Human-Centered Optimization of Mobile Sign Language Video Communication

Jessica Julie Tran

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:
Eve A. Riskin, Co-chair
Richard E. Ladner, Co-chair
Jacob O. Wobbrock, Co-chair

Program Authorized to Offer Degree:

Department of Electrical Engineering

University of Washington

Abstract

User-Centered Optimization of Mobile Sign Language Video Communication

Jessica Julie Tran

Co-Chairs of the Supervisory Committee:

Professor Eve A. Riskin
Department of Electrical Engineering

Professor Richard E. Ladner
Department of Computer Science and Engineering

Associate Professor Jacob O. Wobbrock
The Information School

The proliferation of mobile devices is greater than ever; however, bandwidth and battery life have not grown accordingly to support mainstream use of mobile video communication. This dissertation contributes to the continued effort of making mobile sign language communication more accessible and affordable to deaf and hard-of-hearing people. I am optimizing the lower limits at which mobile sign language can be transmitted to reduce bandwidth and battery life, while maintaining intelligibility. This work presents the *Human Signal Intelligibility Mode*l (HSIM) to address the lack of uniformity in the way that intelligibility and comprehension are operationalized for evaluation. The HSIM influenced the design of four web studies: (1) investigating perceived intelligibility of sign language video transmitted at various low frame rates and low bit rates below the current recommended video transmission standards as prescribed in the International Telecommunication Standardization Sector (ITU-T) Q.26/16 (at least 25 fps and 100 kbps); (2) investigating the relationship between response-time and video intelligibility, which led to the creation of the *Intelligibility Response-Time Method*; (3)

evaluating perceived video quality of different power saving algorithms utilizing qualities unique to sign language; and (4) comparing objective video quality measures to subjective responses. Results revealed an “intelligibility ceiling effect” for video transmission rates, where increasing the frame rate above 10 fps and bit rate above 60 kbps did not improve perceived video intelligibility. These findings suggest that the recommended ITU-T sign language transmission rates can be relaxed while still providing intelligible American Sign Language (ASL) video, thereby reducing bandwidth and network load.

I conducted a laboratory study in which pairs of fluent ASL signers held free-form conversations over an experimental smartphone app transmitting video at frame rates and bit rates well below the ITU-T standard, to investigate how fluent ASL signers adapt to the lower video transmission rates. Participants were successful in holding intelligible conversations across all frame rates, even though they perceived the lower quality of video transmitted at 5 fps/ 25 kbps. Also, video transmitted at 10 fps/50 kbps or higher was not found to significantly improve video intelligibility, which corroborates with web study findings. Finally, I conducted a field study observing everyday use of an experimental smartphone app transmitting video at rates below the ITU-T standard. The field study revealed that gathering in-the-moment information using mobile video chat was preferred over texting because of the faster response-time.

Taken together, the findings from this dissertation support the recommendation that intelligible mobile sign language conversations can occur at video transmission rates far below the ITU-T standard to optimize resources consumption, video intelligibility, and user preferences.

The thesis of my dissertation is:

Mobile sign language video transmitted at frame rates and bit rates below recommended standards (ITU-T vs. 10 fps/50 kbps), which saves bandwidth and battery life by about 30 minutes, is still intelligible and can facilitate real-time mobile video communication.

Table of Contents

List of Figures ... ix
List of Tables ... xiii
Glossary ... xv
Chapter 1 Introduction ... 1
1.1 Contributions ... 2
Chapter 2 Background and Related Work ... 4
2.1 Video Compression ... 4
2.2 Evolution of Mobile Networks ... 6
2.3 Commercial Mobile Video Applications ... 7
2.4 MobileASL Project ... 8
2.4.1 MobileASL for Windows Mobile 6.1 ... 8
2.4.2 Power Saving Algorithms ... 10
2.5 Related Work on Video Quality Evaluations ... 12
2.5.1 Sign Language Comprehension ... 12
2.5.2 Objective Video Quality Measures ... 13
2.5.3 Subjective Video Quality Measures ... 16
2.5.4 Cognitive Load ... 18
2.5.5 Sign Language Linguistic Research ... 19
2.5.6 Surveys for Deaf Participants ... 20
2.5.7 Bandwidth Requirements ... 21
Chapter 3 Human Signal Intelligibility Model (HSIM) ... 24
3.1 Existing Communication Models ... 24
3.1.1 Shannon's Theory of Communication ... 24
3.1.2 Berlos's SMCR Model of Communication ... 25
3.1.3 Barnlund's Transactional Model of Communication ... 27
3.2 Defining Intelligibility ... 28
3.3 HSIM Components ... 28
Chapter 4 Methodology for Creating Web Studies for Deaf People ... 33
4.1 American Sign Language Instructional Videos ... 33
4.2 HSIM Influence ... 34
4.2.1 Establishing Language Fluency ... 35
4.3 Video Stimuli ... 35
4.4 Encoding Videos ... 36
4.5 Survey Components ... 36
4.6 Logging Response-Time ... 38
Chapter 5 Web Study: Relationship between PSNR and Perceived Intelligibility 39
5.1 PSNR Calculations ... 40
5.2 Results ... 46
5.3 Discussion ... 53
Chapter 6 Web Study: Perceived Video Intelligibility ... 55
6.1 Results ... 56
6.2 Discussion ... 61
6.2.1 Frame Rate and Bit Rate ... 61

6.2.2 Comprehension Question Response-Time 62
6.2.3 Signing Speed 63
6.3 Summary 63
Chapter 7 Web Study: Response-Time and Mental Effort Relationship 64
7.1 Existing Workload Evaluations 65
7.2 Intelligibility Response-Time Method 66
7.2.1 Logging Response-Time 69
7.3 Results 69
7.3.1 Comprehension Question Accuracy 70
7.3.2 Perceived Intelligibility 71
7.4 Comprehension Question Response-Time 72
7.5 Correlation between Intelligibility and Response-Time 75
7.6 Discussion 76
7.7 Frame Rate and Bit Rate Recommendations 77
7.8 Summary 77
Chapter 8 Laboratory Study: Effects of Lower Video Transmission Rates 79
8.1 Technology Used 80
8.2 Participants 81
8.3 Study Design 82
8.3.1 Apparatus 82
8.3.2 Conversation Task 83
8.3.3 Subjective Measures 84
8.3.4 Objective Measures 85
8.4 Results 86
8.4.1 Perceived Intelligibility 86
8.4.2 Objective Measures 88
8.5 Exit Interviews 89
8.5.1 5 FPS Video Quality 90
8.5.2 Desire for Larger Screens 90
8.5.3 Adaptation Techniques 91
8.5.4 Comparisons to Commercial Video Apps 92
8.6 Battery Drain 93
8.6.1 Bandwidth Consumption 94
8.7 Discussion 94
8.7.1 Signing Adaptation Techniques 95
8.7.2 Willingness to Use Lower Video Quality 95
8.7.3 Technology Position Adjustments 96
8.7.4 Recommendations 97
8.8 Summary 97
Chapter 9 Field Study: Using the MobileASL Application in the Wild 99
9.1 Technology Used 100
9.2 Study Procedure 101
9.3 Results 102
9.3.1 Unobtrusive Logging 103
9.4 Experience Sampling 103

9.5 Interviews ... 104
9.5.1 Focus Group ... 105
9.6 Summary ... 106
Chapter 10 Power Saving Algorithms ... 108
10.1 Variable Frame Rate ... 108
10.2 Variable Spatial Resolution ... 110
10.3 Combination of Variable Frame Rate and Variable Spatial Resolution ... 112
10.4 Cell Phone Battery Power Study ... 113
10.5 Battery Consumption ... 114
10.6 Current Drain ... 116
10.7 CPU Usage ... 118
10.8 Summary of Battery Power Algorithms ... 121
10.9 IMSDroid Battery Experiment ... 121
10.9.1 Results ... 123
Chapter 11 Web Study: Perception of Power Saving Algorithms ... 126
11.1 Results ... 128
11.2 Discussion ... 132
11.3 Summary ... 133
Chapter 12 Conclusion and Future Work ... 134
12.1 Reflections and Insights ... 135
12.1.1 Distinguishing Video Intelligibility from Quality and Comprehension ... 136
12.1.2 Linguistically Accessible Studies ... 136
12.1.3 User-Centered Approach to Evaluating Video Intelligibility ... 137
12.2 Lessons Learned ... 138
12.3 Future Work ... 140
12.3.1 Context-Aware Video Quality Adaptation ... 140
12.3.2 Region-of-Interest Improvements ... 141
12.3.3 Mobile Video Communication in Emergency Situations ... 141
12.4 Summary of Contributions ... 142
12.5 Final Remarks ... 144
References ... 147
Appendix A-Press Coverage ... 155
Appendix B- Web study: Perceived Video Intelligibility ... 157
English sentences ... 157
Glossed ASL sentences and English sentences used in web-study ... 158
Appendix C- Web Study: Response-Time and Mental Effort Relationship ... 161
Survey Sentences and Questions Used ... 161
Appendix D- Field Study: Using MobileASL Application in the Wild ... 165
Pre-Deployment Questionnaire ... 165
Interview Questions ... 173
Unobtrusive Logging Items ... 174
Experience Sampling Triggers and Questions ... 176
Weekly Questionnaire ... 181

List of Figures

Figure 1: HTC TyTNII cell phone. 9

Figure 2: Block diagram of Shannon's communication system [92]. 25

Figure 3: Berlo's SMCR Model of Communication [15]. 26

Figure 4: Barnlund's Transactional Model of Communication [12]. 27

Figure 5: Block diagram of the Human Signal Intelligibility Model. Note that the components comprising signal intelligibility are a subset of signal comprehension, which is signal intelligibility plus the receiver's mind. 30

Figure 6: (a) Screen shot of one video from web survey evaluating intelligibility of sign language video displayed at 5 frames per second at 15 kilobits per second. (b) Example of question 1 shown in web survey. (c) Multiple choice comprehension question example. Each item was shown one at a time. 37

Figure 7: PSNR(dB) vs. Bit rate (kbps) for spatial resolutions displayed at 320×240 pixels. Higher PSNR means higher objective video quality. Whether it means higher subjective perception of quality is a topic of this research. 41

Figure 8: Screenshot of one 12-second video pair from the paired-comparison experiment. Respondents selected which video they preferred to watch. 43

Figure 9: Q1 was a 7-point Likert scale for the ease of comprehension. Q1 was shown after the video was removed from the screen. 45

Figure 10: Q2 asked a simple comprehension question pertaining to the video shown. Q2 was shown after Q1 was removed from the screen. 45

Figure 11: Percentage of People *vs.* Bit rate (kbps) who selected 320×240 instead of 192×144 spatial resolution in the paired-comparison experiment. Data are from 56 ASL signers and 39 Non-ASL signers. 47

Figure 12: Double *y*-axis plot of 7-point Likert scale. Negative PSNR values of spatial resolutions and bit rates. Lower Likert scores correspond to less difficulty and lower PSNR values correspond to higher video quality. Notice a negative PSNR crossover point occurs at 40 kbps. 50

Figure 13: Double *y*-axis plot of 7-point Likert scale. Positive PSNR values of spatial resolution and bit rate. Higher Likert scores correspond to more ease and higher PSNR values correspond to higher video quality. Notice a positive PSNR crossover point occurs at 40 kbps. 52

Figure 14: Plot of 7-point Likert ratings for participants' ease of understanding the video for each frame rate and bit rate averaged over all participants. Error bars represent ±1 standard error. 59

Figure 15: Double y-axis plot of a 7-point Likert scale rating participants' ease of understanding the video and mean response-time (seconds) for correctly answered comprehension questions for both frame rate (averaged over all four bit rates rates) and bit rate (averaged over all four frame rates). Higher Likert scores correspond to higher perceived intelligibility. 60

Figure 16: Flow diagram example of the Intelligibility Response-Time Method used for video intelligibility analysis. 65

Figure 17: Example of data inclusion using the Intelligibility Response-Time Method. Note a 7-point Likert scale is used, where 7 is very easy and 1 is very difficult for how easy the video was to understand. 68

Figure 18: Plot of the mean Likert scores (higher values are better) and mean response-times (in seconds) for comprehension questions for frame rate (averaged over bit rates rates). 73

Figure 19: Plot of the mean Likert scores (higher values are better) and mean response-times (in seconds) for comprehension questions for bit rate (averaged over frame rates). 74

Figure 20: Experimental setup with two participants separated by a board. A certified ASL interpreter was always present. 83

Figure 21: Count of conversational breakdowns and repair requests that occurred for each session when video was transmitted at 5 fps/25 kbps. 89

Figure 22: Estimated average battery life (in minutes) for sign language video transmitted on IMSDroid at each frame rate/ bit rate. 93

