facial expressions have held back the understandability of ASL animations. This dissertation demonstrates that our data-driven models for the movements of virtual humans in ASL animations improve the automatic synthesis of previously unseen instances of these expressions within novel sentences. Given a detailed input, such as: (i) a script specifying the sequence of the ASL glosses; (ii) start- and end-timing of these glosses and (iii) the associated syntactic facial expression, the technologies described in this dissertation can produce the timing and intensity of facial and head features that correspond to the facial expression.

The state-of-the-art of animation generation software has been advanced in the following ways: (1) This technology could produce flexible facial expression "lexicons" that can be used to synthesize an infinite variety of syntactic phrases. (2) Scripting software could more easily enable users to include facial expressions in a sentence (without requiring the user to manually create a custom animation of the head and other facial components for a particular facial expression). We believe that technologies for automatically synthesizing facial expressions in ASL animations – to increase the naturalness and understandability of those animations – can ultimately lead to better accessibility of online information for people who are deaf with low English literacy.

Although the evaluation methodologies and modeling techniques in this thesis focused on five ASL syntactic facial expressions, they should also be applicable to more ASL syntactic facial expressions. This could be done by collecting sample recordings with instances of these facial expressions and training new models for facial expressions, including for those found in other sign languages used internationally.

The stimuli and comprehension questions engineered in this thesis have been made available to the research community. Ultimately, the field of sign language animation synthesis may benefit from the community identifying a standard set of evaluation stimuli and questions for system evaluation, to better enable comparison of systems and progress in the field.

Our user studies that investigated the effect of video as upper baseline or for presenting comprehension question provided methodological guidance for future researchers who are conducting studies with sign language and facilitated fair comparisons of the results of sign language animation evaluation studies, in which the authors have made different methodological choices. Enabling such comparison of results across published studies should benefit the field by enabling researchers to more accurately identify the most effective animation-synthesis techniques.

Our methodological research on eye-tracking can contribute to future sign language animation researchers designing a user-based evaluation study: (1) They can consider eye tracking as an alternative or complimentary form of measurement in their study. (2) They can use the results presented in this thesis for comparison purposes to understand how to characterize eye metrics they obtain in their studies. (3) They can further investigate other eye-tracking metrics that might better capture signers' eye-gazing for a different sign language.

Our study on demographic and experiential factors influencing acceptance of sign language animation by deaf users provided a deeper understanding of the relationship between participant characteristics and evaluation scores in this field and a concise set of questions that may be useful for researchers who are interested in minimizing the amount of study time spent collecting demographic and technology experience/attitude data. Through collection and publishing of these characteristics of study participants, we anticipate easier comparisons of research results across publications. We also believe that these factors would be useful for researchers to consider if they are balancing or matching participants across treatment conditions in a study.

14.2 Future Work

Part I: The maturing field of sign language animation synthesis can benefit from additional methodological research – especially work that facilitates comparisons across evaluation studies of different animation systems or leads to further consensus in evaluation techniques – and this will ultimately benefit the users of this technology.

Our methodological work in Part I could be extended in the following ways:

- In our eye-tracking study (Chapter 6) we examined one-to-one correlation relationships between the evaluation scores that participants assigned to the stimuli (video and animations) and specific eyetracking metrics. In future work we will focus on sign language animations only, and we will systematically investigate the contribution of *multiple metrics* in indicating the subjective responses that deaf and hard-of-hearing signers assign to ASL animations via multiple regression modeling.
- For our demographic study (Chapter 7) we are also interested in further exploring the variable of Age. This variable was not selected by the exhaustive all-subsets model comparison in this study, but only 10% of our 62 participants were over age 43. In

future work, we would like to conduct additional targeted recruitment of older participants. As we have learned in this study, such participants were the most time-consuming to recruit; so, this must be factored into the data-collection timeline in future work.

- While we believe that studies with ASL signers are the most conclusive way to evaluate the understandability and naturalness of animations of ASL, having a rapid, repeatable method of evaluating the output of facial expression synthesis software is useful for monitoring the development of software, and this evaluation can be performed more frequently than user-based evaluations. In a first attempt at developing an automatic scoring approach (Kacorri and Huenerfauth, 2015b) we have observed some moderate though significant correlations between our scoring approach and the perceived quality of animations by signers. In future work we plan on investigating further variations of scoring techniques that might prove even more effective. For example, instead of Dynamic Time Warping, we may investigate other probabilistic approaches to similarity – and compare them to our findings.

Part II: We explored the Continuous Profile Model in its simplest form, without extensive preprocessing of the data or tuning, based on multiple recordings of a single signer and we found that the results suggest that this is a promising approach for the synthesis of ASL facial expressions. Our facial expression modeling work in Part II could be extended in the following ways:

- To aid CPM convergence to a good local optimum, in future work we will investigate dimensionality reduction approaches that are reversible such as Principal Component Analysis (PCA, Pearson 1901) and other pre-processing approaches similar to (Listgarten, 2007), where the training data set is coarsely pre-aligned and pre-scaled based on the center of mass of the time series. In addition we will fine-tune some of the hyper parameters of the CPM such as spline scaling, single global scaling factor, convergence tolerance, and initialization of the latent trace with a centroid.
- Our linguistically diverse dataset included recordings from a single signer (female) while collaborators at Boston and Rutgers universities are continuing on annotating recordings from two additional signers (male). Given that the MPEG-4 standard allow us to extract signer-independent face and head features, in future work we would like to obtain CPM latent traces on training data from multiple signers for more generalizable models.
- Last, we will explore alternatives for enhancing Continuous Profile Models by incorporating contextual features in the training data set such as timing of hand movements, and preceding, succeeding, and co-occurring facial expressions.

14.3 Summary

In summary, it is our goal to make substantial improvements in ASL animation technologies, which have accessibility benefits for people who are deaf. While high quality broad-coverage generation software for ASL animation is beyond the scope of a single Ph.D. dissertation, we hope that our research will provide motivation for future computational linguistic work on

ASL. We see the methodological research, the sharing of evaluation stimuli and questions, and the dissemination of evaluations of new animation synthesis techniques as laying an important groundwork for future researchers who wish to advance the state-of-the-art of sign language animation technologies. Further, because our animation work is based on MPEG-4 standard, our techniques and models may also have important applications for non-sign language researchers studying modeling, synthesis, and efficient parameterizations of facial expression animations.

APPENDIX A ASL Facial Expression Stimuli Collection

This appendix includes detailed information on the stimuli collection that we shared with the research community in (Huenerfauth and Kacorri, 2014), as described in Chapter 4. Table 14 provides the codenames for each of passages (a total of 48), their ASL transcriptions, and the English translations of two possible interpretations depending on the perceived facial expression. Table 15 includes the comprehension questions for each of the passages and the suggested correct answers if the facial expressions are perceived. There are a total of 192 comprehension questions/answers pairs (4 questions/answers for each of the passages). Screenshots for both the comprehension question and subjective/"notice" question HTML forms included in the collection are illustrated in Figure 83(a) and Figure 83(b), respectively.

Table 14: ASL passages in our stimuli collection including codenames, ASL glosses, and English translation of two possible interpretations based on the facial expression being performed.

ID#	ASL GLOSSES	English Translation (for ASL with facial expression)	English Translation (for ASL without facial expression)	Type	Sub-Type
W1	LAST FRIDAY NIGHT, YOU WATCH #METALLICA WHERE. MUSIC SHOW YOUR FAVORITE.	Where did you saw Metallica last Friday night? Music show is your favorite.	Last Friday night, you saw Metallica. Where was your favorite music show?	question	whq-thing
W2	MY SISTER NAME #MARY. YOUR BOSS WHO. MANAGER YOU TWO_OF_YOU FINISH MEET.	My sister is Mary. Who is your boss? You met the manager.	My sister Mary is your boss. Who was the manager you met?	question	whq-person
W3	MY BROTHER NAME #BILL. YOUR DENTIST WHO. DOCTOR YOU TWO_OF_YOU FINISH MEET.	My brother is Bill. Who is your dentist? You met the doctor.	My brother Bill is your dentist. Who was the doctor you met?	question	whq-person
W4	COMPUTER YOU BOUGHT WHERE. #SALLY FAVORITE SHOPPING CENTER.	Where did you bought the computer? Sally's favorite is the shopping center.	You bought a computer. Where is Sally's favorite shopping center?	question	whq-thing
W5	THAT #MARY HER BIRTHDAY PARTY WHEN. #MARY DRUNK	When is Mary's birthday party? Mary is drunk.	It is Mary's birthday party. When did Mary got drunk?	question	whq-thing
W6	BRIDGE YOU FINISH TOUCH WHAT. MUSEUM YOU PREFER/FAVORITE.	What was the bridge you visit? You prefer museum.	You visited a bridge. What kind of museum do you prefer?	question	whq-thing