Figure 23: Four examples of how participants adjusted the phone position. (a) Original phone setup using a business card holder. (b) Phone propped up with a pen. (c) Increased height and viewing angle. (d) Increased height from table. 96

Figure 24: Example of an experience sampling question appearing in MobileASL. 101

Figure 25: Total number of calls made each day excluding day 1 and 2. 103

Figure 26: Distribution of responses for the experience sampling question, “Which best describes where you are right now?” 104

Figure 27: Depiction of variable frame rate algorithm. The frame rate decreases when the signer is not-signing, resulting in “choppy” video quality [23] 109

Figure 28: Downsampler Block Diagram. 110

Figure 29: Example 4×4 image and YUV 420 representation. 111

Figure 30: Depiction of variable spatial resolution algorithm. The not-signing frames are downsample to 1/4 the original spatial resolution and displayed at the same frame size, resulting in blurry video quality. 111

Figure 31: Example of a not-signing frame downsampled to 1/4 of original size, which produces a blurry video quality. 112

Figure 32: Example of a not-signing frame with default implementation of MobileASL. 112

Figure 33: The implementation of VFR and VSR during not-signing portions of a conversation. The resulting output is a combination of blurry and choppy video. 113

Figure 34: Average battery life (minutes) for each power saving algorithm. 115

Figure 35: Measured current drain (mA) vs. time (minutes) for each encoding algorithm. 116

Figure 36: Total CPU usage (percentage) vs. time (minutes). 119

Figure 37: Experimental setup where two Samsung Galaxy S3 phones are facing a computer screen with a video of a woman signing in ASL. 123

Figure 38: Average battery life for transmitting sign language video at each frame rate/bit rate. 124

Figure 39: Combinations of factors and levels within the web study. 126

Figure 40: Screen shot of ASL video interpretation of survey questions and 5-point Likert scale. 128

Figure 41: VFR×VSR Interaction for Q1. Note the Y-axis is the rank used by the nonparametric analysis procedure. Lower values indicate less perceived choppiness. 130

Figure 42: VFR×VSR Interaction for Q3. Note the Y-axis is the rank used by the nonparametric analysis procedure. Lower values indicate less perceived blurriness. 131

Figure 43: VFR*VSR Interaction for Q4. Note the Y-axis is the rank used by the nonparametric analysis procedure. Lower values indicate less perceived distraction due to blurriness. 132

List of Tables

Table 1: Mean Likert Scale responses (1-7) for difficulty of comprehending video quality. Note *lower* Likert scores correspond to *less* perceived difficulty. 50

Table 2: Mean Likert Scale responses (1-7) for ease of comprehending video quality. Note *higher* Likert scores correspond to *easier* perceived comprehension... 52

Table 3: Mean Likert score responses for ease of understanding video quality. Note higher Likert scores correspond to higher perceived intelligibility. 58

Table 4: Mean Likert score (higher values are better) and mean response-time (in seconds) for correctly answered comprehension questions for both frame rate (averaged over all four bit rates rates) and bit rate (averaged over all four frame rates). Bold values indicate highest mean Likert scores and fastest times to submit answer.. 60

Table 5: Percentage of correctly answered comprehension questions across frame rate and bit rate. Accuracy increases with bit rate but not with frame rate.. 71

Table 6: Mean Likert scores (higher values are better) and mean response-times (in seconds) for comprehension questions for frame rate (averaged over bit rates rates). Bold values indicate highest mean Likert scores and fastest times to submit answer.. 73

Table 7: Mean Likert scores (higher values are better) and mean response-times (in seconds) for comprehension questions for bit rate (averaged over frame rates). Bold values indicate highest mean Likert scores and fastest times to submit answer. 75

Table 8: Predicted Intelligibility Score compared to Mean Likert score for perceived video intelligibility using regression equation where lower response-time and higher Likert and Intelligibility scores are better. 76

Table 9: Count of the number of fingerspelled words and the average, max, min, and standard deviation of the number of characters signed per second. 88

Table 10: Average, min, max, and SD of the bit rate when varying the frame rate as captured by the network traces. 94

Table 11: Average current drain (mA) for each power saving algorithm.................................. 117

Table 12: Comparison of battery life (minutes) and current drain (mA) of each experiment.... 118

Table 13: CPU usage (percentage) for phone only, MobileASL application only, and total system. 120

Table 14: Mean values of applying VFR or VSR for Q1-Q4. 129

Table 15: Summary of MobileASL Project Studies with the Resulting Frame Rate and Bit Rate at which the Intelligibility Ceiling Effect and Diminishing Returns Occurred, respectively..... 146

Glossary

ACTIVITY ANALYSIS OF VIDEO: classification of video into different categories based on the activity recognized in the video

AMERICAN SIGN LANGUAGE (ASL): the primary sign language of Deaf people in the United States

BANDWIDTH: the data capacity of a communication channel, measured in bits per second

(bps) or kilobits per second (kbps)

CHROMINANCE: the color component of an image

FINGER SPELLING: sign language in which each individual letter is spelled

FRAME: a single video image

FRAMES PER SECOND (FPS): unit of measure of the frame rate of a video

FRAME RATE: the rate at which frames in a video are shown, measured in frames per second (fps)

H.264: the latest IEEE standard for video compression

HAND SHAPE: the position the hand is held while making a sign

HUMAN-SIGNAL INTELLIGIBILITY MODEL (HSIM): a new conceptual model that outlines the components comprising signal intelligibility and signal comprehension for the purpose of video intelligibility evaluations

INTELLIGIBILITY RESPONSE-TIME METHOD: a new method using response-time as an indicator of mental effort to further inform video intelligibility evaluations

INTERNATIONAL TELECOMMUNICATION STANDARDIZATION SECTOR (ITU-T): part of the International Telecommunication Union with specific responsibility to research and recommend standards and protocols relating to voice and data transmissions over landline and mobile networks

KILOBITS PER SECOND (KBPS): unit of measure of bandwidth

LUMINANCE: the brightest component of an image

MACROBLOCK: a 16 _ 16 square area of pixels

MOTION VECTOR: a vector applied to a macroblock indicating the portion of the reference frame it corresponds to

PEAK SIGNAL TO NOISE RATIO (PSNR): a measure of the quality of an image

QP: quantizer step size, a way to control macroblock quality

REAL-TIME: a processing speed fast enough so that there is no delay in the video

REGION OF INTEREST (ROI): an area of the frame that is specially encoded

REPAIR REQUEST: a request for repetition

SESSION INITIATION PROTOCOL (SIP): a signaling communications protocol used for controlling multimedia communication sessions such as voice and video calls over Internet Protocol networks

TELETYPEWRITER (TTY): a device that allows users to type messages in real-time over the phone lines

VARIABLE FRAME RATE (VFR): a frame rate that varies based on the activity in the video

VARIABLE SPATIAL RESOLUTION (VSR): a spatial resolution that varies based on the activity in the video

VOICE OVER INTERNET PROTOCOL (VOIP): technology that allows telephone calls to be made over computer networks

X264: an open source implementation of H.264

Acknowledgements

I am thankful for the support and guidance that I have received through my Ph.D. journey from family, friends, colleagues, and mentors.

First, I would like to thank my “dream team” of advisors: Professor Eve A. Riskin, Professor Richard E. Ladner, and Associate Professor Jacob O. Wobbrock, who have given me unconditional support, guidance, and financial stability during this journey. Eve has inspired me to be a strong woman in engineering. Her enthusiasm and encouragement have empowered me to be an independent researcher that always considers “the bigger picture” while welcoming outreach opportunities. Richard has taught me the importance of creating technology to empower people. Jake has influenced my EE+HCI identity, especially in conducting thoughtful user studies and teaching me more about statistics than any formal class. I am a better researcher and public speaker because Eve, Richard, and Jake pushed me to think critically and refine research ideas, while giving me the freedom to pursue independent work.

I am also thankful to my other committee members, Professor Howard Chizeck from EE and Professor Hannah Wiley from the Dance Program. Thank you, Hannah, for making ballet class fun and exciting each quarter!

My work would not have been possible without collaborations with MobileASL alumni: Anna Cavender, Neva Cherniavsky, Rahul Vanam, Jaehong Chon, and Tressa Johnson. Also, I must thank the 15 MobileASL undergraduates who have given me the privilege of mentoring them while working on this project. I would especially like to thank Rafael Rodriguez who worked with me for four years and brought thoughtful questions and ideas to our weekly meetings. I would also like to thank the first group of students I mentored: Joy Kim and Sherri Yin; and my final student: Ben Flowers, for their contributions to this project. Conducting MobileASL research would not have been possible without the support of Tobias Cullins, Coordinator of Deaf & Hard of Hearing Services, Rob Roth, and the numerous American Sign Language interpreters, especially Jennifer Austin and Jason Smith; and last but not least, all the study participants.

Thank you to Pat Johnson and John Davis III for scheduling meetings with my advisors. Also, thank you to Dan Feetham for thoughtful conversations both academically and Husky sports related. I have enjoyed talking Husky Football over the years with Pat, John, and Dan and appreciate their true understanding of bleeding purple and gold. GO HUSKIES!

Thank you to the Electrical Engineering Department staff: Stephen Graham, Erica Coleman, and Travis Sailing for answering all my EE-related questions. Thank you to Professor Radha Poovendran for continued interest and support in my research work. Also, thank you to the ISDL lab mates old and new especially: Pascal Clark, Nicole Nichols, Xing Li, Brian King, Elliot Saba, Scott Wisdom, and Thomas Powers for providing a supportive lab environment.

Numerous friends and colleagues contributed to thoughtful conversations that have sparked new ideas, brainstorming sessions, draft-readings, stress-relieving, and considering the “bigger picture.” In alphabetical order: Shiri Azenkot, Sina Bahram, Theresa Barker, Anke Brock, Jillian Buchan, Marc Carroll, Parmit Chilana, Charles Delahunt, Abigail Evans, Sarah Force, Jeff Huang, Alex Jensen, Shaun Kane, Allison Kang, Alex Marin, Julie Medero, Martez Mott, Katie O’Leary, Rupa Patel, Jennifer Perry, Jan Rey Pioquinto, Kyle Rector, Silvia Rodriguez Vázquez, Fay Shaw, Kelsey Shepard, Kristen Shinohara, Mona Yaptango, and many others!

Finally, I owe a great deal of my success and share this huge accomplishment with my parents, Simon and Julie Tran; brother, Justin Tran; and soon-to-be sister-in-law, Martha Chan. Thank you for your unconditional support and believing that I can accomplish anything that I set my mind to.

To my parents.

Chapter 1 Introduction

With over 1.9 billion smartphone users at the end of 2013, smartphones are rapidly changing the way people communicate and receive information [106]. The growth of smartphone users has led to video being the fastest growing contributor to mobile data traffic [106]. Streaming video providers like YouTube, Hulu, and Netflix contribute to mobile video traffic, consuming 51% of all network traffic. Mobile video telephony is also contributing to the acceleration of video data consumption with the numerous available mobile video chat applications (apps) like FaceTime and Google Hangouts. In 2010, Skype received 7 million downloads onto Apple's iPhone alone [93].

Often high fidelity video quality is top priority for mobile video telephony; however, it is usually at the cost of large bandwidth consumption. Apple's FaceTime app is widely known to provide high quality video over Wi-Fi with an average bandwidth consumption of 5MB of data per minute of conversation [32]. The high data rate cost of using FaceTime over limited data plans can quickly become expensive [19]. Other mobile video chat apps, like Skype, transmit video at lower dynamic transmission rates ranging from 40-450 kbps depending on network traffic [27]. However, wide variations in transmission rates by commercial mobile video applications place a heavy load on the total available network bandwidth, which may lead to packet loss, delay, and blurred video. Video intelligibility is often sacrificed when using commercial mobile video apps that rely on the available network bandwidth to regulate video transmission rates.

Deaf and hard-of-hearing people can benefit significantly from advancements in mobile video communication because they facilitate sign language communication. American Sign Language (ASL) is a visual language with its own grammar and syntax unique from any spoken language. Intelligible video content is required for successful sign language conversations; therefore the Telecommunication Standardization Sector (ITU-T) Q.26/16 recommends at least 25 frames per second (fps) and 100 kilobits per second (kbps) for sign language video transmission [89]. However, total network bandwidth is limited and network congestion can lead to unintelligible content due to delayed and dropped video. Most U.S. cellular networks no longer provide unlimited data plans and throttle network speeds to high data rate consumers [68]. The ITU-T recommendation does not account for the available total bandwidth of cellular networks or consider the lower bounds at which sign language video may be deemed intelligible. Often recommendations are based on evaluations of pre-recorded video and are not intended for real-time mobile video communication.