ID#	ASL GLOSSES	English Translation (for ASL with facial expression)	English Translation (for ASL without facial expression)	Type	Sub-Type
W7	LAST WEDNESDAY NIGHT, YOU WATCH SCARY MOVIE WHAT KIND MOVIE YOU PREFER	Last Wednesday night, you saw a scary movie. What is your favorite kind of movie?	Last Wednesday, which scary movie did you see? That's your favorite kind of movie.	question	wh-q-thing
W8	YOUR UNCLE POINT WOW RICH. NEXT-WEEK, YOUR NEPHEW HIS BIRTHDAY. HE(UNCLE) WILL GIVE FANCY GIFTS. LIST-1-of-2 MAYBE BIKE OR LIST-2-of-2 MAYBE CAR. YOUR NEPHEW OLD++ 10	Your uncle is rich. Next week, it's your nephew's birthday. Your uncle gives fancy presents: perhaps a bike or a car. How old is your nephew? 10-years old?	Your uncle is rich. Next week, it's your nephew's birthday. Your uncle gives fancy presents: perhaps a bike or a car. Your nephew is 10-years old.	question	wh-q-person
W9	PAST, YOU SAY YOUR SCHOOL TRIP MAYBE GO LIST-1-of-2 MOUNTAIN-CLIMBING OR LIST-2-of-2 FANCY RESTAURANT. TOMORROW, YOUR SCHOOL TRIP YOU WEAR HEELS	In the past, you said that your school trip will maybe go mountain climbing or go to a fancy restaurant. Tomorrow, it's your school trip. What will you wear? Heels?	In the past, you said that your school trip will maybe go mountain climbing or go to a fancy restaurant. Tomorrow, it's your school trip. You will wear heels.	question	wh-q-thing
Y1	#BOB'S #DINER THAT YOUR SISTER HER FAVORITE RESTAURANT. ALL FOOD CHEAP (point)	Your sister's favorite restaurant is Bob's Diner. All the food is cheap?	Your sister's favorite restaurant is Bob's Diner. All the food is cheap.	question	ynq-thing
Y2	YESTERDAY, ME FINISH MEET #BOB. HE TEACHER. HIM(point) YOUR SISTER LIKE	Yesterday, I met Bob. He's a teacher. Your sister likes him?	Yesterday, I met Bob. He's a teacher. Your sister likes him.	question	ynq-person
Y3	NEXT YEAR, YOUR SISTER ME VISIT WILL (shake head "yes"). LIVE WASHINGTON_DC SHE(point)	Next year, I will visit your sister. She lives in Washington?	Next year, I will visit your sister. She lives in Washington.	question	ynq-person
Y4	NEW RESTAURANT YOU SUGGEST ME GO (shake head "yes") WILL. THAT RESTAURANT CHINESE (point)	I will go to the new restaurant you suggested. It is Chinese?	I will go to the new restaurant you suggested. It is Chinese.	question	ynq-thing
Y5	HIGHWAY ME DRIVE SPEED ME	I drive too fast on the highway?	I drive too fast on the highway.	question	ynq-person
Y6	FISH RESAURANT (point) PHILADELPHIA ME GO (shake head "yes") WILL. NICE	I will go to fish restaurant in Philadelphia. It is nice?	I will go to the fish restaurant in Philadelphia. It is nice.	question	ynq-thing
Y7	ME WILL GO YOUR GRADUATION NEXT YEAR YES YOU LIVE ROCHESTER?	I will attend your graduation next year. Do you live in Rochester?	I will attend your graduation next year. You live in Rochester.	question	ynq-person
R1	TODAY ALEX SICK WHY. JACKET NOT WEAR	(Rh-q) Why is Alex sick today? Because he don't wear a jacket.	(Wh-q) Alex is sick today. Why don't he wear a jacket?	question	rhq-why
R2	BEST MOVIE I MISS WHY. SOPHIE LATE	(Rh-q) Why did I miss the best movie? Sophie was late.	(Wh-q) I missed the best movie. Why was Sophie late?	question	rhq-why
R3	YESTERDAY I BOUGHT POTATOS WHY. YOU COOK FANCY DINNER	(Rh-q) Why did I bought potatos yesterday? Because you will cook a fancy dinner.	(Wh-q) I bought potatos yesterday. Why will you cook a fancy dinner?	question	rhq-why
R4	ME CHRIS TWO_OF_US BEST-FRIEND. NEXT SUNDAY I GO FISHING WHY. HE(point) LOVE SEAFOOD	(Rh-q) Me and Chris, we are best friends. I will go fishing next Sunday because he loves seafood.	(Wh-q) Me and Chris, we are best friends. I will go fishing next Sunday. Why does Chris love so much seafood?	question	rhq-why
R5	SOPHIE ALEX TWO_OF_THEM TAKE SAME CLASS WHY. TWO_OF_THEM BECOME PILOTS WANT	(Rh-q) Sophie and Alex take the same classes because they both want to become pilots.	(Wh-q) Sophie and Alex take the same classes. Why do they want to become pilots?	question	rhq-why

ID#	ASL GLOSSES	English Translation (for ASL with facial expression)	English Translation (for ASL without facial expression)	Type	Sub-Type
R6	ALEX NOW GO-GO (index finger with two hands) PARTIES WHY. FINISH DIVORCE	(Rh-q) Why did Alex started going to parties? He is divorced.	(Wh-q) Alex started going to parties. Why is he divorced?	question	rhq-why
R7	THIS YEAR ASL I LEARN HOW. I PRACTICE	(Rh-q) This year I learn ASL how? I practice.	This year I learn ASL. How do I practice?	question	rhq-how
R8	TWO-OF-US SAME CLASS. MANY STUDENTS GOSSIP, MAYBE HAVE POPQUIZ TOMORROW I KNOW+ WHO INVENT LIGHTBULB?	You and I are in the same class. Many students are gossiping that there might be a pop quiz tomorrow. I know (there will be). Who invented the light bulb?	You and I are in the same class. Many students are gossiping that there might be a pop quiz tomorrow. I know who invented the light bulb.	question	rh-q-thing
R9	YOU ENJOY DECORATE CRAFTS. I PREFER KNITTING. TODAY, ME GO STORE BUY CRAFT SUPPLIES WHY YOU DECORATE WINE BOTTLE FOR PARTY.	You enjoy decorating / crafting; I prefer knitting. Today, I'm going to the store to buy craft supplies because you are decorating a wine bottle for the party.	You enjoy decorating / crafting; I prefer knitting. Today, I'm going to the store to buy craft supplies. Why are you decorating a wine bottle for the party?	question	rq-why
R10	#MARISSA POINT NOW TRAVEL OFTEN WHY. SHE(POINT) FINISH QUIT HER JOB. SHE(POINT)DON'T-LIKE COMPUTERS.	Marissa started traveling a lot recently because she quit her job. Doesn't she like computers?	Why did Marissa start traveling a lot recently? She quit her job. Doesn't she like computers?	question	rhq-why
R11	YOU RICH. YOU ENJOY SHOPPING. I HAVE NEW BUSINESS. ME HOPE ACHIEVE MY BUSINESS HOW BUY 5 HOUSES EVERY YEAR.	You are rich and enjoy shopping. I have a new business. I hope to run a business by buying five houses a year.	You are rich and enjoy shopping. I have a new business. I hope to achieve my business (goals). How (do you) buy 5 houses every year?	question	rh-how
N1	THIS MORNING, ALEX COOKED VEGETABLES. NOW, neg:{ PASTA READY }. HE(point) HUNGRY WANT LUNCH NOW.	(Neg) This morning Alex cooked vegetables. Now the pasta are not ready. He is hungry. Now he wants to have lunch.	This morning Alex cooked vegetables. Now the pasta are ready. He is hungry. Now he wants to have lunch.	negative	
N2	ALEX TEND TAKE-UP MATH CLASS. NOW SEMESTER, SCHOOL neg{HAVE SCIENCE CLASS}. ALEX TAKE-UP TWO CLASS.	(Neg) Alex always takes math classes. This semester, the school doesn't have any science classes. Alex is taking two classes.	Alex always takes math classes. This semester, the school has science classes. Alex is taking two classes.	negative	
N3	NOW CRISTMAS VISIT SISTER neg{ME}. UNDERSTAND (sign this word twice only) VISIT GRANDMOTHER WANT ME (shake head "yes").	(Neg) This Christmas I am not visiting my sister. However I want to visit grandmother.	This Christmas I am visiting my sister. However I want to visit grandmother.	negative	
N4	YESTERDAY, MATH HOMEWORK ME FIGURE-OUT FINISH. NOW, MY ENGLISH HOMEWORK neg{FINISH}. ME HOMEWORK SUBMIT.	(Neg) Yesterday I figured out my math homework. Now, my English homework is not ready. I submitted my homework.	Yesterday I figured out my math homework. Now, my English homework is ready. I submitted my homework.	negative	

ID#	ASL GLOSSES	English Translation (for ASL with facial expression)	English Translation (for ASL without facial expression)	Type	Sub-Type
N5	LAST WEEK, MY SISTER HER BIRTHDAY. WILL WEEKEND SATURDAY, HAVE PARTY. FEW PEOPLE ME INVITE. neg{MY BIRTHDAY}.	(Neg) Last week, it was my sister's birthday. Next weekend, on Saturday, there's a party. I invited a few people. It's not my birthday.	Last week, it was my sister's birthday. Next weekend, on Saturday, there's a party. I invited a few people. It's my birthday.	negative	
N6	IF WANT GO CANADA, CAN LIST-1-of-2 RIDE BUS OR LIST-2-of-2 TRAIN. #JUSTIN POINT WANT RIDE TRAIN. YESTERDAY, HE(Justin) BUY TICKET. TODAY, TIME 11 45, HE ARRIVE TRAIN-STATION. HE(JUSTIN) BRING TICKET. NOON, TRAIN DEPART.	If someone wants to go to Canada, they can ride a bus or a train. Justin wants to ride the train. Yesterday, he bought a ticket. Today, at 11:45, he arrive at the train station. He did not bring his ticket. At noon, the train left.	If someone wants to go to Canada, they can ride a bus or a train. Justin wants to ride the train. Yesterday, he bought a ticket. Today, at 11:45, he arrive at the train station. He brought his ticket. At noon, the train left.	negative	
T1	ME SELL-SELL CARS BOATS. AIRPLANES MY BROTHER MECHANICAL SPECIALIZE	(Topic) I sell cars and boats. My brother is a mechanical expert in airplanes.	I sell cars, boats and airplanes. My brother is a mechanical expert.	topic	
T2	#ALEX(point) HIS FRIENDS BEST #BOB #MARY. #CHRIS HE(point) ANGRY	(Topic) Alex friends are Bob and Mary. Alex is very angry at Chris.	Alex friends are Bob, Mary and Chris. Alex is very angry.	topic	
T3	GO-GO RESTAURANT (two “b” hands in opposite directions) MY SISTER LIKE (shake head “yes”). HER FAVORITE ITALIAN CHINESE. FRENCH MUSIC SHE(point) LOVE	(Topic) My sister likes to go to restaurants. Her favorite are Italian and Chinese. She loves French music.	My sister likes to go to restaurants. Her favorite are Italian, Chinese and French. She loves music.	topic	
T4	NEW RESTAURANT INCLUDE PASTA PIZZA. SWEETS MY SISTER COOK EXPERT	(Topic) The new restaurant has pasta and pizza. At sweets my sister is an expert cook.	The new restaurant has pasta, pizza and sweets. My sister is an expert cook.	topic	
T5	EVERYDAY ME SCHOOL GO-GO(two index fingers). BACK-AND-FORTH ME TRAIN. BUS TAKE FOREVER.	(Topic) Everyday I go to school. I take the train. Bus takes me a really long time.	Everyday I go to school. I take the train the bus. It takes me a really long time.	topic	
T6	EVERYMORNING BREAKFAST ME EAT ORANGE. BACON DELICIOUS	(Topic) Everymorning I eat oranges for breakfast. Bacon taste really good.	Everymorning I eat orange and bacon for breakfast. It taste really good.	topic	
T7	I STUDENT MEDICAL SCHOOL EVERY DAY, I GO TO LIBRARY STUDY BOOKS PRACTICE AT HOSPITAL TOUGH.	I’m a student at medical school. Every day, I go to the library and study books. Practice at the hospital is really tough.	I’m a student at medical school. Every day, I go to the library, study books, and practice at the hospital. It is really tough.	topic	
E1	RECENT YOU STORY (point) #ALEX ME WHAT.	<IN AN ANGRY VOICE> What did you tell Alex about me?	What did you just tell Alex about me?	emotion	anger-at-viewer
E2	YESTERDAY MY SISTER CAT BRING.	<IN A SAD VOICE> Yesterday, my sister brought a cat.	Yesterday, my sister brought a cat.	emotion	sad-event

ID#	ASL GLOSSES	English Translation (for ASL with facial expression)	English Translation (for ASL without facial expression)	Type	Sub-Type
E3	YESTERDAY, I WORK. MY SISTER VISIT ME.	<IN AN ANGRY VOICE> Yesterday, my sister visited me at work.	Yesterday, my sister visited me at work.	emotion	anger-at-3rd-person
E4	TOMORROW, MY BIRTHDAY 30TH. I EXCITED.	<IN A IRONIC VOICE> Tomorrow is my 30th birthday. I am excited.	Tomorrow is my 30th birthday. I am excited.	emotion	ironic-event
E5	MY COMPUTER, YOU DO-DO.	<IN AN ANGRY VOICE> What did you do with my computer?	What did you do with my computer?	emotion	anger-at-viewer
E6	YESTERDAY, NEW PANTS SHIRT MY SISTER BUY GIVE_ME. SHIRT BLUE.	<IN A SAD VOICE> My sister bought me new pants and a new shirt. The shirt was blue.	My sister bought me new pants and a new shirt. The shirt was blue.	emotion	sad-thing
E7	LAST FRIDAY, MY BROTHER TAKE MY CAR. DRIVE SCHOOL.	<IN AN ANGRY VOICE> Last Friday, my brother drove my car to school.	Last Friday, my brother drove my car to school.	emotion	anger-at-3rd-person
E8	YESTERDAY, MY SISTER BOTH_OF_US GO_OUT MUSIC SHOW. MUSIC COUNTRY.	<IN A IRONIC VOICE> Yesterday, my sister and I went to a concert. It was country music.	Yesterday, my sister and I went to a concert. It was country music.	emotion	ironic-thing

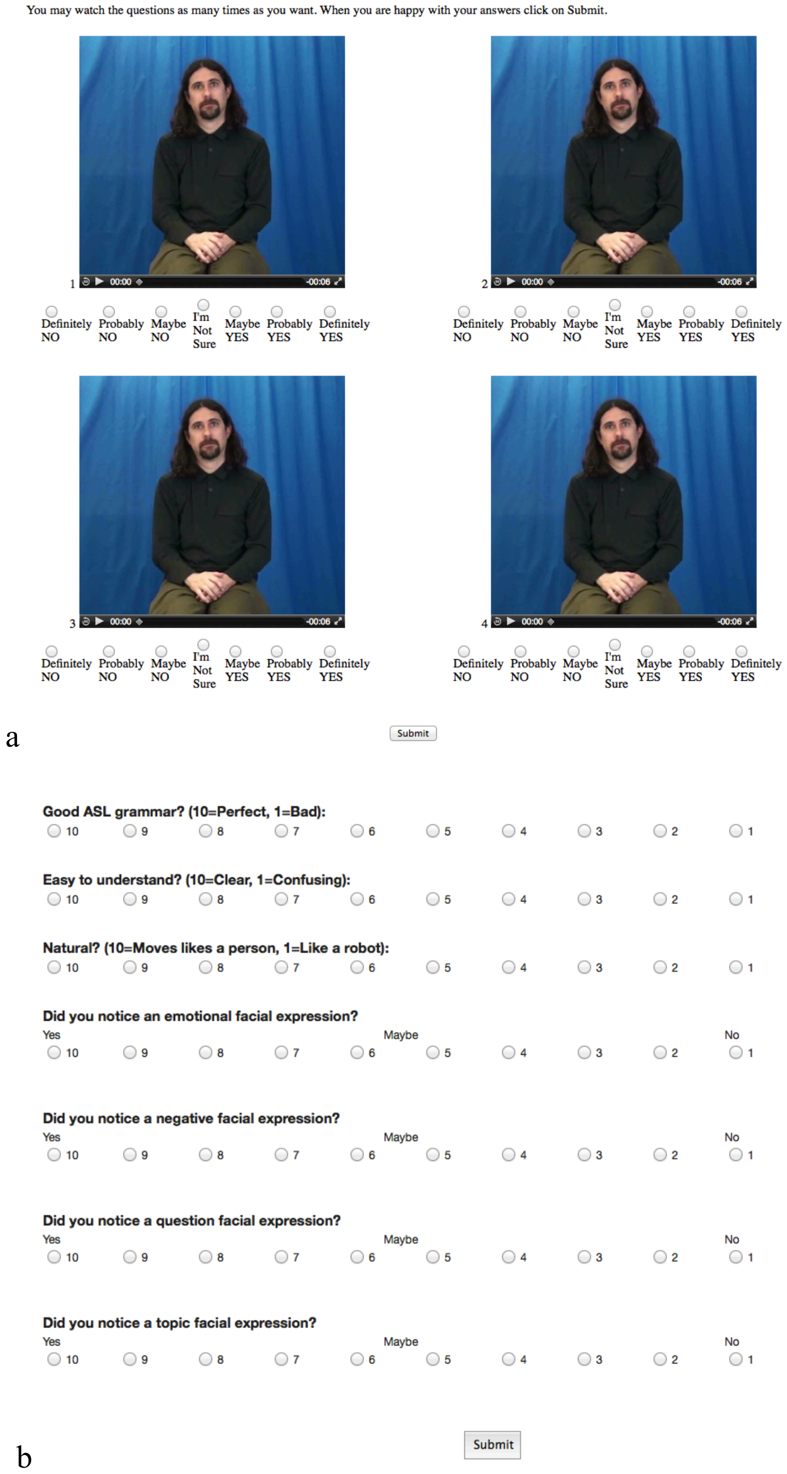
Table 15: Comprehension Questions and Their Suggested Answers based on the stimuli meaning when the ASL facial expression is correctly perceived. (In this table the questions are phrased in English.)

ID#	Question 1	Question 2	Question 3	Question 4	Q1 Answer	Q2 Answer	Q3 Answer	Q4 Answer
W1	Is Charlie asking you a question?	Does Charlie know where you saw Metallica?	Does Charlie think that music events are your favorite?	Does Charlie know where your favorite music show was?	NO	NO	YES	YES
W2	Is Charlie asking you a question?	Does Charlie know who your boss is?	Does Charlie think the manager is your boss?	Is Charlie's sister your boss?	YES	NO	NO	NO
W3	Is Charlie asking you a question?	Does Charlie know who your dentist is?	Does Charlie think that you met a dentist?	Did Charlie's brother your dentist?	YES	NO	NO	NO
W4	Is Charlie asking you a question?	Does Charlie know where you bought the computer?	Does Charlie think that Sally prefers shopping centers?	Did Charlie know where Sally's favorite shopping center is?	YES	NO	YES	YES-MAYBE
W5	Is Charlie asking you a question?	Is Mary drunk now?	Does Charlie know when the party is?	Did Mary's birthday party already happen?	YES	YES	NO	YES-MAYBE
W6	Is Charlie asking you a question?	Does Charlie know what bridge you visited?	Does Charlie think you prefer museums?	Did Charlie know what your favorite museum is?	YES	NO	YES	YES-MAYBE
W7	Is Charlie asking you a question?	Does Charlie know you saw a scary movie?	Does Charlie know the title of the movie you saw?	Does Charlie think that scary movies are your favorite kind of movie?	YES	YES	YES-MAYBE	NO-MAYBE

ID#	Question 1	Question 2	Question 3	Question 4	Q1 Answer	Q2 Answer	Q3 Answer	Q4 Answer
W8	Is Charlie asking you a question?	Does Charlie know how old your nephew is?	Will your uncle probably give your nephew a car for his birthday?	Does Charlie know what year your nephew was born?	YES	NO-MAYBE	YES-MAYBE	NO-MAYBE
W9	Is Charlie asking you a question?	Does Charlie know what kind of shoes you are wearing?	Does Charlie know where your school trip is going?	Does Charlie think your school trip is going mountain-climbing?	YES	NO-MAYBE	NO-MAYBE	YES-MAYBE
Y1	Is Charlie asking you a question?	Does Charlie know if the restaurant is expensive?	Has Charlie ever gone to the restaurant Bob's Diner?	Is your sister's favorite restaurant Bob's Diner?	YES	NO	NO	YES
Y2	Is Charlie asking you a question?	Does Charlie know if your sister likes Bob?	Does your sister like Bob?	Is Bob a teacher?	YES	NO	MAYBE	YES
Y3	Is Charlie asking you a question?	Does Charlie know where your sister lives?	Does Charlie think your sister lives in Washington?	Will Charlie visit your sister?	YES	NO	NO-MAYBE	YES
Y4	Is Charlie asking you a question?	Does Charlie know what kind of restaurant it is?	Did you already tell Charlie that the restaurant is Chinese?	Will Charlie go to the new restaurant?	YES	NO	NO	YES
Y5	Is Charlie asking you a question?	Does Charlie know whether he drives too fast?	Does Charlie think he drives too fast?	Will Charlie drive in a lower speed?	YES	NO	MAYBE	MAYBE
Y6	Is Charlie asking you a question?	Does Charlie know the restaurant is nice or not?	Does Charlie think Philadelphia is nice?	Will Charlie go to the fish restaurant?	YES	NO	NO	YES
Y7	Is Charlie asking you a question?	Does Charlie know where you live?	Does Charlie think you live in Rochester?	Will Charlie go to your graduation?	YES	NO	NO-MAYBE	YES
R1	Is Charlie asking you a question?	Does Charlie know why Alex is sick?	Does Charlie know why Alex is not wearing a jacket?	According to Charlie, is it true that when people don't wear jackets, then they get sick?	NO	YES	YES-MAYBE	YES
R2	Is Charlie asking you a question?	Did Sophie's lateness cause Charlie to miss the movie?	Does Charlie know why Sophie is late?	Did Charlie see the best movie?	NO	YES	YES-MAYBE	NO
R3	Is Charlie asking you a question?	Is Charlie buying groceries for you?	Does Charlie know why you are cooking a fancy dinner?	Are you cooking the potatos for your fancy dinner?	NO	YES	YES	YES
R4	Is Charlie asking you a question?	Is Charlie the one who loves seafood?	Is Charlie wondering why Chris like seafood?	Soes Charlie go for fishing because of Chris?	NO	NO	NO	YES
R5	Is Charlie asking you a question?	Are Sophie's and Aex's classes about piloting?	Is Charlie wondering why Sophie and Alex want to be pilots?	Does Charlie know why both Sophie and Alex take sam classes?	NO	YES	NO	YES
R6	Is Charlie asking you a question?	Is Alex divorced?	Is Charlie wondering why was Alex divorced?	Does Charlie know why Alex started going to parties?	NO	YES	NO-MAYBE	YES