1.1 Contributions

This dissertation contributes to the continued effort of making mobile sign language communication more accessible and affordable to deaf and hard-of-hearing people. I am optimizing how much mobile sign language video transmission rates can be reduced to save resources (bandwidth and battery life) while maintaining intelligibility. This work includes the creation of the *Human Signal Intelligibility Model* (HSIM), a new conceptual model addressing the lack of uniformity signal intelligibility and signal comprehension have been operationalized for evaluation. The HSIM influences the design and execution of four web studies I conducted: (1) investigating perceived intelligibility of sign language video transmitted at various low frame rates and low bit rates (Chapter 6); (2) investigating the relationship between response-time and video intelligibility, which lead to the creation of the *Intelligibility Response-Time Method* (Chapter 7.2); (3) evaluating perceived video quality of different power saving algorithms that utilize qualities unique to sign language (Chapter 8, Chapter 10, and Chapter 11); and (4) comparing the effectiveness of objective video quality measures to subjective responses for video transmitted at low transmission rates (Chapter 5). Results revealed an intelligibility ceiling effect for video transmission rates, where increasing the frame rate above 10 fps and bit rate above 60 kbps did not improve perceived video intelligibility.

I conducted a laboratory study in which pairs of fluent ASL signers held free-form conversations over an experimental smartphone app transmitting video at frame rates of 5, 10, 15, and 30 fps and bit rates of 25, 50, 75, and 150 kbps, well below the ITU-T standard, to investigate how fluent ASL signers adapt to the lower video transmission rates (Chapter 8). Finally, I conducted a field study evaluating everyday use of a mobile video chat app, MobileASL, in the wild to observe how mobile video communication can improve everyday communication (Chapter 9). Findings from this dissertation support the recommendation that intelligible mobile sign language conversations can occur at video transmission rates far below the ITU-T standard to optimize resources consumption, video intelligibility, and user preferences. These findings also have the potential to influence global use of mobile video communication, especially across developing network infrastructures (Chapter 12).

Chapter 2 Background and Related Work

2.1 Video Compression

Successful real-time mobile video telephony requires little to no latency during transmission; therefore video compression must be applied to make video content manageable for network data transmissions. Video compression is the application of algorithms that convert video files into a format that takes fewer bits to represent the data. Compression can take on two forms: lossless and lossy. Lossless compression compresses data without losing any information, but at the expense of using more resources such as processing time and file storage space. Lossy video compression uses spatial correlation and temporal motion compensation to reduce redundancy in video data. The benefit of lossy video compression is reduced file size, but at the expense of video quality such as introducing visible or distracting artifacts that may impact video intelligibility.

H.264/MPEG-4 AVC is a standard for lossy video compression that is a commonly used format for recording, compressing, and decompressing video [43, 88]. H.264 is best known as the codec standards for Blu-ray Discs and different streaming internet sources like YouTube, Vimeo, iTunes store, and web software like Adobe Flash Player and Microsoft Silverlight. The MobileASL application, discussed in Chapter 2, uses x264 which is an open source version of H.264 [3].

H.264 is a block-based motion-compensation codec. Motion estimation is used to create motion vectors for intra- and inter- frame coding. Intra-frame coding only uses information contained in the current frame to process (no temporal processing). Inter-frame coding takes advantage of temporal redundancy between neighboring frames which allows for higher compression rates. The higher compression rates are achieved by the encoder dividing each frame into blocks of pixels, called macroblocks. The encoder uses a block matching algorithm which tries to find a closely matching block in the previous decoded and up-sampled frame. If a matching block cannot be found, then that block is intra-coded (I-block); otherwise the difference between the new block and the previous one is transformed, via Discrete Cosine Transform (DCT), and the resulting DCT coefficients are quantized. All of this information is losslessly compressed and sent to the decoder for video reconstruction.

The DCT separates an image into parts of different frequencies. The DCT has a strong "energy compaction" property where the signal information is concentrated in a few low-frequency coefficients and the high frequency components are quantized to zero. Tradeoffs in video quality can be made by varying the quantization parameter (QP) and frame rate. For instance, the QP may vary from frame to frame or may be fixed for the entire video. A low QP value requires more bits to encode than a high QP value, but the resulting video has higher quality. Conversely, a high QP value results in the DCT coefficients being quantized more heavily, which sets more coefficients to zero. With fewer coefficients to send and fewer bits per zero coefficients, fewer bits are used, which leads to reduced video quality. In addition to the QP, the frame rate can be varied. Typically, a lower frame rate with the same value of QP requires fewer bits to encode than a higher frame rate. For a fixed bit rate there is a trade-off between frame quality (typically objectively measured by Peak Signal-to-Noise Ratio), which is controlled by the QP parameter and frame rate. More frame per second means that the individual frames will have to be of lower quality to maintain the same bit rate.

2.2 Evolution of Mobile Networks

A new mobile generation wireless system is introduced in the United States approximately every ten years. The first 1G system, Nordic Mobile Telephone, was introduced in 1981 and was the first fully automatic cellular phone system transmitting data at 1200 bps. The next generation known as 2G, Global System for Mobile Communication (GSM), started rolling out in 1992 and became the de facto global standard for mobile communications, transmitting data at 14.4 kbps. 3G (EDGE and CDMA) started becoming available in 2011 and provided an upload data rate of 118.4 kbps and download data rate of 296 kbps. 3G was slow to be adopted globally due to some 3G networks not using the same radio frequencies as 2G; as a result, network providers needed to build new networks and license new frequencies to achieve higher data transmission rates. 4G Long Term Evolution (LTE) began appearing in 2012 and provides download data peak rates of 300 Mbps and upload peak rates of 75 Mbps. The quality of service aims for a data transfer latency of less than 5 ms. Today, major cellular phone companies like Sprint, T-Mobile, AT&T, and Verizon are expanding their 4G LTE networks to provide higher data speeds in more locations across the U.S.; however, consistent access to 4G LTE service is currently location-dependent.

Even though network providers are continually growing their data services, total network bandwidth is still limited. Many cellular phone companies no longer offer unlimited data plans and have switched to tiered data plans ranging from 2-4 GB per month depending on the data plan [10, 101, 105]. The average U.S. consumer uses 733 MB of data per month; however, those users generally check websites and email [38]. Smartphone users who stream music or video on their mobile devices can quickly use up their data allowance in a few hours. For instance, streaming music with average quality (160 Kbps) requires 1.2 MB per minute or 72 MB per hour; music streaming at 320 Kbps is equivalent to 2.4 MB per minute or 144 MB per hour; a Netflix video in standard definition can consume up to 0.7 GB per hour and a Netflix video in HD can consume 1 GB-2.8 GB per hour [53].

2.3 Commercial Mobile Video Applications

Commercial mobile video applications have evolved with the growing networks. Skype is a free voice-over-IP service that allows people to communicate through instant message, voice, and video on computers and mobile devices [94]. Skype video calls transmit video at high bit rates with mobile-to-mobile Skype calls transmit at 500 kbps and video calls between mobile phone and a computer transmit at 600 kbps [54]. Before 2013, Apple’s FaceTime mobile video chat application could only work over Wi-Fi networks. Once Apple devices supported iOS6 (initially released in September 2012), FaceTime began working on AT&T’s tiered data plans, at the data consumption rate of 3 Mb of data per minute, for tiered data plans only [1].

Video relay services allow deaf, hard-of-hearing, and speech-impaired people to communicate over video telephone with a hearing person in real-time via a sign language interpreter. Major VRS companies like Purple Communications, Inc. [82]; Sorenson VRS [96]; ConvoRelay [31]; and ZVRS [119] provide VRS apps for mobile devices. In compliance with the ITU-T standard, these applications attempt to transmit video at least 25 fps and 100 kbps or higher, which may lead to video delay or dropped video calls. VRS users tend to use video phones or computers with access to broadband connection to utilize interpreting services without the worry of dropped video calls.

All of these aforementioned commercial mobile video apps provide reasonable video quality for intelligible conversations at the expense of larger bandwidth consumption and more aggressive battery consumption than voice calls or texting. Those who use video chat or VRS consume network bandwidth more rapidly than average data users, which leads to increased cost for all mobile users. Cellular phone companies do not currently offset the extra cost of mobile video communication used by deaf and hard-of-hearing people. Instead, network providers begin throttling down network speeds after 2 GB of data usage per month [68]. This dissertation contributes to the MobileASL project's goal of providing deaf and hard-of-hearing people equal access to mobile video communication without needing to pay more for services.

2.4 MobileASL Project

MobileASL is a video compression project at the University of Washington and Cornell University that began in 2005 with the goal of making wireless cell phone communication through sign language a reality in the United States [69]. One of the goals was to transmit real-time, two way video using the GSM EDGE network that has 296 kbps download and 118 kbps upload speeds. In 2008, a major milestone was met with a working prototype of MobileASL, an experimental smartphone application that provides two-way, real-time sign language video at very low bandwidth (30 kilobits per second at 8-12 frames per second) [17].

2.4.1 MobileASL for Windows Mobile 6.1

MobileASL was developed using the Windows Mobile 6.1 platform for the HTC TyTNII cellular phone [24]. This phone, shown in Figure 1, was selected because it has a front-facing camera and screen which can prop itself up on a table at an angle during conversations. The phone weighs 6.7 oz; has a 400 MHz processor; and 1350 mAH battery life. The MobileASL app uses the open source x264 implementation of the H.264 standard [3] with ARMv6 SIMD instruction set [7] and a NAT-enabled protocol [24]. The app uses a peer-to-peer networking application that allows video transmission on both Wi-Fi and AT&T 3G/4G cellular networks.

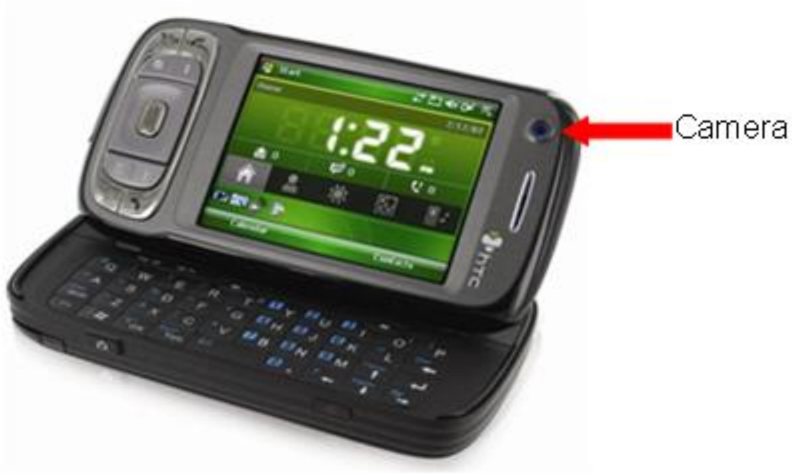


Figure 1: HTC TyTNII cell phone.

Since intended users of MobileASL are deaf or hard-of-hearing, characteristics unique to sign language were used to reduce the total amount of data needed for transmission. For example, an algorithm called Region-of-Interest (ROI) encoding, that differentiates between skin pixels and background, was implemented [22]. When MobileASL transmits video, more bits are devoted to skin pixels, such as a person's hands and face, making those regions appear clearer than the background.

Intelligible ASL video is more important than ASL video quality because people can perceive changes in video quality before content intelligibility is compromised. Cavender *et al.* [17] conducted a focus group in 2006 investigating intelligibility of sign language video constrained by mobile phone technology. In the focus group they explored the need and/or desire for mobile video phones and addressed potential challenges with using such technology. Some notable findings were: participants desired the device to have the ability to be propped up for two-hand communication; the software interface needs to have an easy and intuitive display; and the ability to make video calls between different video software. Cavender *et al.* [17] also conducted a laboratory study evaluating video intelligibility at two frame rates (10 and 15 fps), three bit rates (15, 20, and 25 kbps), and three region-of-interest (ROI) encoding levels (0, -6, and -12 ROI) where participants viewed pre-recorded videos and were asked to subjectively rate perceived intelligibility. (The ROI was an approximation of where the signers' face and hands were located.) They discovered a frame rate preference at 10 fps for viewing ASL video at a fixed bit rate of 25 kbps.

Masry and Hemami [66] evaluated subjective video quality perception of non-ASL streaming video content transmitted at 10, 15, and 30 fps and six bit rates (40, 100, 200, 300, 600, and 800 Kbps). Respondents viewed fifteen 30-second video clips consisting of low, medium, and high motion sequences. After each video, respondents rated video quality on a slider ranging from 0 (worst) to 100 (best). The researchers found that respondents favored video shown at 15 fps over 10 fps when shown at a fixed bit rate of 800 Kbps.

Findings from this dissertation along with those from prior work [17] demonstrate that there is a threshold over which increasing the frame rate and bit rate at which video is transmitted does not significantly improve video intelligibility. The research presented in this dissertation builds upon Cavender *et al.*'s [17] findings by developing a web study based on their study design and more rigorously investigates intelligibility of sign language video, which includes discovering how much video quality can be reduced before sign language intelligibility is compromised, a goal not approached by prior MobileASL related research.