ID#	Question 1	Question 2	Question 3	Question 4	Q1 Answer	Q2 Answer	Q3 Answer	Q4 Answer
R7	Is Charlie asking you a question?	Did Charlie say how he learned ASL?	Does Charlie know how to practice ASL?	Did Charlie learn ASL by taking classes?	NO	YES	YES	NO
R8	Is Charlie asking you a question?	Does Charlie know there will be pop quiz tomorrow?	Does Charlie know who invented light bulb?	Does Charlie believe the rumors?	YES	YES	NO	YES
R9	Is Charlie asking you a question?	Did Charlie buy knitting supplies at the store?	Does Charlie buy supplies for you?	Does Charlie know why you are decorating the wine bottle?	NO	YES	YES	YES
R10	Is Charlie wondering why Marissa quit her job?	Is Charlie wondering why Marissa started traveling more often?	Did Marissa’s job allow her to travel a lot?	Did Marissa’s job involve computers?	NO-MAYBE	NO	NO	NO-MAYBE
R11	Is Charlie asking you a question?	Does Charlie think you buy five houses every year?	Is Charlie planning to buy five houses this year?	Does Charlie know how he will achieve his business (goals)?	NO	NO	YES	YES
N1	Very soon, is Alex eating?	Will Alex eat Pasta soon?	Will Alex eat vegetables only?	Is the pasta ready now?	NO	NO	YES	NO
N2	Is Alex taking a math class this semester?	Is Alex taking a math class this semester?	Does the school have science classes this semester?	Is Alex taking two math classes?	YES	NO	NO	YES
N3	This Christmas, is Charlie visiting his sister?	This Christmas, is Charlie visiting his grandmother?	This Christmas, is Charlie visiting someone?	Does Charlie want to visit his sister?	NO	YES	NO	NO
N4	Did Charlie submit his English homework today?	Did Charlie submit his Math homework today?	Is Charlie's English homework ready?	Did Charlie finish all of his homework?	NO	YES	NO	NO
N5	This weekend, on Saturday, is the party for Charlie?	This weekend, Saturday, is it Charlie's birthday?	Will Charlie's sister have more friends at the party than Charlie?	Will Charlie get presents at the party?	NO	NO	YES	NO
N6	Did Justin forget his ticket?	Did Justin miss his train today?	Will Justin ride a bus today?	Is Justin stuck here in America?	YES	YES	YES-MAYBE	YES-MAYBE
T1	Does Charlie sell airplanes?	Is Charlie's brother an expert mechanic in boats?	Is Charlie asking you a question?	Does Charlie mention what kind of expert mechanic his brother is?	NO	NO	NO	YES
T2	Does Alex likes Mary more than Chris?	Is Charlie askinf you a question?	Is Chris Alex's best friend?	Does Charlie know to whom Alex is angry at?	YES	NO	NO-MAYBE	YES
T3	Does Charlie's sister like French restaurants?	Does Charlie's sister prefer French music?	Does Charlie's sister like Chinese music?	Does Charlie's sister like Chinese restaurants?	NO	YES	NO	YES

ID#	Question 1	Question 2	Question 3	Question 4	Q1 Answer	Q2 Answer	Q3 Answer	Q4 Answer
T4	Is Charlie's sister an expert at cooking sweets?	Is Charlie's sister an expert at cooking pizza?	Does the new restaurant have sweets?	Does the new restaurant only have pizza and pasta?	YES	NO	NO	YES
T5	Does train take a long time?	Does Charlie everyday ride the bus?	Everyday, does Charlie ride the train?	Does it take for Charlie a really long time to get to school?	NO	NO	YES	NO
T6	Do oranges taste good?	Does Charlie eat bacon for breakfast everymorning?	Every morning, does Charlie eat oranges for breakfast?	Does Charlie think that the combination of oranges with bacon is tasty?	NO-MAYBE	NO	YES	NO-MAYBE
T7	Does Charlie practice at the hospital every day?	Does Charlie think studying books is tough?	Does Charlie think practicing at the hospital is more difficult than studying books?	Does Charlie think medical school is tough?	NO	NO	YES	NO-MAYBE
E1	Is Charlie accusing you of something?	Is Charlie upset about something?	Does Charlie care about Charlie's opinion about him?	Did you just tell Charlie something?	YES	YES	YES	YES
E2	Is Charlie happy about the cat?	Is Charlie sad?	Was it Charlie that brought the cat?	Did Charlie sister bring a pet?	NO	YES	NO	YES
E3	Is Charlie angry at his sister?	Is Charlie upset about something?	Does Charlie like having visitors at work?	Did Charlie's sister visit him at work?	YES	YES	NO	YES
E4	Is Charlie happy about his birthday?	Is Charlie sad?	Is Charlie actually excited?	Is tomorrow Charlie's birthday?	NO	YES	NO	YES
E5	Is Charlie accusing you of something?	Is Charlie upset about something?	Does Charlie think you lost his computer?	Is Charlie asking about his computer?	YES	YES	YES	YES
E6	Does Charlie like the shirt?	Does Charlie like blue things?	Is Charlie disappointed?	Did Charlie's sister buy him new pants?	YES	NO	YES	YES
E7	Is Charlie angry at his brother?	Is Charlie upset about something?	Does Charlie like his brother driving his car?	Did Charlie's brother drive his bike?	YES	YES	NO	YES
E8	Did Charlie like the concert?	Does Charlie like country music?	Is Charlie disappointed?	Did Charlie go to a concert?	NO	NO	YES	YES



You may watch the questions as many times as you want. When you are happy with your answers click on Submit.

1 00:00 -00:06

○ Definitely NO ○ Probably NO ○ Maybe NO ○ I'm Not Sure ○ Maybe YES ○ Probably YES ○ Definitely YES

2 00:00 -00:06

○ Definitely NO ○ Probably NO ○ Maybe NO ○ I'm Not Sure ○ Maybe YES ○ Probably YES ○ Definitely YES

3 00:00 -00:06

○ Definitely NO ○ Probably NO ○ Maybe NO ○ I'm Not Sure ○ Maybe YES ○ Probably YES ○ Definitely YES

4 00:00 -00:06

○ Definitely NO ○ Probably NO ○ Maybe NO ○ I'm Not Sure ○ Maybe YES ○ Probably YES ○ Definitely YES

Good ASL grammar? (10=Perfect, 1=Bad):

○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Easy to understand? (10=Clear, 1=Confusing):

○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Natural? (10=Moves likes a person, 1=Like a robot):

○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Did you notice an emotional facial expression?

Yes ... Maybe ... No

○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Did you notice a negative facial expression?

Yes ... Maybe ... No

○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Did you notice a question facial expression?

Yes ... Maybe ... No

○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Did you notice a topic facial expression?

Yes ... Maybe ... No

○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Figure 83: Screenshots of (a) the comprehension questions and (b) the subjective/“notice” questions HTML forms.

APPENDIX B Best Practices For Conducting Evaluations of Sign Language Animations[24]

Our lab has conducted several research projects, prior and during this Ph.D. dissertation, to investigate experimental methodologies on learning to generate understandable sign language animations (surveyed in Huenerfauth, 2014). Informed by this prior work, including hundred of hours of studies with deaf participants, this appendix summarizes best practices for conducting such evaluations.

Identifying and Screening Participants

When humans evaluate a language generation system, it is important for them to be native speakers of that language: proper screening is needed to ensure that these judges are sufficiently critical of the system's output (Neidle et al., 2000). An ideal participant is a "native signer," someone who learned ASL in early childhood through interactions at home or through significant time in a school environment using ASL. We have effectively advertised for such participants in metropolitan areas through distributing messages to online groups and email lists and through hiring recruiters from the local Deaf community. We have also found that it is ineffective to screen potential participants by asking questions such as "How well do you sign?," "Are you a native signer?," or "Is ASL your first language?" (Huenerfauth et al., 2007). Such questions could be misinterpreted as asking whether the individual feels personally oriented toward Deaf culture. We have instead found it effective to ask whether the potential participant has had life experiences typical of a native signer: "Did you grow up using

[24] Part of this appendix has first appeared as an article in (Huenerfauth and Kacorri, 2015a).

ASL as a child?," "Did your parents use ASL at home?," "Did you attend a residential school where you used ASL?," etc.

We found that demographics and technology experience factors can influence the evaluation scores that participants assign to sign language animation during a study (Chapter 7). We encourage researchers to report these factors (shared as ASL videos and English text in our lab: http://latlab.ist.rit.edu/assets2015/). When evaluating animation approaches with subtle differences, e.g., variations of facial expression models, it might be useful to screen for participants whose demographics factors in Chapter 7 indicated 'harsher criticism': participants who self-identified as "deaf/Deaf" or "hard-of-hearing," who had grown up using ASL at home or had attended an ASL-based school as a young child, such as a residential or daytime school.

Controlling the Experimental Environment

When seeking grammaticality judgments from signers, it is important to minimize environmental characteristics which may prompt signers to code-switch to more spoken-language-like forms of signing or accept such signing as grammatically correct (Neidle et al., 2000). Many signers are accustomed to switching to such signing in interactions with hearing individuals. To avoid this, participants should be exposed only to fluent sign language during the study (Huenerfauth et al., 2007). Instructions should be signed by another native signer. If possible, participants should be immersed in a sign language environment prior to the study, e.g., engaging in conversation in fluent ASL prior to the study. If interpreters are required, they should possess near-native sign language fluency (Huenerfauth et al., 2007).

As with any study, users must feel comfortable criticizing the system being evaluated. In this context, it is important that the participant not feel that anyone responsible for the system is sitting with them while they critique it – or else they may not feel as comfortable offering negative opinions about the system. If a native signer is “hosting” the study, it is helpful for this person to present themselves as an “outsider” to the technical team that had created the animations being evaluated (Huenerfauth et al., 2007).

Engineering Stimuli for Studies

Inventing stimuli that contain specific linguistic phenomena and measure whether participants understand the intended information is challenging – but necessary for effectively evaluating ASL animations. For instance, in Chapter 4, we have described how to engineer animation stimuli that can be interpreted (ambiguously) in different ways, depending on whether a particular aspect of the sentence was successfully understood by the participant (e.g., whether a particular ASL facial expression was correctly perceived). In this way, comprehension questions can be invented that specifically measure whether this aspect of the animation was correct, thereby enabling an evaluation of that specific issue. To aid researchers, we have published our methods for designing stimuli for a variety of linguistic phenomena in ASL, and we have also released a collection of stimuli for evaluating ASL facial expressions (Chapter 4 and Appendix A).

Engineering Subjective Evaluation Questions for Studies

To measure user’s satisfaction with sign language animations after viewing stimuli, we have asked participants to answer subjective questions concerning grammatical correctness, ease of understanding, and naturalness of movement of the virtual human character. To ensure that they are clearly communicated, these questions are explained in sign language, and participants select answer choices on 1-to-10 scales. In (Huenerfauth et al., 2007), we observed that the scores measuring grammaticality, naturalness, and understandability were moderately correlated, which was understandable since the grammaticality and naturalness of an animation could affect its perceived understandability. In other studies, we have also asked participants to rate on a 1-to-10 scale how confident they were that they had noticed specific phenomena of interest in the animations, e.g., a specific facial expression. A questionnaire with both types of subjective/notice questions and their answer choices was released in Appendix A.

Engineering Comprehension Evaluation Questions for Studies

While it is relatively easy to ask a participant to rate subjectively whether they believe a particular animation stimulus was understandable, we have observed low correlation between a user's subjective impression of the understandability of a sign language animation and his/her actual success at answering comprehension questions about that animation, e.g., (Huenerfauth et al., 2007). It is for this reason that we have made efforts to include an actual comprehension task (either a comprehension question about information content in the stimulus or a matching task that the user must perform based on this information). We have discussed how users' perceived understandability scores are not an adequate substitute for this actual comprehension data.

To obtain reliable scores, researchers must ensure that spoken-language skills are not necessary for participants to understand comprehension questions or answer choices. In prior work, we have presented comprehension questions in sign language, e.g. using videos of a native singer or high quality animations created by a native signer. We found that presenting questions as video or high-quality animation did not affect comprehension scores (Chapter 5). To present answer choices, we have successfully used image matching (Huenerfauth et al., 2007), clip-art illustrations for answer choices (Huenerfauth, 2008), or definitely-no-to-definitely-yes 7-point scalar responses (Appendix A).

As discussed above, for comprehension scores to be meaningful, they must be engineered to probe whether participants have understood the intended information specifically conveyed by the aspect of the animation that the researcher wishes to evaluate. This is particular challenging for non-manual components of animation, e.g. facial expressions (Chapter 4). To aid other researchers in conducting studies, we have released to the research community a set of 192 comprehension questions for ASL stimuli with facial expressions (Appendix A).

Use of Baselines for Comparison

In general, the *absolute* scores recorded from questions in a study are difficult to interpret unless they can be considered *relative* to some baselines for comparison. This is because the absolute scores in a study may depend on a variety of factors beyond the animation-quality, e.g., the difficulty of the stimuli and the comprehension questions, participants' memory skills, etc. Thus, in addition to the to-be-evaluated version of an animation stimulus, we include other stimuli in a study so that the relative scores can be compared.

As a "lower baseline" for comparison, we have found it effective to present users with a version of the ASL animation that differs from the stimuli by excluding only the features being evaluated, e.g., if we are evaluating a method to add a particular facial expression to an animation, the lower baseline will lack this facial expression. A good "upper baseline" should represent an "ideal" system output and may consist of a high-quality computer animation or a video recording of a human signer (performing identical sentences to the virtual human in the animations). We compare both approaches in (Chapter 5).

Eye-tracking metrics in Evaluation Studies

Researchers sometimes need to measure users' reactions to animations without obtrusively directing participants' attention to the new features being incorporated; in such cases, we have investigated the use of eye-tracking technologies to evaluate stimuli (Chapter 6). We divided the screen region where the stimuli appear to three areas of interest: "Upper Face", "Lower Face", and "Hands". We found that the time-normalized fixation trail length metric should be utilized if seeking an eye metric that correlates with participants' subjective judgments about ASL videos or animations.

APPENDIX C CPM Training Set Per ASL Facial Expression Subcategory

This appendix includes detailed information on the linguistically diverse dataset that was used to train the CPMs for the ASL facial expressions, discussed in Chapter 12 and Chapter 13. Per each subcategory, we report the filenames of the videos in the ASLLRP corpora[25], and portion of the extracted facial expression with a reference to the start frame, end frame, and duration.

Table 16: Video recordings (9) in the CPM training set for YesNo_A.

Video	Start Frame	End Frame	Duration
2011-12-01_0037-cam2-00	22	72	51
2011-12-01_0037-cam2-02	12	29	18
2011-12-01_0037-cam2-03	20	50	31
2011-12-01_0037-cam2-05	20	64	45
2011-12-01_0037-cam2-10	52	77	26
2011-12-01_0037-cam2-13	56	94	39
2011-12-01_0042-cam2-10	36	70	35
2012-01-27_0051-cam2-04	49	86	38
2012-01-27_0052-cam2-18	62	96	35

[25] ASL videos were created and annotated by linguists and computer scientists (Neidle and Sclaroff at BU; Athitsos, now at U. Texas, Arlington; and Metaxas at Rutgers) as part of the "Corpora Through the American Sign Language Linguistic Research Project" (ASLLRP) and the NSF-funded collaborative project "Generating Accurate Understandable Sign Language Animations Based on Analysis of Human Signing."

Table 17: Video recordings (10) in the CPM training set for YesNo_B.

Video	Start Frame	End Frame	Duration
2011-12-01_0036-cam2-20	35	76	42
2011-12-01_0037-cam2-01	6	60	55
2011-12-01_0037-cam2-06	5	82	78
2011-12-01_0037-cam2-07	6	61	56
2011-12-01_0037-cam2-08	11	60	50
2011-12-01_0037-cam2-09	6	61	56
2011-12-01_0037-cam2-11	0	38	39
2011-12-01_0037-cam2-12	5	49	45
2012-01-27_0051-cam2-09	70	90	21
2012-01-27_0052-cam2-02	7	45	39

Table 18: Video recordings (4) in the CPM training set for WhQuestion_A.

Video	Start Frame	End Frame	Duration
2011-12-01_0038-cam2-04	38	57	20
2011-12-01_0038-cam2-05	44	64	21
2011-12-01_0038-cam2-06	24	47	24
2011-12-01_0038-cam2-10	58	75	18

Table 19: Video recordings (8) in the CPM training set for WhQuestion_B.

Video	Start Frame	End Frame	Duration
2011-12-01_0038-cam2-00	34	59	26
2011-12-01_0038-cam2-01	5	42	38
2011-12-01_0038-cam2-02	26	65	40
2011-12-01_0038-cam2-03	29	69	41
2011-12-01_0038-cam2-07	36	68	33
2011-12-01_0038-cam2-08	46	76	31
2011-12-01_0038-cam2-09	15	51	37
2011-12-01_0038-cam2-11	37	77	41

Table 20: Video recordings (2) in the CPM training set for Rhetorical_A.

Video	Start Frame	End Frame	Duration
2011-12-01_0041-cam2-04	42	54	13
2011-12-01_0041-cam2-05	62	77	16

Table 21: Video recordings (8) in the CPM training set for Rhetorical_B.

Video	Start Frame	End Frame	Duration
2011-12-01_0041-cam2-00	19	41	23
2011-12-01_0041-cam2-02	5	53	49
2011-12-01_0041-cam2-06	1	44	44
2011-12-01_0041-cam2-07	2	34	33
2011-12-01_0041-cam2-08	14	68	55
2011-12-01_0041-cam2-09	3	35	33
2011-12-01_0041-cam2-10	7	44	38
2012-01-27_0051-cam2-25	18	57	40

Table 22: Video recordings (29) in the CPM training set for Topic_A.

Video	Start Frame	End Frame	Duration
2011-12-01_0038-cam2-03	7	24	18
2011-12-01_0038-cam2-04	6	16	11
2011-12-01_0038-cam2-10	23	34	12
2011-12-01_0039-cam2-01	7	20	14
2011-12-01_0039-cam2-06	9	27	19
2012-01-27_0050-cam2-00	27	39	13
2012-01-27_0050-cam2-01	34	47	14
2012-01-27_0050-cam2-03	38	50	13
2012-01-27_0050-cam2-04	36	47	12
2012-01-27_0050-cam2-05	23	41	19
2012-01-27_0050-cam2-08	33	46	14
2012-01-27_0050-cam2-20	2	16	15
2012-01-27_0050-cam2-21	15	25	11
2012-01-27_0051-cam2-01	21	42	22
2012-01-27_0051-cam2-06	8	15	8
2012-01-27_0051-cam2-11	7	17	11
2012-01-27_0051-cam2-20	7	24	18
2012-01-27_0051-cam2-21	1	39	29
2012-01-27_0051-cam2-22	12	28	17
2012-01-27_0051-cam2-23	0	25	26
2012-01-27_0051-cam2-24	0	15	16
2012-01-27_0052-cam2-00	16	32	17
2012-01-27_0052-cam2-04	14	33	20
2012-01-27_0052-cam2-07	9	20	12
2012-01-27_0052-cam2-09	7	21	15
2012-01-27_0052-cam2-13	10	22	13
2012-01-27_0052-cam2-14	4	17	14

Table 23: Video recordings (15) in the CPM training set for Topic_B.

Video	Start Frame	End Frame	Duration
2011-12-01_0038-cam2-07	3	24	22
2011-12-01_0039-cam2-02	9	24	16
2011-12-01_0039-cam2-04	5	36	32
2011-12-01_0039-cam2-07	12	35	24
2011-12-01_0039-cam2-08	22	53	32
2011-12-01_0039-cam2-09	25	42	18
2011-12-01_0039-cam2-10	9	42	34
2011-12-01_0039-cam2-11	77	90	14
2011-12-01_0042-cam2-11	2	22	21
2012-01-27_0050-cam2-02	29	57	29
2012-01-27_0050-cam2-10	29	50	22
2012-01-27_0051-cam2-09	2	46	45
2012-01-27_0051-cam2-10	54	79	26
2012-01-27_0051-cam2-14	7	32	26
2012-01-27_0052-cam2-08	1	19	19

Table 24: Video recordings (16) in the CPM training set for Negative_A.

Video	Start Frame	End Frame	Duration
2011-12-01_0036-cam2-09	37	93	57
2011-12-01_0039-cam2-00	30	69	40
2011-12-01_0040-cam2-01	34	71	38
2011-12-01_0040-cam2-03	27	64	38
2011-12-01_0040-cam2-06	42	89	48
2011-12-01_0040-cam2-10	45	72	28
2012-01-27_0050-cam2-06	41	88	48
2012-01-27_0050-cam2-07	30	50	21
2012-01-27_0050-cam2-08	46	85	40
2012-01-27_0050-cam2-19	22	88	67
2012-01-27_0051-cam2-03	22	41	20
2012-01-27_0051-cam2-08	24	50	27
2012-01-27_0051-cam2-12	30	88	59
2012-01-27_0051-cam2-25	57	74	18
2012-01-27_0051-cam2-26	37	60	24
2012-01-27_0051-cam2-27	77	102	26

Table 25: Video recordings (25) in the CPM training set for Negative_B.

Video	Start Frame	End Frame	Duration
2011-12-01_0036-cam2-03	9	45	37
2011-12-01_0036-cam2-03	63	101	39
2011-12-01_0039-cam2-02	34	51	18
2011-12-01_0039-cam2-03	104	134	31
2011-12-01_0039-cam2-10	48	98	51
2011-12-01_0039-cam2-11	104	147	44
2011-12-01_0040-cam2-00	28	92	65
2011-12-01_0040-cam2-02	13,	59	47
2011-12-01_0040-cam2-05	11	76	66
2011-12-01_0040-cam2-07	5	69	65
2011-12-01_0040-cam2-09	44	79	36
2011-12-01_0040-cam2-11	18	56	39
2011-12-01_0042-cam2-06	51	87	37
2012-01-27_0050-cam2-05	46	86	41
2012-01-27_0050-cam2-11	41	98	58
2012-01-27_0050-cam2-16	95	160	66
2012-01-27_0050-cam2-17	113	153	41
2012-01-27_0050-cam2-22	5	80	76
2012-01-27_0051-cam2-02	5	65	61
2012-01-27_0051-cam2-04	25	43	19
2012-01-27_0051-cam2-10	11	46	36
2012-01-27_0051-cam2-13	26	56	31
2012-01-27_0051-cam2-28	7	56	50
2012-01-27_0051-cam2-30	17	54	38
2012-01-27_0052-cam2-15	97	128	32

Bibliography

Adamo-Villani, N., and Wilbur, R. 2010. Software for math and science education for the deaf. *Disability and Rehabilitation: Assistive Technology*, 5(2), 115–124.

Ahlberg, J., Pandzic, I.S., and You, L. 2002. Evaluating face models animated by MPEG-4. In *I.S. Pandzic, R. Forchheimer (eds.), MPEG-4 facial animation: the standard, implementations and applications*, Wiley & Sons, 291–296.

Allbritton, D.W., Mckoon, G., and Ratcliff, R. 1996. Reliability of prosodic cues for resolving syntactic ambiguity. *J. Exp. Psychol.-Learn. Mem. Cogn.*, 22, 714–735.

Applied Science Labs. 2013. Homepage. http://www.asleyetracking.com.

Asghar, A., and Rao, N.I. 2008. Color image segmentation using multilevel clustering approach. In *Digital Image Computing: Techniques and Applications (DICTA)*, 519–524. IEEE.

Baker-Shenk, C. 1983. A micro analysis of the nonmanual components of American Sign Language. *Unpublished Doctoral Dissertation*, U of California, Berkeley, U.S.A.