2.4.2 Power Saving Algorithms

Methods to save battery power while using MobileASL were important for wide adoption of mobile video communication. Cherniavsky *et al.* [23] used qualities unique to sign language, specifically identifying when someone was signing or not-signing, to vary the frame rate at which real-time sign language video was transmitted on MobileASL. This technique, called variable frame rate (VFR), uses the sum of pixel differencing between frames to identify when a person is signing or not-signing. The frame rate drops from 15 fps to 1 fps when a person is classified as not-signing. This frame rate reduction produces perceived choppy video quality. In a laboratory study, Cherniavsky *et al.* [22] investigated the effects of VFR on battery life consumption and intelligibility of real-time sign language communication. They found that battery life was increased by 47% which resulted in a 68 minute gain of additional talk time. Chapter 10 describes two new power saving algorithms, which build upon Cherniavsky's work, while successfully increasing battery life and reducing the perceived negative effects introduced by each algorithm.

Cherniavsky *et al*. also conducted a laboratory study in which pairs of fluent ASL signers were video recorded signing over MobileASL with the VFR algorithm implemented. They found that applying VFR led to degradation in video quality which resulted in respondents having to guess more frequently during conversations. Overall, participants expressed that having the VFR algorithm applied during their conversations did not deter their potential adoption of MobileASL for mainstream mobile video communication. With these and other findings demonstrating the potential lower transmission limits in which intelligible mobile sign language video communication can occur, Chapter 8 describes a new laboratory study investigating intelligibility of real-time sign language video transmitted at frame rates and bit rates below the recommended ITU-T standard for the purposes of saving resources on a new experimental smartphone application.

2.5 Related Work on Video Quality Evaluations

The effects of frame rate and bit rate reductions on objective video quality have been widely researched for sign language learning and comprehension; evaluating subjective video quality; creating video quality measures; and evaluating video intelligibility. However, unlike the present work, none of this prior work has been intended for facilitating real-time mobile sign language conversations or considering the bandwidth needed to support such communication. The work here in fills this gap by identifying the lower limits of intelligible mobile sign language communication.

2.5.1 Sign Language Comprehension

Sign language learning is more nuanced than holding sign language conversations. The former requires linguistic accuracy to correctly convey signs, while the latter does not require absolute accuracy of signs in order for the overall message to be understood in a conversation. The effect of frame rate reduction on sign language learning has been extensively researched [20, 52, 57, 99] but not so for holding sign language conversations. Johnson and Caird [57] investigated whether perceptual ASL learning was affected by video transmitted at 1, 5, 15, and 30 fps. In a discrimination task, participants made a *yes-no* decision about whether the displayed sign and the English word shown matched. They found that frame rates as low as 1 fps and 5 fps were sufficient for novice ASL learners to recognize learned ASL gestures. Although this work suggests frame rates as low as 1 and 5 fps can support sign language recognition, it does not evaluate conversational sign language, which this dissertation investigates.

Hooper *et al*. [52] define comprehension as the ability for respondents to accurately retell stories verbatim. They investigated the impact on ASL comprehension when ASL video was presented at 6, 12, and 18 fps and displayed at 240×180, 320×240, and 480×360 pixels at 700 kbps. Hooper *et al.* found video display size did not affect comprehension, but varying frame rates did. Students performed better after viewing video at 12 fps than at 6 fps, and at 18 fps than at 6 fps; however, there was no significant difference in performance between 18 fps *vs.* 12 fps.

Sperling *et al.* [99] defines intelligibility as the ability to correctly recognize signs. Under this operationalization, they investigated ASL video intelligibility transmitted at 10, 15, and 30 fps displayed at 96×64, 48×32, and 24×16 pixels, while applying a grayscale image transformation. They found that common isolated ASL signs shown at 96×64 pixels at 15 fps and 30 fps did not have a noticeable difference in intelligibility, but lowering the frame rate to 10 fps did. While prior work showed that lower frame rates can impact isolated sign recognition, these results may not hold true for mobile sign language video conversations because the spatial resolutions investigated were small and may have influenced respondents ability to recognize signs shown at 10 fps. Also, Sperling *et al.'s* work was conducted in 1985 where the video compression algorithms were not as efficient as today; therefore more visual artifacts may have been introduced in the stimuli used. This dissertation goes beyond sign recognition and investigates video intelligibility to support two-way real-time mobile sign language conversations.

2.5.2 Objective Video Quality Measures

Measuring subjective video quality is time consuming, content-specific, and requires many subjects to produce generalizable findings. By contrast, the peak signal-to-noise ratio (PSNR) is commonly used in video compression to measure *objective* video quality after lossy compression [112]. However, the PSNR has been shown to not always accurately represent humans' subjective judgments about video quality [37, 73, 100, 102, 109]. A short video sequence with a few frames that are heavily distorted may reduce the PSNR; however, the overall video sequence was understood. Numerous researchers have attempted to map PSNR to subjective responses by creating new objective video quality perception metrics [77, 108, 114, 118]; however, these objective measures have been content-dependent and not evaluated on sign language videos.

Numerous metrics and algorithms have been created in an attempt to bridge the gap between PSNR and subjective video quality. However, the PSNR has not been shown to accurately represent subjective video quality [37, 73, 100, 114] and a standard subjective metric has not yet been adopted.

Feghali *et al*. [37] created a subjective quality model that takes into account encoding parameters (quantization error and frame rate) and motion speed of video during calculation of their new subjective quality metric. They used Pearson's correlation *r*, as a measure of how well their subjective model matches subjective video quality, where values closer to 1.0 indicate a stronger positive linear relationship. They were able to achieve, on average (across five videos with different motion levels) an $r = .93$ when comparing the assessed subjective quality to their new objective quality metric. For high motion video, such as a football game, the assessed subjective quality compared to the PSNR resulted in $r = .57$, while the new quality metric resulted in $r = .95$; however, a smaller difference in *r* was found for slow motion video. Nemethova *et al*. [18] created a different rule-based algorithm that adapts the PSNR curve to mean opinion scores (MOS) by scaling, clipping, and smoothing the PSNR results. The new MOS adapted from the PSNR curve was compared to the assessed subjective MOS whose results demonstrated an average $r = .89$. Both algorithms demonstrated success in increasing the accuracy of measuring subjective video quality; however, both researchers recognize that their algorithms are content-dependent and have higher performance with fast motion video, of which sign language video would be considered one type.

Related research by Ciaramello and Hemami [25] developed an objective measure for ASL video intelligibility, which relies on region-of-interest (ROI) encoding of different areas of video. They encoded ASL video at three different bit rates (20, 45, and 80 kbps) and five ROI settings that vary the allocation of bits to the background and the signer in the foreground during video encoding. This varying resulted in video with the background appearing blurrier than the ASL signer depending on the bit rate and ROI combinations. In a paired comparison experiment with 12 respondents, they found that at higher bit rates, respondents preferred the background and signer in the foreground to be equal in blurriness; however, at lower encoding bit rates, respondents preferred the signer to be less blurry than the background. The experiments and studies I present in this dissertation are different than prior work since I evaluate both subjective video quality *and* video intelligibility while others only evaluated subjective video quality. This work will reveal how user preferences and video intelligibility may change with varying spatial resolutions and bit rates; a person may not prefer the video quality but still finds the content intelligible.

A related research topic is investigating tolerance of image artifacts when lowering bit rates and image resolutions. Bae *et al*. [11] conducted a 7-respondent experiment that assessed absolute perceived quality and relative perceived quality of compressed images at different bit rates. In the absolute perceived quality assessment, respondents were shown uncompressed images and asked to score the image on a 5-point Likert scale on video quality (excellent to poor). Next, compressed sets of images were presented to the participant, who selected the one image that they preferred the most. Bae *et al*. discovered that as bit rates decrease, respondents prefer to maintain image quality by selecting a lower image resolution. Respondents were willing to accept an increase in image distortion (compression noise) introduced by the coding algorithms when shown an image at smaller spatial resolutions.

Video intelligibility is most important for successful mobile sign language video communication; therefore, objective video evaluations are not the most appropriate way to characterize video quality. Ciaramello and Hemami [26] recognized that sign language video needs to be evaluated in terms of subjective intelligibility. They created a computational intelligibility model (CIM) for ASL called CIM-ASL, which measures the perceptual distortions of video regions deemed important for conveying information, specifically the hands, face, and torso of a signer. The CIM-ASL model has been shown to have statistically significant improvements over PSNR when estimating distortions in the CIM-ASL-defined signing region. However, the CIM-ASL model relies on video quality perception with the assumption that greater video quality in the signing region leads to higher intelligibility. The Human Signal Intelligibility Model (HSIM) (described in Chapter 3) is different from prior models because the HSIM defines the components comprising signal intelligibility for the purpose of evaluation.

2.5.3 Subjective Video Quality Measures

Part of this work aims to discover whether frame rate or bit rate has more impact on ASL video intelligibility. A subjective experiment, conducted by Yadavalli *et al.*[117], evaluated frame rate preferences passively viewed for low, medium, and high motion sequences displayed at 352×240 pixels; three frame rates (10, 15, and 30 fps); and three bit rates (100, 200, and 300 kbps). Viewers preferred video at 15 fps when bit rate was averaged across all bit rates and video sequences, which suggests that 15 fps represents a compromise rate between frame and motion quality. At 300 kbps, respondents preferred video at 30 fps, suggesting that motion quality is more important once adequate frame quality is achieved. Similar to Yadavalli *et al.*'s work, my research aims to determine whether ASL video is made more intelligible by increasing the frame rate once frame quality (determined by bit rate) is adequate. But unlike this prior work, respondents are required to actively watch and understand ASL video content.

Subjective Evaluations of Auditory Signals

Frequently, video quality assessment is based on objective or subjective evaluations of the video itself, and the mental effort required of viewers is overlooked. Part of my research also utilizes response-time to a multiple choice comprehension question as an indicator of mental effort along with self-reported perception of video intelligibility and comprehension question accuracy to evaluate video intelligibility, a combination of analyses not previously used. Chapter 7 describes the Intelligibility Response-Time Method (IRTM), which draws a relationship between mental effort and response-time to multiple choice comprehension questions as an additional measure for video intelligibility evaluations.

The IRTM is based on both speech communication and cognitive load evaluations. Drawing from these existing evaluations, the IRTM uses response-time to a multiple choice comprehension question as an indicator of mental effort in video intelligibility evaluations. This is similar to response-time used in both auditory signal and cognitive load evaluations.

The assessment of speech communication is well established with the ISO 9921 standard defining speech intelligibility "as a measure of effectiveness of understanding speech" [56].

Subjective and objective assessment are defined, the former based on the use of speakers and listeners and the latter based on physical parameters of the transmission channel. In these evaluations, high speech intelligibility is defined as the number of speech items recognized correctly such as correctly answered phonemes, words, and sentences.

Researchers have also conducted speech intelligibility evaluations by using response-time on subjective tests as a more accurate measurement [81, 91, 116]. For example, Gatehouse and Gordon evaluated amplification used in hearing aids by measuring the ease of listening using auditory response-times to speech stimuli of single words and sentences [41]. Other researchers have demonstrated correlations between ease of listening and response-time on verification tests [47, 81]. Specifically, ease of listening increases as response-time to stimuli decreases. As part of this dissertation, Chapter 7 presents research drawing a parallel between response-time and video intelligibility, which to my knowledge has not been explored.

2.5.4 Cognitive Load

In psychology, cognitive load is “a multidimensional construct representing the load that performing a specific task places on the learner’s cognitive system” [78] and can be measured by assessing mental load, mental effort, and/or performance. Mental effort is defined as the cognitive capacity allocated to the demands placed by the task; this is considered to reflect the actual cognitive load [78]. Mental effort is measured while participants are working on the task and can be captured in performance. Performance measures can be defined as learner’s achievements, such as the number of correct test items, number of errors, and time on task. In my laboratory study presented in Chapter 7, mental effort is defined as performance in the form of comprehension question response-time. The details of how response-time is measured is in the description of the Intelligibility Response-Time Method (Chapter 7, Section 7.2).

Lebeter and Saunders [61] evaluated the effects of time compression (i.e. speedup in time) on the comprehension of natural and synthetic speech using response-time. They found that people responded more quickly to comprehension questions spoken with natural speech than with synthetic speech. Chapter 7 presents research investigating if faster response-time to video comprehension suggests less mental effort, which in turn could correspond to higher video intelligibility.