Baker-Shenk, C.L. 1991. American Sign Language: A teacher's resource text on grammar & culture. Gallaudet U.

Baldassarri, S., Cerezo, E., and Royo-Santas, F. 2009. Automatic translation system to Spanish Sign Language with a virtual interpreter. In *Proc of the 12th IFIP TC Int'l Conf on Human-Computer Interaction: Part I (INTERACT),* 196-199. Springer-Verlag, Berlin, Heidelberg

Baldassarri, S., and Cerezo, E. 2012. Maxine: Embodied conversational agents for multimodal emotional communication, *Computer Graphics*. InTech.

Bartels, R.H., Beatty, J.C., and Barsky, B.A. 1987. An introduction to splines for use in computer graphics and geometric modeling. Morgan Kaufmann.

Bartels, M., and Marshall, S.P., 2006. Eye tracking insights into cognitive modeling. In *Proc. of the Symp on Eye Tracking Research & Applications*. ACM.

Bergmann, K. 2012. The production of co-speech iconic gestures: empirical study and computational simulation with virtual agents. *Dissertation*, Bielefeld U, Germany.

Boster, C. 1996. On the quantifier-noun phrase split in American Sign Language and the structure of quantified noun phrases. In *Int'l Review of Sign Linguistics*. 159–208. Mahwah, NJ: Lawrence Erlbaum Associates.

Boulares, M., and Jemni, M. 2012. Toward an example-based machine translation from written text to ASL using virtual agent animation. In *Proc. of CoRR*.

Catmull, E., and Rom, R. 1974. A class of local interpolating splines. *Computer aided geometric design*, 74, 317-326.

Cavender, A.C, Bigham, J.P., and Ladner, R.E. 2009. ClassInFocus: enabling improved visual attention strategies for deaf and hard of hearing students. In *Proc of the 11th Int'l ACM SIGACCESS Conf on Computers and Accessibility (ASSETS)*, 67-74. ACM.

Cavender, A.C., Rice, E.A., and Wilamowska, K.M. 2005. SignWave: Human perception of sign language video quality as constrained by mobile phone technology. Retrieved from:

http://www.cs.washington.edu/education/courses/cse510/05sp/ project-reports/cse510-signwave.pdf

Center for Research and Education on Aging and Technology Enhancement (CREATE). 2015. Retrieved on May 6, 2015 from http://create-center.gatech.edu/resources.php.

Chapdelaine, C., Gouaillier, V., Beaulieu, M., and Gagnon, L., 2007. Improving video captioning for deaf and hearing-impaired people based on eye movement and attention overload. In *Electronic Imaging*, 64921-64921. Int'l Society for Optics and Photonics.

Cox, S., Lincoln, M., Tryggvason, J., Nakisa, M., Wells, M., Tutt, M., and Abbott, S. 2002. TESSA, a system to aid communication with deaf people. In *Proc of the 5th Int'l ACM Conf on Assistive Technologies,* 205-212. ACM.

Crabb, M., and Hanson, V. L. 2014. Age, technology usage, and cognitive characteristics in relation to perceived disorientation and reported website ease of use. In *Proc. of the 16th Int'l ACM SIGACCESS Conf on Computers and Accessibility (ASSETS),* 193-200, ACM.

Crasborn, O., Sloetjes, H., Auer, E., and Wittenburg, P. 2006. Combining video and numeric data in the analysis of sign languages within the ELAN annotation software. In *2nd Workshop on the Representation and Processing of Sign Languages (LREC)*, 82–87.

Davidson, M.J., Alkoby, K., Sedgwick, E., Berthiaume, A., Carter, R., Christopher, J., Craft, B., Furst, J., Hinkle, D., Konie, B., Lancaster, G., Luecking, S., Morris, A., McDonald, J., Tomuro, N., Toro, J., and Wolfe, R. 2000. Usability testing of computer animation of fingerspelling for American Sign Language. *Presented at the 2000 DePaul CTI Research Conference*, Chicago, IL.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society*. Series B (methodological), 1-38.

DeWitt, L.C., Lam, T.T., Silverglate, D.S., Sims, E.M., and Wideman, C.J. 2003. Method for animating 3-D computer generated characters. *U.S. Patent No. 6,535,215*. Washington, DC. U.S. Patent and Trademark Office.

DGS-KORPUS. 2014. HamNoSys. Retrieved on August 10, 2014 from http://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/hamnosys-97.html.

Dicta-Sign. 2014. Corpus by Task and Languages. Retrieved on August 5, 2014 from http://www.sign-lang.uni-hamburg.de/dicta-sign/portal/task.html.

Duarte, K. and Gibet, S. 2010. Corpus design for signing avatars. In *Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (LREC)*, 1-3.

Duchowski A., 2002. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, 34(4), 455-470.

Ebling, S. and Glauert, J. 2013. Exploiting the full potential of JASigning to build an avatar signing train announcements. In *3rd Int'l Symp on Sign Language Translation and Avatar Technology (SLTAT)*, Chicago, IL, USA.

Ebling, S. and Glauert, J. 2015. Building a Swiss German Sign Language avatar with JASigning and evaluating it among the Deaf community. In *Univ Access Inf Soc.*

Efthimiou, E., Fontinea, S.E., Hanke, T., Glauert, J., Bowden, R., Braffort, A., Collet, C., Maragos, P., and Goudenove, F. 2010. Dicta-sign–sign language recognition, generation and modeling: a research effort with applications in deaf communication. In *Proc of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies* (LREC), 80-83.

Ekman, P., and Friesen, W. 1978. Facial action coding system: a technique for the measurement of facial movement. *Consulting Psychologists*, San Francisco.

Elliott, R., Glauert, J.R.W., Jennings, V., and Kennaway, J.R. 2004. An overview of the SiGML notation and SiGMLSigning software system. In *Proc of the 4th Int'l Conf on Language Resources and Evaluation (LREC)*, 98–104, Lisbon, Portugal.

Elliott, R., Glauert, J., Kennaway, J., Marshall, I., and Safar, E. 2008. Linguistic modeling and language-processing technologies for avatar-based sign language presentation. *Univ Access Inf Soc* ,6(4), 375-391. Berlin: Springer.

Elliott, R., Glauert, J. R., Kennaway, J.R., and Marshall, I. 2000. The development of language processing support for the ViSiCAST project. In *Proc of the 4th Int'l ACM Conf on Assistive Technologies*, 101-108. ACM.

Elliott, R., Glauert, J.R.W., Kennaway, J.R., Marshall, I., and Safar, E. 2007. Linguistic modelling and language-processing technologies for avatar-based sign language presentation. Universal Access in the Information Society, 6(4):375–391.

Emmorey, K., Thompson, R., and Colvin, R. 2009. Eye gaze during comprehension of American Sign Language by native and beginning signers. *J of Deaf Studies and Deaf*

Education, 14(2), 237-243.

Filhol, M., Delorme, M., and Braffort, A. 2010. Combining constraint-based models for sign language synthesis. In *Proc of 4th Workshop on the Representation and Processing of Sign Languages, Language Resources and Evaluation (LREC)*, Malta.

Fotinea, S.E., Efthimiou, E., Caridakis, G., and Karpouzis, K. 2008. A knowledge-based sign synthesis architecture. *Univ Access Inf Soc,* 6(4), 405-418. Berlin: Springer.

Fox, J. and Monette, G. 1992. Generalized collinearity diagnostics. *JASA* 87, 178-183.

Garau, M., Slater, M., Bee, S., and Sasse, M.A. 2001. The impact of eye gaze on communication using humanoid avatars. In *Proc of the SIGCHI Conf on Human Factors in Computing Systems (CHI)*, 309-316. ACM.

Garnier, S., Ross, N., and Rudis, B. 2015. Viridis. R package version 0.3.2.

Gelman, A. 2008. Scaling regression inputs by dividing by two standard deviations. *Stat Med* 27(15), 2865-2873.

Gibet, S., Courty, N., Duarte, K., and Naour, T.L. 2011. The SignCom system for data-driven animation of interactive virtual signers: Methodology and Evaluation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(1), 6.

Giorgino, T. 2009. Computing and visualizing dynamic time warping alignments in R: the dtw package, *Journal of Statistical Software*, 31(7), 1–24.

Goldberg, J.H., Stimson, M.J., Lewenstein, M., Scott, N., and Wichansky, A.M. 2002. Eye tracking in web search tasks: design implications. In *Proc of the Symp on Eye-Tracking*

Research & Applications, 51-58. ACM.

Goldberg, J.H., and Kotval, X.P., 1999. Computer interface evaluation using eye movements: methods and constructs. *Int'l J Ind Ergonom,* 24(6) 631-645.

Grömping, U. 2006. Relative importance for linear regression in R: the package relaimpo. *Journal of statistical software* 17(1), 1-27.

Grossman, R.B., and Shepard-Kegl, J.A. 2006. To capture a face: A novel technique for the analysis and quantification of facial expressions in American Sign Language. *Sign Language Studies*, 6(3), 273-305.

Guan, Z., and Cutrell, E. 2007. An eye tracking study of the effect of target rank on web search. In *Proc of the SIGCHI Conf on Human Factors in Computing Systems*, 417-420. ACM.

Hagberg, A.A., Schult, D.A., and Swart, P.J. 2008. Exploring network structure, dynamics, and function using NetworkX. In *Proc of the 7th Python in Science Conference (SciPy)*, Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), 11–15.

Halverson, T., and Hornof, A. J. 2007. A minimal model for predicting visual search in human-computer interaction. In *Proc of the SIGCHI Conf on Human Factors in Computing Systems (CHI)*, 431-434. ACM.

Ham, R.T., Theune, M., Heuvelman, A. and Verleur, R. 2005. Judging Laura: perceived qualities of a mediated human versus an embodied agent. In *Intelligent Virtual Agents*, 381-393. Springer Berlin / Heidelberg.

Hanke, T. 2001. ViSiCAST Deliverable D5-1: Interface Definitions. *Technical Report.*

ViSiCAST project. http://www.visicast.co.uk/members/milestones/D5-1rev1.pdf

Hanke, T. 2010. HamNoSys–Hamburg Notation System for Sign Languages, *The 3rd Workshop of the Sign Linguistics Corpora Network*, Germany.

Hayward, K., Adamo-Villani, N., and Lestina, J. 2010. A computer animation system for creating deaf-accessible math and science curriculum materials. In *Eurographics'10*.

Heloir, A., and Kipp, M. 2009. EMBR–A realtime animation engine for interactive embodied agents. In *Intelligent Virtual Agents*, 393-404. Springer Berlin Heidelberg.

Heloir. A., Nguyen, Q., and Kipp, M. 2011. Signing avatars: a feasibility study. *The 2nd Int'l Workshop on Sign Language Translation and Avatar Technology (SLTAT),* Dundee, Scotland, United Kingdom.

Hsu, E., da Silva, M., and Popović, J. 2007. Guided time warping for motion editing. In *Proc of the ACM SIGGRAPH/Eurographics symposium on computer animation*, 45-52. Eurographics Association.

Huenerfauth, M. 2004. Spatial and planning models of ASL classifier predicates for machine translation. In *Proc of the 10th Int'l Conf on Theoretical and Methodological Issues in Machine Translation (TMI)*.