2.5.5 Sign Language Linguistic Research

Linguistic research has shown that ASL is not a visual code for English [63]; ASL has a distinct, unrelated grammar and lexicon that has developed over time. There are conversational similarities between ASL and English, such as multiple people “holding the floor” at once [29] and feedback through back-channeling [30] (the latter refers to one person acknowledging understanding to the other, which could take the form of a muttered “uh-huh” in English or a head nod in ASL). The rate of finger spelling is normally several letters per second with skilled users approaching a rate of 10 letters per second [16].

Previous research has found that hand and face movements are key linguistic features of ASL that contribute to the intelligibility of a message [103]. Peripheral low-resolution vision is a key component in the perception of movement. Muir and Richardson [71] explored the eye movement patterns of Deaf people as they viewed sign language in video and then applied their findings to the design of video communication systems. They found that a Deaf viewer’s focus is placed on the facial region of a signer in order to pick up the small detailed movements in the signer's facial expression and lip shapes. This region is of interest because it conveys linguistic information to the receiver. Part of this dissertation explores the human perception of video quality when different power saving algorithms are applied during not-signing sections of a conversation (Chapter 11).

2.5.6 Surveys for Deaf Participants

Instructions in both English text and ASL videos have not, to my knowledge, been used in web-based user surveys intended for Deaf participants. Previous studies [4, 50] have been conducted to examine electronic communication among the Deaf population, but the medium in which researchers primarily chose to gather data from Deaf participants was based on English text.

Hogg *et al.* [50] researched the use of communication technology, gathering data from Deaf participants through an online English text-based survey. In the analysis of the respondents' responses, Hogg *et al.* recognized the limitations of their survey due to a "high proportion" of participants that did not complete the survey. Hogg *et al.* suggested that the participants' variance in reading levels may have contributed to the incomplete surveys. They further suggested that an ASL version of their survey might have produced better results.

A study conducted by Akamatsu *et al.*[4] used a text-based survey to collect data for their investigation of texting between Deaf high school students and their hearing parents. When the researchers reviewed the survey results, they recognized that the textual surveys were not linguistically accessible, and determined that interviews conducted in sign language were necessary to ensure complete and accurate responses from their deaf participants. Chapter 4 presents the methodology used to create linguistically accessible surveys by presenting instructions in both English text and ASL videos.

2.5.7 Bandwidth Requirements

Consideration of the bandwidth requirements for transmission of sign language has been ongoing since the early 1990s. Sperling [98] investigated the ability for deaf people to transcribe ASL and finger spelling from reduced television displays at bandwidths of 86 kHz, 21 kHz, 4.4 kHz, and 1.1 kHz. He wanted to address 1) what are the bandwidth requirements for ASL communication by video telephone; and 2) to what extent could such a video telephone use existing telephone channels to communicate ASL and finger spelling. The data rate for a telephone channel is 4 kHz or 33.6 kbps. Intelligibility was found to drop to 90% at 21 kHz and to 10% at 4.4 kHz. Finger spelling was found to be more sensitive to bandwidth reduction. Sperling discovered that most subjects could interpret ASL sentences with little loss at a bandwidth of 21 kHz. Expert signers received sentences at 40-50% correct at 4.4 kHz. This required bandwidth is four times greater for auditory speech on a telephone bandwidth of 3 kHz.

Foulds [39] introduced a method to temporally compress sign language animation on the order of 5:1 by separating kinematic or biomechanical bandwidth necessary to represent continuous movements in sign language before choppy video is perceived. His work evaluated the kinematic bandwidth needed for stick figure animations of individual signs. Nine participants evaluated twenty signs in isolation that were transmitted at 6 fps and 30 fps. Participants were asked to write the English word equivalent of the sign viewed. He found that the kinematics of human movement can found intelligible at 6 fps; however, a limitation of this work was that evaluations were performed on individual signs, which are not representative of real-time conversations.

Pearson [80] attempted to maintain smooth and intelligible video motion while reducing the frame rate. He used a frame repeating technique which felt unnatural to the signers and led to recommending frame rates of 15-30 fps for intelligible sign language. Sosnowski and Hsing [97] evaluated moving images, finding that reducing the frame rate from 30 to 15 fps only produced slightly less intelligible video; however, video displayed below 15 fps resulted in intelligibility dropping dramatically. Harkins *et al*. [44] compared the outline of signers to a videotaped control, which consisted of the video transmitted at the original recording rate, and found that video shown below 10 fps resulted in poor intelligibility. Ultimately, these prior works suggest that frame rates between 15-30 fps are the recommended rates at which video should be transmitted to maintain intelligibility. My dissertation will demonstrate that intelligible sign language conversations can occur below 15 fps.

Manoranjan and Robinson [65] investigated a method to reduce bandwidth consumption by transmitting binary sketches of cartoon signers. They implemented their video processing technique on a computer that simulated the bandwidth used over telephone lines. In a laboratory study with two total participants, participant 1 signed a sentence and participant 2 wrote down what he viewed. Participants evaluated four picture sizes of video displayed at 80×60, 160×120, 120×160, and 320×240 pixels/frame with video transmitted at 8 fps. The computer simulated transmission rates at 33.5 kbps for phone lines and 100 Mbps for the LAN data rate. Participants were unable to complete the task at 320×240 pixels/frame because of the low number of bits allocated per pixel. At such a low frame rate, participants preferred to view the binary sketches of the signer at the 80×60 pixels/frame resolution. A major limitation of this prior work was the small sample size of 2 total participants, which made results hard to generalize to mobile video communication.

Chapter 8 describes a major component of this work, a laboratory study investigating how fluent ASL signers adapt to lower video transmission rates; and the goal is to identify a lower threshold at which intelligible conversations could be held. Prior work has evaluated isolated words, animations of signs, and finger spelling. Although all this early work has demonstrated that there are limitations to the temporal reduction of frame rate, these results may not directly translate to real-time mobile video conversations. Video intelligibility may still persist, even if people perceive lower video quality. This dissertation will demonstrate that mobile sign language video transmitted at frame rates and bit rates below recommended standards saves bandwidth and battery life, is still intelligible and can facilitate real-time mobile video communication.

Chapter 3 Human Signal Intelligibility Model (HSIM)

In evaluating mobile sign language video intelligibility, I discovered a lack of uniformity in the way that “signal intelligibility” and “signal comprehension” are operationalized in human-centered evaluations. Often, intelligibility and comprehension are loosely defined and used interchangeably in evaluations of video quality. Some researchers focused on measuring signal intelligibility with the assumption that if one finds the signal intelligible, then comprehension of content follows [8, 45, 48, 52, 75]. As part of this dissertation, I present the *Human Signal Intelligibility Model* (HSIM), a new conceptual model informing video intelligibility evaluations and disentangling intelligibility from comprehension.

3.1 Existing Communication Models

Before introducing the components comprising the HSIM, I first discuss three existing conceptual models used to explain the human communication process: Shannon’s Theory of Communication [92]; Berlo’s Source-Message-Channel-Receiver model [15]; and Barnlund’s transactional model of communication [12]. Shannon’s Theory of Communication originates from information theory, while Berlo’s and Barnlund’s model of communication originates from communication theory. This section will also address the limitations of existing communication models and how intelligibility is defined, which led to the creation of the HSIM, described below.

3.1.1 Shannon’s Theory of Communication

In his famous work, Shannon [92] created a simple abstraction for communication called the *channel*, consisting of a sender (the information source), a transmission medium with noise and distortion, and a receiver (Figure 2). As this model stands, one could argue that objective metrics could be used to measure video quality and high quality scores may imply intelligible content. However, I argue that there are more components to intelligibility and comprehensibility of a video signal and using objective measures alone is not enough. The environment in which video is recorded and displayed as well as the humans sending and receiving video also need to be considered.

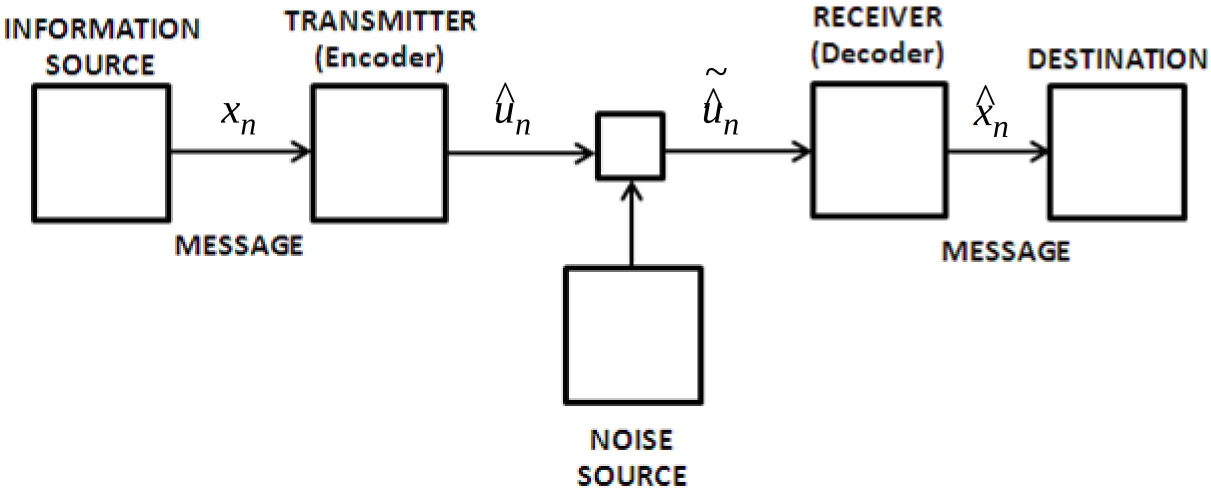


Figure 2: Block diagram of Shannon's communication system [92]**.**

Shannon's channel model only focuses on the communication channel itself without considering the surrounding environment or properties of a human sender and receiver.

3.1.2 Berlos's SMCR Model of Communication

Existing communication models [12, 15] attempting to distinguish intelligibility from comprehension are poorly defined. Berlo viewed communication as a coordination or synchronization process to allow people to deal with the environment in which they live [15]. He created the source, message, channel, receiver (SMCR) model of communication, as shown in Figure 3, to represent an exchange of ideas that may hold influence and authority with one's culture.

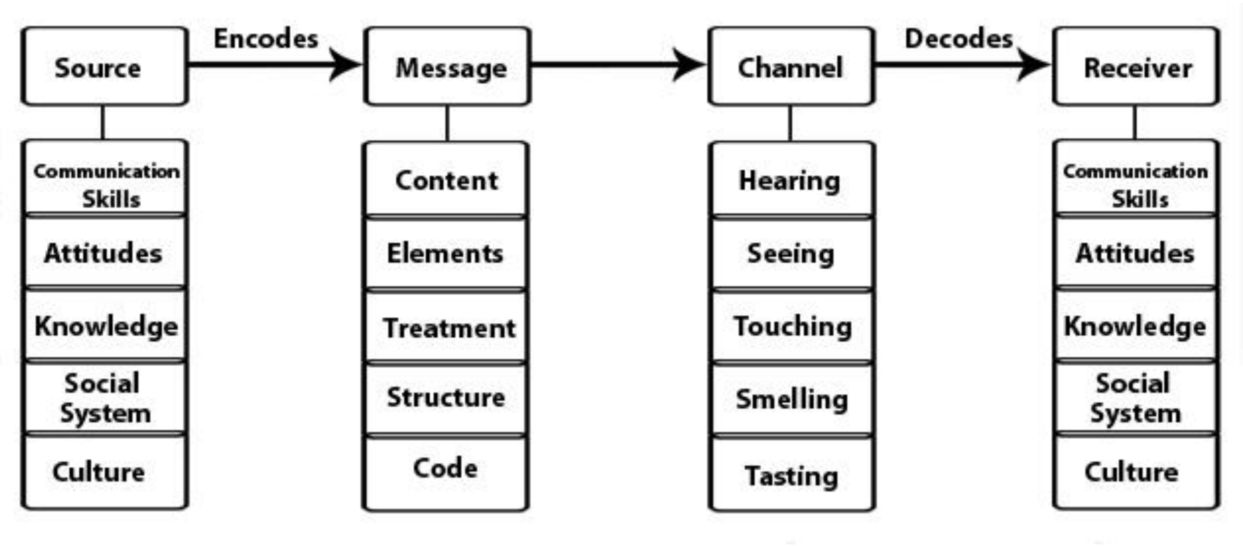


Figure 3: Berlo's SMCR Model of Communication [15]**.**

The SMCR model consists of the source, which includes the sender's communication skills, attitudes, knowledge, social system, and culture. The message is the physical product of the sender. The channel represents how the information is transmitted to the receiver's senses. Finally, the intended person of the message is the receiver with his own communication skills, attitudes, knowledge, social system, and culture. The SMCR model relies on the response of the receiver to determine if the message is successfully transmitted.

The SMCR model also has many limitations when used to evaluate intelligibility of mobile sign language communication. First, both the source and receiver list culture as a component to account for. Culture could be classified as a component of the human sending and receiving information, which has no direct impact on video transmission. Second, the channel components consist of the human senses, which are not representative of data being transmitted across mobile devices. While this model attempts to model human communication with twenty different components, the SMCR model does not clearly identify which elements produce intelligible communication.

3.1.3 Barnlund’s Transactional Model of Communication

Barnlund [12] proposed a Transactional Model of Communication with seven communication postulates suggesting individuals are simultaneously engaging in the sending and receiving of messages, as shown in Figure 4.

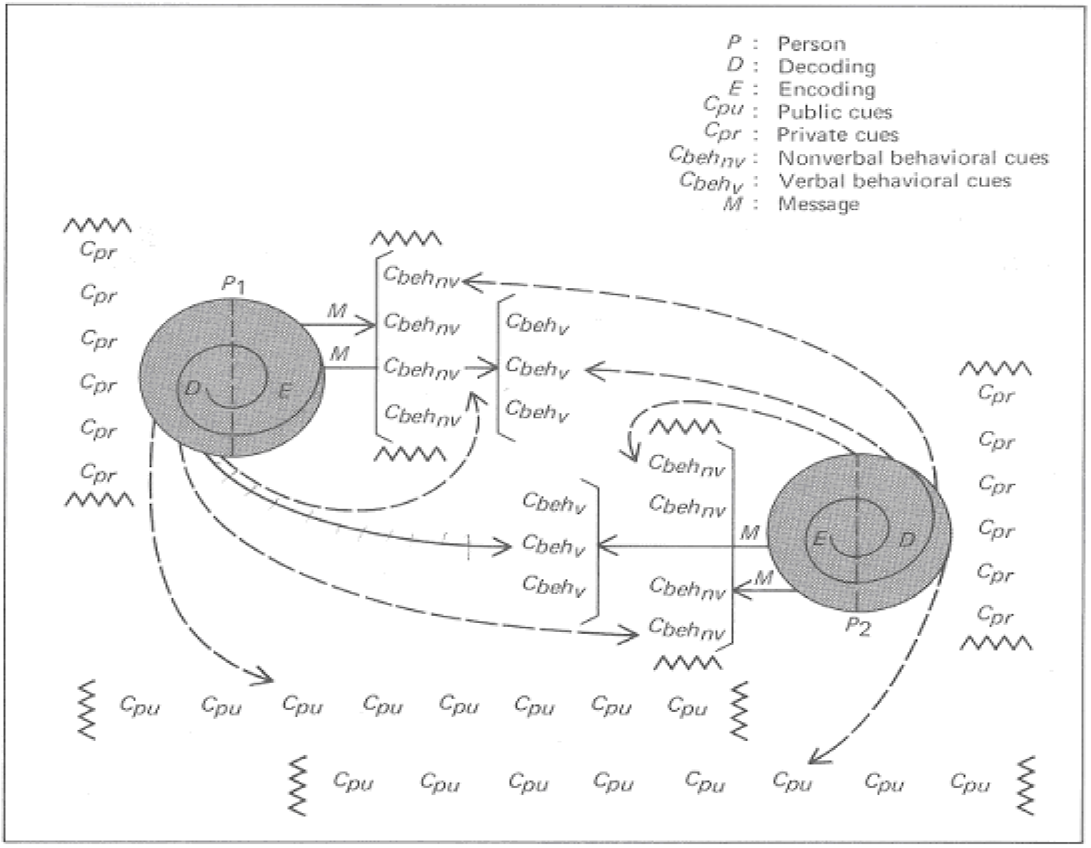


Figure 4: Barnlund’s Transactional Model of Communication [12]**.**

The Transactional Model of Communication states that giving and receiving messages is reciprocal; therefore, both the sender and receiver are responsible for the effectiveness of the communication. This model also divides communication into intrapersonal, which consists of encoding and decoding messages within one’s self, and interpersonal, which is encoding and decoding messages with another. There are seven communication postulates [12]: (1) communication describes the evolution of meaning; (2) communication is dynamic; (3)

communication is continuous; (4) communication is circular; (5) communication is unrepeatable; (6) communication is complex; and (7) communication is irreversible. Ultimately, this model emphasizes that people need to build a shared meaning of message. While this model focuses on how information is transferred and the relationship of the message between the sender and receiver, it does not attempt to distinguish intelligibility from comprehension. Also, this model does not consider the medium in which communication occurs and how it affects communication overall, which is included in the HSIM.

3.2 Defining Intelligibility

Signal intelligibility and signal comprehension need to be differentiated for the purpose of evaluating the lower limits at which intelligible sign language video can be transmitted. Intelligibility is defined as the *capability* of a signal to be understood [67], including how well the signal was articulated, captured, transmitted, received, and perceived by the receiver, including the environmental conditions affecting these steps. Comprehension relies on signal intelligibility *and* the human receiver having the prerequisite knowledge to understand the information. Both intelligibility and comprehension are human-centered concepts, unlike objective video quality measures such as peak signal-to-noise ratio (PSNR).These insights lead to the creation of the HSIM, described next.

3.3 HSIM Components

I present the HSIM to address the lack of uniformity in the way that signal intelligibility and signal comprehension have been operationalized, especially in contrast to objective video quality measures. This model distinguishes subjective video intelligibility from objective video quality and video comprehension, which are three usefully distinct and separable things.

The HSIM (1) extends Shannon's theory of communication [92] to include the human and environmental influences on signal intelligibility and signal comprehension, and (2) identifies the components that make up the *intelligibility* of a communication signal, while separating those from the *comprehension* of a communication signal. Signal intelligibility and signal comprehension are separable concepts because an intelligible signal does not necessarily lead to comprehension if the receiver lacks the requisite knowledge for understanding.

The capability of a signal (*e.g.*, video) to be comprehended is different than whether a signal is *actually* comprehended in any given instance, and this capability is the intelligibility of a signal. In the case of sign language video, intelligibility is affected by the human articulation of the signal; the environment affecting that articulation; the channel capturing, transmitting, receiving, and portraying that signal (the items in Shannon's model); the human perception of that signal; and the environment affecting that perception all affect intelligibility. Figure 5 shows a block diagram illustrating the components comprising intelligibility within the HSIM.

Whether or not the signal is *actually* understood involves all of the components comprising intelligibility *and* one additional component: the knowledge of the human receiver being adequate to understand the information, that is, to make sense of it. Because whether or not the signal is understood by the receiver is a part of the signal's ability to be *comprehended*, the receiver's mind is included in the components comprising comprehension in Figure 5. The knowledge of the human sender is irrelevant to comprehension by the receiver. For example, the sender could be a robot articulating ASL signs, but having no knowledge of ASL. The HSIM's definition of signal intelligibility and signal comprehension builds upon Koul's definition of speech signal quality. Koul [58] defines intelligibility of a speech signal as the individual's ability to recognize phonemes and words presented in isolation. Comprehension is defined as the listener's ability to process the linguistic message as a whole.

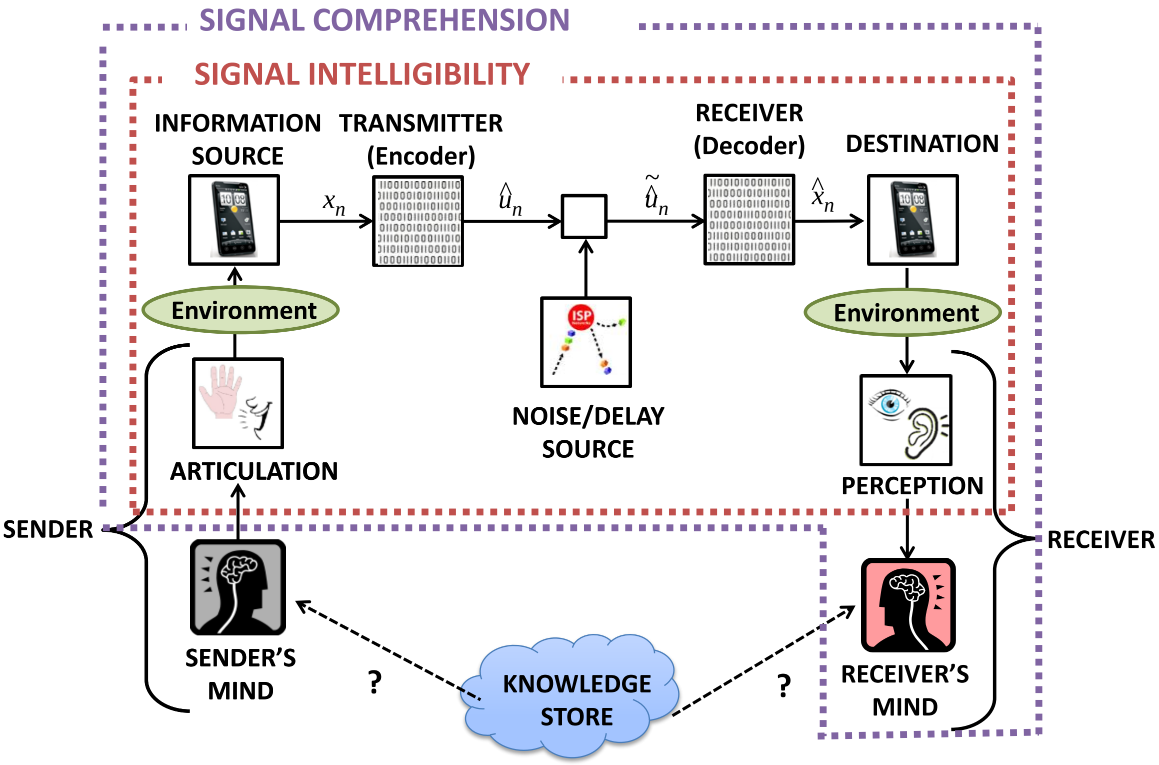


Figure 5: Block diagram of the Human Signal Intelligibility Model. Note that the components comprising signal intelligibility are a subset of signal comprehension, which is signal intelligibility plus the receiver's mind.

The HSIM goes beyond Koul to include environmental influences in which a signal is transmitted and received. Lighting is an example of an environmental factor that may influence signal intelligibility. For instance, viewing sign language video on a mobile device outside on a sunny day would make the screen appear dark. This environmental factor would clearly affect the ability for the video to be perceived by the receiver, compromising its intelligibility. (By contrast, the video's objective quality (PSNR) would be unaffected by sunny outdoor conditions.) Recognizing that the environment can influence signal intelligibility is why the environment is included in the HSIM.

The HSIM also explicitly separates the sender into two parts, the sender's mind and the sender's articulation. Similarly, the HSIM separates the receiver into two parts, the receiver's mind and the receiver's perception. The sender's articulation impacts intelligibility and comprehension because for sign language video, the quality with which information is conveyed influences the receiver's ability to understand the content. For example, a fluent ASL signer could have a motor impairment that would limit his ability to sign clearly. The physical limitation impacts the sender's signal articulation, which impacts the intelligibility of that signal to the receiver.

The receiver's perception also influences his or her ability to process information. For instance, the sender could sign perfectly clear ASL, but if the receiver has low vision, the signal would be unintelligible to that receiver. However, since the sign language video was clearly signed, it may be intelligible to other receivers. Moreover, measuring perception alone is not sufficient to infer intelligibility. Perceiving a change in video quality does not necessarily reflect the understandability of content. These and other examples illustrate the importance of recognizing human factors and environmental influences on signal intelligibility and signal comprehension. Intelligibility, then, is inherently a *contextualized* concept, unlike objective signal quality as measured by PSNR.

The HSIM reveals an important fact about signal intelligibility: it cannot be measured directly, as the ability to be comprehended cannot be easily separated from the actual comprehension of a signal. Fortunately, intelligibility can be inferred by measuring signal comprehension in the presence of fully capable receivers' minds with more than adequate linguistic knowledge to understand the signals they receive. Such minds remove any chance that a lack of knowledge affects comprehension, leaving only intelligibility to explain any comprehension difficulties.

One may wonder why signal perception is not used as a measure of signal intelligibility. Perception is defined as the ability to see, hear, or become aware of a change. Therefore, measuring awareness of changes in video quality alone is not sufficient to infer intelligibility. Using a just-noticeable difference evaluation [111] would not be appropriate because difference in video quality will be more evident at lower transmission rates before a signal becomes unintelligible.

The HSIM informs the web study designs in this dissertation, which presents research on evaluating the extreme lower transmission rate limits at which mobile sign language video can be transmitted before intelligibility is compromised. Owing to the need to ensure all receivers' minds are fully capable of comprehension, participants were screened for ASL fluency. Thereafter, differences in comprehension can be attributed to differences in intelligibility and not knowledge.