Huenerfauth, M. 2006. Generating American Sign Language classifier predicates for English-to-ASL machine translation. *Doctoral Dissertation, Computer and Information Science*, U of Pennsylvania.

Huenerfauth, M. 2008. Evaluation of a psycholinguistically motivated timing model for animations of American Sign Language. In *Proc of the 10th Int'l ACM SIGACCESS Conf on Computers and Accessibility (ASSETS)*, 129-136. ACM.

Huenerfauth, M. 2014. Learning to Generate Understandable Animations of American Sign Language. *2nd Annual Effective Access Technology Conference.*

Huenerfauth, M., and Hanson, V. 2009. Sign Language in the interface: access for deaf signers. In *Universal Access Handbook*, 38,1-18. NJ: Erlbaum.

Huenerfauth, M., and Kacorri, H. 2014. Release of experimental stimuli and questions for evaluating facial expressions in animations of American Sign Language. In *Proc of the 6th Workshop on the Representation and Processing of Sign Languages (LREC)*, Reykjavik, Iceland.

Huenerfauth, M. and Kacorri, H. 2015a. Best Practices for Conducting Evaluations of Sign Language Animation. *Journal on Technology and Persons with Disabilities,* Volume 3, September 2015, California State University, Northridge.

Huenerfauth, M. and Kacorri, H. 2015b. Augmenting EMBR virtual human animation system with MPEG-4 controls for producing ASL facial expressions. *The 5th Int'l Workshop on Sign Language Translation and Avatar Technology (SLTAT),* Paris, France.

Huenerfauth, M., and Lu, P. 2010. Modeling and synthesizing spatially inflected verbs for American Sign Language animations. In *Proc of the 12th Int'l ACM SIGACCESS Conf on Computers and Accessibility (ASSETS),* 99-106. ACM.

Huenerfauth, M., Lu, P. 2012. Effect of spatial reference and verb inflection on the usability of

American Sign Language animation. In *Univ Access Inf Soc*. Berlin: Springer.

Huenerfauth, M., Lu, P., and Rosenberg, A. 2011. Evaluating Importance of Facial Expression in American Sign Language and Pidgin Signed English Animations. In *Proc of the 13th Int'l ACM SIGACCESS Conf on Computers and Accessibility (ASSETS)*, 99-106. ACM.

Huenerfauth, M., Zhao, L., Gu, E., and Allbeck, J. 2008. Evaluation of American Sign Language generation by native ASL signers. *ACM Transactions on Accessible Computing (TACCESS)*, 1(1), 3.

Huenerfauth, M., Zhou, L., Gu, E., and Allbeck, J. 2007. Evaluating American Sign Language generation through the participation of native ASL signers. In *Proc of the 9th Int'l ACM SIGACCESS Conf on Computers and Accessibility (ASSETS)*. New York: ACM Press.

Hurdich, J. 2008. Utilizing Lifelike, 3D Animated Signing Avatar Characters for the Instruction of K-12 Deaf Learners. In *Proc of the Exploring Instructional and Access Technologies Symp National Tech Institute for the Deaf*, New York, USA.

ISO/IEC 14496-2. 1999. Information technology -- Coding of audio-visual objects -- Part 2: Visual.

Jacob, R.J.K., and Karn, K.S. 2003. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *The Mind's Eye (First Edition). J. Hyönä, R. Radach and H. Deubel*, 573-605. Amsterdam.

Jennings, V., Elliott, R., Kennaway, R., and Glauert, J. 2010. Requirements for a signing avatar. *In 4th Workshop on the Representation and Processing of Sign Languages:*

Corpora and Sign Language Technologies (LREC), 33-136. Valetta, Malta.

Jensen, S.S., and Pedersen, T. 2011. Eye tracking deaf people's metatcognitive comprehension strategies on the Internet. *Master's Thesis*. Retrieved from http://projekter.aau.dk/projekter/files/52662579/SamletProjekt.pdf

Kacorri, H. 2015. TR-2015001: A Survey and Critique of Facial Expression Synthesis in Sign Language Animation. *Computer Science Technical Reports.* Paper 403.

Kacorri, H., Harper, A., and Huenerfauth, M. 2013a. Comparing native signers' perception of American Sign Language animations and videos via eye tracking. In *Proc of the 15th Int'l ACM SIGACCESS Conf on Computers and Accessibility (ASSETS)*, 9. ACM.

Kacorri, H., Lu, P., and Huenerfauth, M. 2013b. Effect of displaying human videos during an evaluation study of American Sign Language animation. *ACM Transactions on Accessible Computing (TACCESS)*, 5(2), 4.

Kacorri, H., Lu, P., and Huenerfauth, M. 2013c. Evaluating facial expressions in American Sign Language animations for accessible online information. In *Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion*, 510-519. Springer Berlin Heidelberg.

Kacorri, H., Harper, A., and Huenerfauth, M. 2014. Measuring the perception of facial expressions in American Sign Language animations with eye tracking. In *Universal Access in Human-Computer Interaction. Design for All and Accessibility Practice*, 553-563. Springer International Publishing.

Kacorri, H. and Huenerfauth, M. 2014. Implementation and evaluation of animation controls sufficient for conveying ASL facial expressions. In *Proc of the 16th Int'l ACM SIGACCESS Conf on Computers and Accessibility (ASSETS).* ACM.

Kacorri, H. and Huenerfauth, M. 2015a. Comparison of Finite-Repertoire and Data-Driven Facial Expressions for Sign Language Avatars. *Universal Access in Human-Computer Interaction. Access to Interaction*, Vol. 9176 of the series Lecture Notes in Computer Science, 393-403.

Kacorri, H. and Huenerfauth, M. 2015b. Evaluating a dynamic time warping based scoring algorithm for facial expressions in ASL animations. In *Proc of the 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), INTERSPEECH 2015*, Dresden, Germany.

Kacorri, H., Huenerfauth, M., Ebling, S., Patel, K., and Willard, M. 2015. Demographic and experiential factors influencing acceptance of sign language animation by Deaf users. In *Proc of the 17th ACM SIGACCESS Conf on Computers and Accessibility (ASSETS).* Lisbon, Portugal. New York: ACM Press.

Kennaway, J.R., Glauert, J.R., and Zwitserlood, I. 2007. Providing signed content on the Internet by synthesized animation. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 14(3), 15.

Khan, M.A. and Ohno, Y, 2007. An automated video data compression algorithm by Cardinal spline fitting. *NICOGRAPH Int'l Conf*, Toyota, Japan.

Kipp, M., Heloir, A., and Nguyen, Q. 2011a. Sign language avatars: Animation and comprehensibility. In *Intelligent Virtual Agents*, 113-126. Springer Berlin Heidelberg.

Kipp, M., Nguyen, Q., Heloir, A., and Matthes, S. 2011b. Assessing the deaf user perspective on sign language avatars. In *Proc of the 13th Int'l ACM SIGACCESS Conf on Computers and Accessibility (ASSETS)*, 107-114. ACM.

Kowler, E. 2011. Eye movements: The past 25years. *Vision Research*, 51(13), 1457-1483.

Kreyszig, E. (2005). Advanced Engineering Mathematics (9 ed.), 816. Wiley

Krnoul, Z., Kanis, J., Zelezny, M., and Muller, L. 2008. Czech text-to-sign speech synthesizer. In *Machine Learning for Multimodal Interaction*, 180-191. Springer Berlin Heidelberg.

Lillo-Martin, D. 2000. Aspects of the syntax and acquisition of WH-questions in American Sign Language. In *The Signs of Language Revisited*, 401-413. Mahwah, NJ: Lawrence Erlbaum.

Lindeman, R.H., Merenda, P.F., and Gold, R.Z. 1980. Introduction to bivariate and multivariate analysis. Scott Foresman, Glenview, IL.

Listgarten, J. 2007. Analysis of sibling time series data: alignment and difference detection. *Doctoral dissertation*, University of Toronto.

Listgarten, J., Neal, R.M., Roweis, S.T., and Emili, A. 2004. Multiple alignment of continuous time series. In *Advances in neural information processing systems*, 817-824.

López-Colino, F., and Colás, J. 2011. The Synthesis of LSE classifiers: from representation to evaluation. *J. UCS*, 17(3), 399-425.

Lu, P. 2013. Data-driven synthesis of animations of spatially inflected American Sign

Language verbs using human data. *Doctoral Dissertation, Computer Science*, The Graduate Center, City University of New York.

Lu, P., and Huenerfauth, M. 2009. Accessible motion-capture glove calibration protocol for recording sign language data from deaf subjects. In Proc of the 11th Int'l ACM SIGACCESS Conf on Computers and Accessibility (ASSETS), 83-90. ACM.

Lu, P., and Huenerfauth, M. 2010. Collecting a motion-capture corpus of American Sign Language for data-driven generation research. In Proc of the 1st Workshop on Speech and Language Processing for Assistive Technologies (HLT-NAACL), Los Angeles, CA, USA.

Lu, P., and Huenerfauth, M. 2011. Synthesizing American Sign Language spatially inflected verbs from motion-capture data. *2nd Int'l Workshop on Sign Language Translation and Avatar Technology (SLTAT)*, Dundee, Scotland.

Lu, P., and Huenerfauth, M. 2014. Collecting and evaluating the CUNY ASL corpus for research on American Sign Language animation. Computer Speech & Language, 28(3), 812-831.

Lu, P., and Huenerfauth, M. 2012. Learning a vector-based model of American Sign Language inflecting verbs from motion-capture data. In *Proc of the 2nd Workshop on Speech and Language Processing for Assistive Technologies (HLT-NAACL),* Montreal, Canada.

Lu, P., and Kacorri, H. 2012. Effect of presenting video as a baseline during an American Sign Language animation user study. In *Proc of the 14th Int'l ACM SIGACCESS Conf on Computers and Accessibility (ASSETS),* 183-190. ACM.

Lumley, T. 2009. leaps: regression subset selection. R package version 2.9.

Mana, N., and Pianesi, F. 2006. HMM-based synthesis of emotional facial expressions during speech in synthetic talking heads. In *Proc of the 8th Int'l Conf on Multimodal Interfaces*, 380-387. ACM.

Marschark, M., Pelz, J., Convertino, C., Sapere, P., Arndt, M.E., and Seewagen, R. 2005. Classroom interpreting and visual information processing in mainstream education for deaf students: Live or Memorex®? *Am Educ Res J*, 42, 727–762.

McDonnell, R., Jörg, S., Mchugh, J., Newell, F., and O'Sullivan, C. 2008. Evaluating the emotional content of human motions on real and virtual characters. In *Proc of the 5th Symp on Applied Perception in Graphics and Visualization (APGV)*, 67-74 ACM.

McDonald, J., Wolfe, R., Moncrief, R., and Baowidan, S. 2014. Analysis for synthesis: Investigating corpora for supporting the automatic generation of role shift. In *Proc of the 6th Workshop on the Representation and Processing of Sign Languages (LREC)*, Reykjavik, Iceland.

Mitchell, R., Young, T., Bachleda, B., and Karchmer, M. 2006. How many people use ASL in the United States? Why estimates need updating. *Sign Lang Studies*, 6(3), 306-335.