Chapter 4 Methodology for Creating Web Studies for Deaf People

There are two opposing conceptualizations of deafness, each with a unique impact on the design of a survey and the way in which it is received by Deaf participants. The first defines deafness as a pathological condition, while the second views deafness as a social identifier. The pathological model focuses on people's audiological status and considers deafness a medical condition requiring treatment. This perspective classifies people with hearing loss as "disabled" or "handicapped," and is marked by negative stereotypes and prejudice [34, 72]. Under this paradigm, deafness is perceived as the dominant quality of a group of people who share a "condition."

The social model, in contrast, holds that Deaf people are disabled more by their interactions with hearing people than by the physical condition that determines their perception of sounds. This view recognizes the linguistic [63, 64] and sociological [79, 83] research that has identified ASL as a unique language distinct from English, and Deaf Culture as a legitimate culture distinct from the mainstream.

Given the historical dominance of the pathological view of deafness [59], designing web studies that demonstrated respect for the language and culture of Deaf people was deemed of paramount importance. Taking into consideration both the values identified as defining characteristics of Deaf Culture, and the recorded experiences of deaf individuals who do not identify themselves as members of that culture, I identify two issues requiring explicit attention: linguistic accessibility, and respect for the autonomy and intelligence of the Deaf individual.

4.1 American Sign Language Instructional Videos

Ensuring the accessibility of an online survey is paramount to its success. Three factors were taken into consideration with regard to the accessibility of the web studies: (1) the intended audience of Deaf signers; (2) linguistic research determining the grammar and lexicon of ASL distinct from that of English [20, 25]; and (3) the value Deaf Culture places on both linguistic accessibility and self-determination [20]. For the web studies presented in this dissertation, I include an alternative to textual English by incorporating *ASL instructional videos,* to both increase accessibility and demonstrate my respect both for the individual participants and for Deaf Culture. Creating bilingual surveys widened the audience to include both ASL signers and those who prefer to communicate visually (potential MobileASL users) but who are not fluent in ASL (example: late-deafened individuals.)

Neither words nor signs have absolute equivalents in other spoken languages. What makes ASL/English interpretation possible is that both languages have the capacity to express identical meanings. The process of interpreting the surveys in ASL began with analyzing the text for explicit and implicit meaning, English-based discourse patterns, and cultural influences. A certified ASL interpreter was consulted to interpret the instructions with equivalent meaning while utilizing ASL-based discourse patterns and cultural influences.

4.2 HSIM Influence

The HSIM, presented in Chapter 3, informs the design of the web studies created to evaluate how much mobile sign language video transmission rates such as frame rate, bit rate, and spatial resolutions can be reduced before intelligibility is compromised. A mobile web survey was considered, but at the time of survey development, there was too much variability across mobile devices and mobile web browsers, which could not be controlled as an unwanted influence. The HSIM outlines that all receivers' minds need to be fully capable of sign language comprehension; therefore, sign language fluency must be established. Thereafter, I can attribute virtually all differences in comprehension to differences in intelligibility and not language fluency.

4.2.1 Establishing Language Fluency

Each web survey began by asking participants to self-report their fluency in ASL. Demographic questions were presented at the end of the survey to further identify language fluency. Examples of questions asked include: “Are you a native ASL signer?”; “From whom did you learn ASL?”; and “How many years have you signed ASL?” Instructions to the web study were provided in both ASL and English. ASL interpretations of the English text instructions were shown side-by-side throughout the web survey to increase accessibility. A professional ASL interpreter was consulted before filming.

4.3 Video Stimuli

Users of mobile sign language video communication are limited by the front-facing camera angle and confined signing space. Since the web survey would display pre-recorded video on a computer screen, the videos used in the survey simulated the 45 degree angle and signing space that would typically be displayed on small mobile devices.

A male native ASL signer/consultant signed 16 short ASL sentences that included various amounts of finger spelling and descriptive lexicons. The ASL signer sat in front of a solid dark blue background and was asked to sign all signs within the allowable signing space. Video length ranged from 15-30 seconds. For the web study described in Chapter 6, the ASL signer was asked to sign slowly. In the web study in Chapter 7, the ASL signer was asked to sign at a normal comfortable signing speed within the allowable signing space. The sentences used for each study are listed in Appendices B and C.

Technology Used for Recoding Video

When using prerecorded video, it is important to use videos that are representative of what is viewed on a mobile phone, especially when it comes to the signing space and angle at which video is displayed. For the web study described in Chapter 6, an Acer Iconic tablet running Android Honeycomb 3.2.1 was used to record the stimuli video. At the time of video recording, the front-facing camera of smartphones, like Sprint's EVO phone, only recorded compressed video in 3GP file format. At the time, recording video from a smartphone was not an option due to added video compression. A tablet was selected to record the videos because it simulated the allowable signing space and display angle. For the web study described in Chapter 7, a Google Nexus phone at 30 fps at 640×480 spatial resolution was used to recreate the angle and confined signing space imposed by the phone during mobile video conversations. In post-processing, these videos were further downsampled to 320×240 spatial resolution to further simulate the screen size on smartphones.

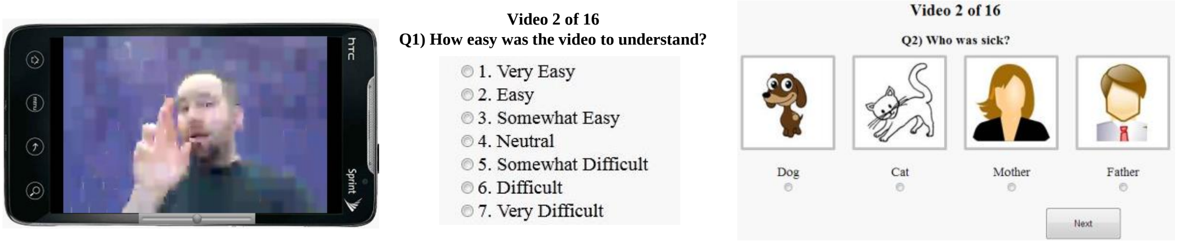
4.4 Encoding Videos

The YUV videos recorded were encoded to specified frame rates and bit rates used in each web study using the open source H.264 encoder [88]. The encoded videos were converted to MPEG-4 using a publicly available converter [14] that does not contribute additional artifacts. The web survey displayed the videos using Apple's QuickTime media player [6] since no additional artifacts were contributed by this player.

4.5 Survey Components

Each survey consisted of three parts and respondents were instructed to complete the survey over the course of one session. Part 1 had two practice videos to allow familiarization with the survey layout. Part 2 was the survey evaluating intelligibility of 16 different videos shown in a single-stimulus experiment. The same layout was used in part 1 and part 2 of the survey to reduce mental load imposed by the survey structure. Part 3 contained demographic questions to further identify language fluency.

All videos were displayed at 320×240 pixels in the middle of the web page. A picture of the Sprint EVO phone was placed behind each video simulating the mobile video appearance, as shown in Figure 6(a). Each video was shown once, *without* the option to repeat or enlarge, and then removed from the screen and replaced by two questions shown one at a time. Figure 6(b) is an example of question 1, which asked respondents to rate their agreement on a 7-point Likert scale with "How easy was the video to understand?" The 7-point Likert scale was shown in descending vertical order from *very easy* to *very difficult*. Figure 6(c) is an example of a comprehension question pertaining to the video shown. A four-option multiple choice question appeared with corresponding images. Participants were prompted to answer each question "as quickly as possible."



Video 2 of 16

Q1) How easy was the video to understand?

○ 1. Very Easy
○ 2. Easy
○ 3. Somewhat Easy
○ 4. Neutral
○ 5. Somewhat Difficult
○ 6. Difficult
○ 7. Very Difficult

Video 2 of 16

Q2) Who was sick?

Dog	Cat	Mother	Father
○	○	○	○

(a) (b) (c)

Figure 6: (a) Screen shot of one video from web survey evaluating intelligibility of sign language video displayed at 5 frames per second at 15 kilobits per second. (b) Example of question 1 shown in web survey. (c) Multiple choice comprehension question example. Each item was shown one at a time.

4.6 Logging Response-Time

Using the HSIM's definition of intelligibility, Chapter 7 introduces the Intelligibility Response-Time Method (IRTM), which uses response-time as a measure of mental effort in evaluating video intelligibility. The time to answer the comprehension question was unobtrusively logged for all comprehension questions presented in the web studies, which represents the mental effort exerted. In cognitive psychology [78], response-time is measured from the moment the stimulus is presented to the first action made by the participant. In an IRTM study, response-time is measured from the time the comprehension question is first presented to the time the comprehension question answer is *submitted*, which includes memory recall and answer selection by the respondent. The IRTM's measurement of response-time does not stop after the first selection made by the respondent because multiple selections may occur before her answer is submitted. The additional time used by the respondent to select an answer may reflect an increase in mental effort to deem the video intelligible, as investigated in Chapter 6 and Chapter 7.

Chapter 5 Web Study: Relationship between PSNR and Perceived Intelligibility

Video and image quality is often objectively measured using peak signal-to-noise ratio (PSNR), but for sign language video, human comprehension is most important. Yet the relationship of human comprehension to PSNR has not been studied. Using the methodology described in Chapter 4, a web survey was created investigating how well PSNR matches perceived intelligibility of sign language video. Six low bit rates (10-60 kbps) and two low spatial resolutions (192×144 and 320×240 pixels) were used, which may be typical of video transmission on mobile phones using 3G networks.

A national web survey was created investigating user preferences and comprehension when varying the bit rates (10-60 kbps in increments of 10 kbps) and spatial resolutions (192×144 and 320×240) of ASL video that would be transmitted for mobile video phone communication. This study seeks to answer four questions:

1) When users are shown ASL video encoded at different spatial resolutions and bit rates, which combinations do they prefer?

2) How does the objective video quality measure (PSNR) compare to the subjective video quality preferences for varying bit rates and spatial resolutions?

3) For respondents who are fluent in ASL, how does video quality preference influence comprehension of video content with varied spatial resolutions and bit rates?

4) For respondents who are fluent in ASL, how do varied spatial resolutions and bit rates affect their perceived ease/difficulty of comprehension?

The findings from this study will demonstrate that intelligible sign language video can occur at transmission rates below recommended standards and demonstrate how perceived video intelligibility compares to PSNR.

5.1 PSNR Calculations

Selecting a specific spatial resolution and bit rate combination to transmit video on an experimental mobile video app is important because there are tradeoffs with computational complexity, video quality, and resource availability on smartphones. Larger video resolutions and higher bit rates result in higher video quality at the expense of increased computational power to transmit the data in real-time. Before the investigation of how resource allocation is affected by video transmission, there is a need to determine at which bit rates and spatial resolutions video can be transmitted for intelligible conversations.

Despite the fact that PSNR may not be suitable for measuring subjective video quality, it still is a reasonable measure of objective video quality when used across the same content [100]. PSNR was calculated for two different spatial resolutions (192×140 and 320×240 pixels) and 15 bit rates (10-150 kbps in increments of 10 kbps) of the same 12-second video clip of a local deaf woman signing at her natural signing pace with a stationary background. The original video was recorded at 320×240 pixels at 15 fps. Duplicate videos were created at the smaller spatial resolution before calculating the PSNR. The smaller spatial resolution was transmitted at 192×140 pixels and then enlarged and displayed at 320×240 pixels using bilinear interpolation [35] before PSNR was calculated. As Figure 7 demonstrates, the PSNR values for each spatial resolution increase monotonically with increasing bit rate.

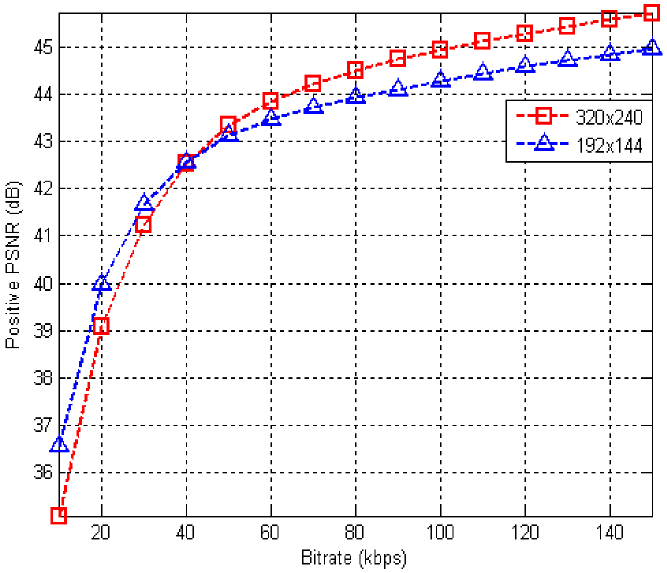


Figure 7: PSNR(dB) vs. Bit rate (kbps) for spatial resolutions displayed at 320×240 pixels. Higher PSNR means higher objective video quality. Whether it means higher subjective perception of quality is a topic of this research.