Mlakar, I., and Rojc, M. 2011. Towards ECA's animation of expressive complex behaviour. In analysis of verbal and nonverbal communication and enactment. *The Processing Issues*, 185-198. Springer Berlin Heidelberg.

Moemedi, K.A. 2010. Rendering an avatar from sign writing notation for sign language animation. *Doctoral dissertation*, University of the Western Cape.

Morimoto, K., Kurokawa, T., Kentarou, U., Teruyo, K., Kazushi, T., Katsuo, N., and Tamotsu, F. 2003. Design of an agent to represent Japanese Sign Language for hearing-impaired people in stomach x-ray inspection. *Asia Design Conference*.

Morimoto, K., Kurokawa, T., and Kawamura, S. 2006. Improvements and evaluations in sign animation used as instructions for stomach x-ray examination. In *Proc of the 10th Int'l Conf on Computers Helping People with Special Needs (ICCHP)*, 607-614. Springer Berlin Heidelberg.

Muir, L.J., and Richardson, I.E. 2005. Perception of sign language and its application to visual communications for deaf people. *J Deaf Stud Deaf Educ*, 10(4), 390-401.

Neidle, C., Kegl, J., MacLaughlin, D., Bahan, B., and Lee, R. 2000. The Syntax of American Sign Language: Functional categories and hierarchical structure. Cambridge, MA: MIT Press.

Neidle, C., Liu, J., Liu, B., Peng, X., Vogler, C., and Metaxas, D. 2014. Computer-based tracking, analysis, and visualization of linguistically significant nonmanual events in American Sign Language (ASL). In *Proc of the 6th Int'l Workshop on the Representation and Processing of Sign Languages (LREC)*, Reykjavik, Iceland.

Oates, T., Firoiu, L., and Cohen, P.R. 1999. Clustering time series with hidden markov models and dynamic time warping. In *Proc of the IJCAI-99 workshop on neural, symbolic and reinforcement learning methods for sequence learning*, 17-21.

Och, F.J., and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1), 19-51.

Ow, S.H. 2009. User evaluation of an electronic Malaysian Sign Language dictionary: e-Sign dictionary. In *Proc of Computer and Information Science*, 34-52.

Ouhyoung, M., Lin, H. S., Wu, Y. T., Cheng, Y. S., and Seifert, D. 2012. Unconventional approaches for facial animation and tracking, In *SIGGRAPH Asia*, 24.

Pandzic, I.S., and Forchheimer, R. 2002. MPEG-4 facial animation. The standard, implementation and applications. Chichester, England: John Wiley & Sons.

Pejsa, T., and Pandzic, I. S. 2009. Architecture of an animation system for human characters. In *Proc 10th Int'l Conf on Telecom (ConTEL),*171-176. IEEE.

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2 (11), 559–572.

Poritz, A.B. 1988. Hidden markov models: A guided tour. In *Proc of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7–13. Morgan Kaufmann, 1988.

Pražák, M., McDonnell, R., and O'Sullivan, C. 2010. Perceptual evaluation of human animation timewarping. In *ACM SIGGRAPH ASIA 2010 Sketches*, 31-32.

Price, P., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C. 1991. The use of prosody in syntactic disambiguation. *J of the Acoustical Society of America*, 90(6), 2956-2970.

Prillwitz, S., Leven, R., Zienert, H., Hanke, T, and Henning, J. 1989. HamNoSys: v2.0:

Hamburg Notation System for Sign Languages: an introductory guide. Signum, Hamburg.

Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372-422.

Rayner, K. 2009. Eye movements in reading: models and data. *Journal of Eye Movement Research*, 2(5), 1-10.

Reilly, J., and Anderson, D. 2002. FACES: The acquisition of non-manual morphology in ASL.

Rosen, L.D., Whaling, K., Carrier, L.M., Cheever, N.A., and Rokkum, J. 2013. The media and technology usage and attributes scale: An empirical investigation. *Comput Human Behav,* 29(6), 2501-2511.

Russell, M., Kavanaugh, M., Masters, J., Higgins, J., and Hoffmann, T. 2009. Computer-based signing accommodations: comparing a recorded human with an avatar. *Journal of Applied Testing Technology*, 10(3), 21.

Sakoe, H. and Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition, *Acoustics, Speech and Signal Processing*, 26(1), 43–49.

Salvucci, D.D., and Anderson, J.R. 2001. Automated eye-movement protocol analysis. *Hum-Comput Interact*, 16(1), 39-86.

San-Segundo Hernández, R. 2010. Improvement and expansion of a system for translating text to sign language - Representation of the signs (chapter 5).Retrieved on August 10, 2014 from vhg.cmp.uea.ac.uk/tech/hamnosys/An%20intro%20 to%20eSignEditor%20and%20HNS.pdf.

San-Segundo, R., Montero, J.M., Córdoba, R., Sama, V., Fernández, F., D'Haro, L.F., López-Ludeña, V., Sánchez, D., and García, A. 2012. Design, development and field evaluation of a Spanish into sign language translation system. *Pattern Analysis and Applications*, 15(2), 203-224.

San-Segundo, R., Barra, R., Córdoba, R., D'Haro, L.F., Fernández, F., Ferreiros, J., Lucas, J.M., Macías-Guarasa, J., Montero, J.M., and Pardo, J.M. 2008. Speech to sign language translation system for Spanish. *Speech Commun*. 50(11), 1009-1020.

Sims, E., and Silverglate, D. 2002. Interactive 3D characters for web-based learning and accessibility. In *ACM SIGGRAPH Conference Abstracts and Applications*, 314-314. ACM.

Sims, E. 2000. Virtual communicator characters. *ACM SIGGRAPH Computer Graphics*, 34(2), 44.

Sheikh, Y.A., Khan, E.A., and Kanade, T. 2007. Mode-seeking by medoidshifts. In *IEEE 11th Int'l Conf on Computer Vision (ICCV)*, 1-8. IEEE.

Schmidt, C., Koller, O., Ney, H., Hoyoux, T., and Piater, J. 2013. Enhancing gloss-based corpora with facial features using active appearance models. In *3rd Int'l Symp on Sign Language Translation and Avatar Technology (SLTAT)*. Chicago, IL, USA.

Schnepp, J. and Shiver, B. 2011. Improving deaf accessibility in remote usability testing. In *Proc. of 13th Int'l ACM SIGACCESS Conf on Computers and Accessibility (ASSETS)*, Dundee, Scotland.

Schnepp, J., Wolfe, R., and McDonald, J. 2010. Synthetic corpora: A synergy of linguistics and computer animation. *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (LREC)*, Valetta, Malta.

Schnepp, J., Wolfe, R., Shiver, B., McDonald, J., and Toro, J. 2011. SignQUOTE: a remote testing facility for eliciting signed qualitative feedback. *2nd Int'l Workshop on Sign Language Translation & Avatar Technology*, Dundee, UK.

Schnepp, J., Wolfe, R., McDonald, J.C., and Toro, J. 2012. Combining emotion and facial nonmanual signals in synthesized American Sign Language. In *Proc of 14th Int'l ACM SIGACCESS Conf on Computers and Accessibility (ASSETS)*, 249-250.

Schuirmann, D.J. 1987. A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. *J Pharmacokin Biopharm*, 15, 657–680.

Smith, R., and Nolan, B. 2013a. Emotional facial expressions in synthesized sign language avatars: A manual evaluation. *ITB Journal*, 4(24).

Thangali, A., Nash, J. P., Sclaroff, S., and Neidle, C. 2011. Exploiting phonological constraints for handshape inference in ASL video. In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition (CVPR),* 521-528. IEEE.

Tran, J.J., Johnson, T.W., Kim, J., Rodriguez, R., Yin, S., Riskin, E.A., Ladner, R.E., and Wobbrock, J.O. 2010. A web-based user survey for evaluating power saving strategies for deaf users of mobileASL. In *Proc of the 12th Int'l ACM SIGACCESS Conf on Computers and Accessibility (ASSETS)*, 115-122, ACM.

Traxler, C.B. 2000. The Stanford Achievement Test: National norming and performance standards for deaf and hard-of-hearing students. *Journal of deaf studies and deaf education*, 5(4), 337-348.

Valstar, M.F., Mehu, M., Jiang, B., Pantic, M., and Scherer, K. 2012. Meta-analysis of the first facial expression recognition challenge. *, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(4), 966-979.

Vcom3D. 2015. Homepage. http://www.vcom3d.com/.

Velichko, V. M. and Zagoruyko, N. G. 1970. Automatic recognition of 200 words, *Int'l Journal of Man-Machine Studies*, 2(3), 223–234.

Verlinden, M., Tijsseling, C., and Frowein, H. 2001. Sign language on the WWW. In *Proc of 18th Int'l Symposium on Human Factors in Telecommunication*.

Visage Technologies. 2014. Visage Technologies Face Tracking. Retrieved on February 4, 2014 from http://www.visagetechnologies.com/products/visagesd k/facetrack/.

Wallraven, C., Breidt, M., Cunningham, D.W., and Bülthoff, H.H. 2008. Evaluating perceptual realism of animated facial expressions. *ACM Trans. Appl. Percept. (TAP)*, 4(4), 4.

Watanabe K, Matsuda T, Nishioka T, and Namatame M. 2011. Eye Gaze during observation of static faces in deaf people. *PLoS ONE,* 6(2), e16919.

Watson, K.L. 2010. WH-questions in American Sign Language: Contributions of non-manual marking to structure and meaning. *Master Thesis, Linguistics*, Purdue U.

Weast, T.P. 2008. Questions in American Sign Language: A quantitative analysis of raised and lowered eyebrows. *ProQuest*.

Weise, T., Bouaziz, S., Li, H., and Pauly, M. 2011. Realtime performance-based facial animation. *ACM Transactions on Graphics (TOG),* 30(4), 77.

Wells, J. 1997. SAMPA computer readable phonetic alphabet. In *Gibbon, R.W., and Moore, R. (Eds.), Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Berlin, Germany.

Wilbur, R.B. 2003. Modality and the structure of language: Sign languages versus signed systems. *Oxford handbook of deaf studies, language, and education*, 332-346.

Wolfe, R., Cook, P., McDonald, J.C., and Schnepp, J. 2011. Linguistics as structure in computer animation: Toward a more effective synthesis of brow motion in American Sign Language. *Sign Language & Linguistics*, 14(1), 179-199.

World Federation of the Deaf. 2014. Retrieved on July 5, 2014 from http://wfdeaf.org/

human-rights/crpd/sign-language.

Yang, O., Morimoto, K., and Kuwahara, N. 2014. Evaluation of Chinese Sign Language animation for mammography inspection of hearing-impaired people. In *Advanced Applied Informatics (IIAI)*, 831-836. IEEE.

Yu, X., Huang, J., Zhang, S., Yan, W., and Metaxas, D.N. 2013. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *Proc of the*

IEEE International Conference on Computer Vision (ICCV), 1944-1951. IEEE.

Zhang, Y., Ji, Q., Zhu, Z., and Yi, B. 2008. Dynamic facial expression analysis and synthesis with MPEG-4 facial animation parameters. *Circuits and Systems for Video Technology*, 18(10), 1383-1396.

Zhou, F., and De la Torre, F. 2009. Canonical time warping for alignment of human behavior. *In NIPS*, 2286-2294.