The PSNR curves demonstrate a crossover point at 40 kbps where, at lower bit rates, the smaller spatial resolutions have higher PSNR values than the larger spatial resolution. Visual inspection of the same ASL video (displayed at the same size) transmitted at lower bit rates (10-40 kbps) shows more blocky artifacts in videos sent at 320×240 pixels than at 192×144. The crossover in the PSNR plots occurs because at very low bit rates, the higher resolution video is quantized more heavily and thus has very poor visual quality (such as blockiness and loss of fine details). The same videos at lower spatial resolutions are not quantized as heavily which results in higher measured video quality. As bit rates increase, the higher resolution has higher measured video quality than the smaller spatial resolutions. This is due to blurriness from enlarging the video. The crossover of PSNR curves has been found in other video compression techniques [62, 74, 107], but the results, to my knowledge, have not been used to evaluate human comprehension, which, along with subjective quality measures, is the focus of this online survey.

Comparing PSNR to Perceived Video Intelligibility Web Study Design

From a computational perspective, transmitting video at the smaller spatial resolution and at the lowest bit rates takes the least amount of power and resources; however, without user feedback, it is uncertain whether mobile sign language communication at these transmission rates is intelligible.

Chapter 4 details the basic framework and motivation for the study design. Bit rates higher than 60 kbps were not considered since the larger spatial resolution always had higher perceived video quality than the smaller spatial resolution upon visual inspection.

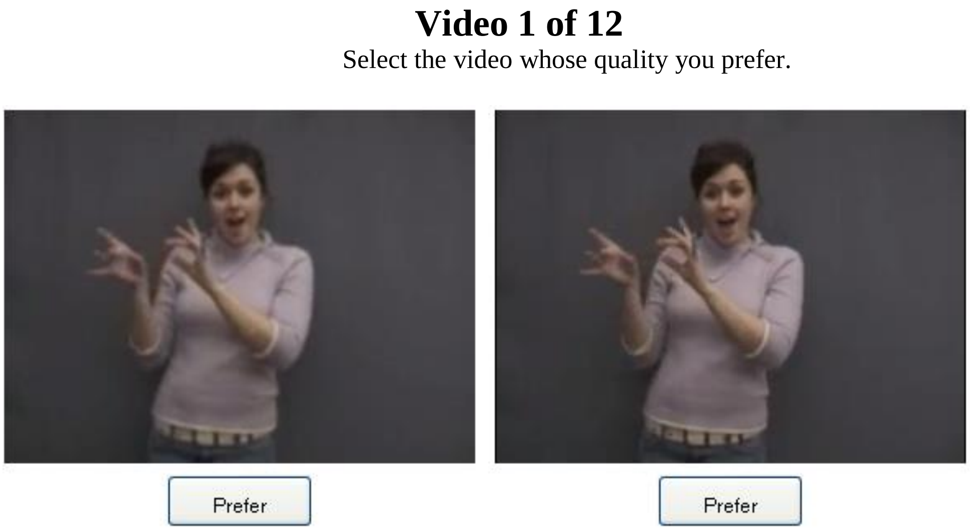
The online survey began by asking participants to self-report their fluency in ASL. The survey asked different questions depending on the response to this question. Part 1 was a paired-comparison experiment which investigated the subjective video quality preferences of ASL signers and non-ASL signers (see Figure 8). Part 2 was a single-stimulus experiment which examined comprehension of ASL video of varying bit rates and spatial resolutions (ASL signers only) (see Figure 9 and Figure 10). Finally, part 3 asked demographic questions.

To determine how subjective video quality preference differs between fluent ASL and non-ASL signers, it was important to get an equal number of ASL and non-ASL signing respondents. An online survey was selected over a laboratory study because an online survey is accessible to most people with Internet access, so more respondents could be included from across the nation.

Videos Used in Online Survey

Part 1 Videos

The same 12-second video clips used to measure PSNR of ASL video were used in part 1 of the survey. A 12-second video duration was used because it was long enough for respondents to make a video preference selection while keeping the overall survey duration to 4-7 minutes. Recall that all videos were transmitted at their respective spatial resolution (192×144 and 320×240) at varied bit rates, and then displayed at 320×240 pixels (with the smaller spatial resolution enlarged using bilinear interpolation, a standard method for enlarging video).


Video 1 of 12

Select the video whose quality you prefer.

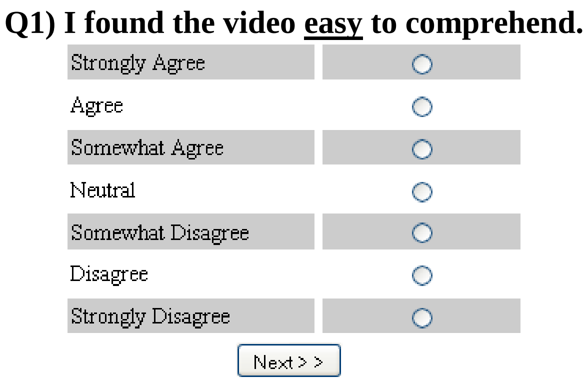
Figure 8: Screenshot of one 12-second video pair from the paired-comparison experiment. Respondents selected which video they preferred to watch.

Part 2 Videos

Twelve different video clips of the same local deaf woman signing different short stories she created at her natural signing pace were used. All videos were recorded with the same parameters. Each video was again truncated to the first 12-seconds of the story to keep the overall duration of the survey manageable and to test respondents with comprehension question about that segment. A duplicate set of the twelve videos were created and downsampled to a spatial resolution of 192×144 pixels.

Paired-Comparison Experiment

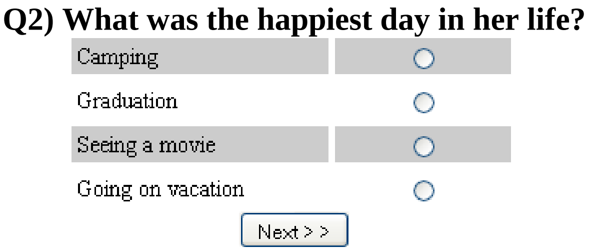
As Figure 8 demonstrates, part 1 of the survey used a paired-comparison method with simultaneous presentation as described in prior work [14]. For each of the six bit rates, a pair of videos (each at the two different spatial resolutions) was shown. This yields six pair-wise combinations, one at each bit rate. The videos were shown side-by-side on the same screen with synchronous playback. Respondents could watch the video pairs repeatedly until a selection was made. Each of the six pairs was presented twice, switching the left/right display order to counterbalance and prevent bias from video placement. None of the test pairs contain videos at different bit rates, since previous research [17] confirmed that higher bit rates were always selected when given the option. This study design resulted in twelve trials per participant. Randomization was done with an algorithm that randomly selected the next video after eliminating the previous selection. During each trial, respondents were asked to select the video whose quality they preferred. To make sure respondents watched the video pairs, they could not select a preferred video until at least four seconds after a video pair began playing. The time to select an answer was unobtrusively logged.



Q1) I found the video easy to comprehend.

Strongly Agree	○
Agree	○
Somewhat Agree	○
Neutral	○
Somewhat Disagree	○
Disagree	○
Strongly Disagree	○

Figure 9: Q1 was a 7-point Likert scale for the ease of comprehension. Q1 was shown after the video was removed from the screen.



Q2) What was the happiest day in her life?

Camping	○
Graduation	○
Seeing a movie	○
Going on vacation	○

Figure 10: Q2 asked a simple comprehension question pertaining to the video shown. Q2 was shown after Q1 was removed from the screen.

Single Stimulus Experiment

A single stimulus experiment, whose design is described in Chapter 4, was used to evaluate comprehension of ASL video transmitted and encoded at each combination of spatial resolution and bit rate. These combinations yielded twelve videos in the single stimulus experiment. Before beginning part 2, fluent ASL signers were shown a practice video to familiarize themselves with the layout.

In this web study, participants were asked to rate their agreement/disagreement on a 7-point Likert scale with the statement, “I found the video easy to comprehend.” The 7-point Likert scale was shown in descending vertical order from *strongly agree* to *strongly disagree*. The word ‘difficult’ replaced the word ‘easy’ for every other respondent, but always remained the same within a respondent. This approach prevented bias from respondents' interpretations of "easy" or "difficult." Figure 10 is an example of question 2 which asked a trivial comprehension question pertaining to the video shown. Since the ease/difficulty of comprehension varied with each 12-second video segment, the comprehension questions were only used as a way to confirm that the participant had been paying attention to the video. Finally, after respondents completed parts 1 and 2, they were asked background questions to confirm language fluency.

5.2 Results

Recall that at the start of the survey, respondents self-declared their fluency in ASL. Part 1 of the survey investigated (1) the preferences of both ASL and non-ASL signers for spatial resolution as bit rates varied, and (2) how subjective video quality preferences compared to measured PSNR values. Part 2 of the survey investigated whether comprehension of ASL video content by respondents fluent in ASL was affected by transmission bit rate and spatial resolution.

A total of 103 respondents completed the survey; however, in part 1, results were eliminated from those who used internet browsers that were survey-incompatible. Results were included from respondents who completed part 1 but failed to finish the entire survey (part 2 and demographics sections). In part 1, data were analyzed from 95 respondents: 56 ASL signers (30 men, 15 women, and 11 who did not specify) and 39 non-ASL signers (13 men, 25 women, and 1 who did not specify). Their age ranged from 18-71 years old (mean: 37 years). Of the respondents who self-reported fluency in ASL, 41 were deaf, 35 self-declared using ASL as their daily language, and the number of years of signing experience ranged from 3-58 years (mean: 26 years). Seventy-eight respondents (43 ASL, 35 non-ASL) owned a cell phone, and 72 of those cell phone owners (43 ASL, 29 non-ASL) used it to text message.

In part 2 of the survey, data were analyzed from 53 respondents (33 men, 18 women, and 2 who did not specify). Their age ranged from 18-71 years old (mean: 27 years) and all but five respondents were deaf. The self-reported number of years they have signed ASL ranged from 3-58 years (mean: 27 years). Forty-one respondents indicated they use ASL as their daily language. Finally, 48 respondents indicated they own a cell phone, with all of them using texting, and all but three respondents said they use video phones and/or video relay services.

Subjective Video Quality Preferences

Respondents were asked to select which video they preferred when presented with two videos playing simultaneously side-by-side at the same bit rates. Figure 11 shows the percentage of people *vs.* bit rate who selected the 320×240 spatial resolution over the 192×144 spatial resolution by ASL and non-ASL signing respondents.

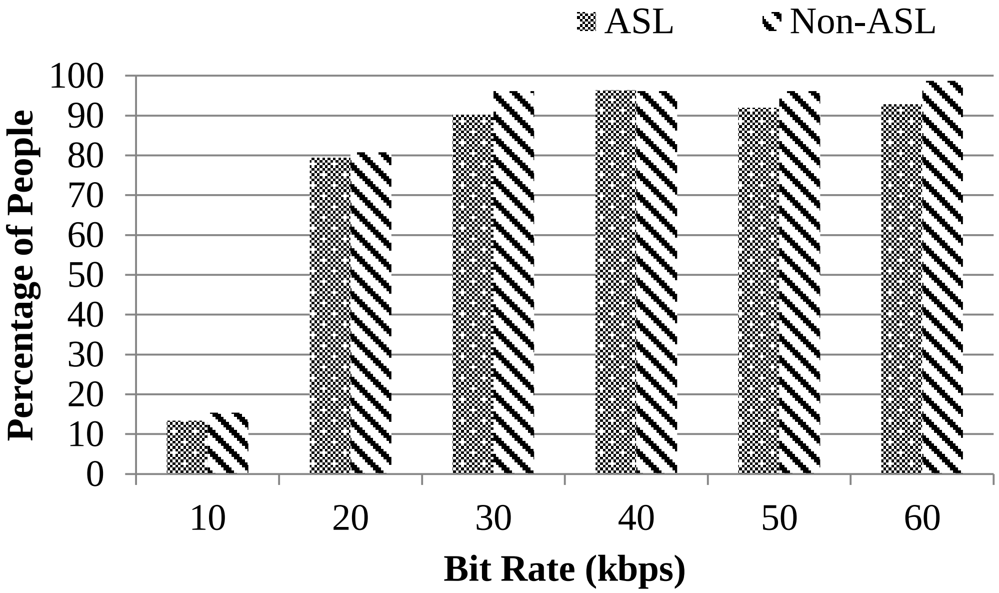


Figure 11: Percentage of People *vs.* Bit rate (kbps) who selected 320×240 instead of 192×144 spatial resolution in the paired-comparison experiment. Data are from 56 ASL signers and 39 Non-ASL signers.

A one-sample Chi-Square test was performed to test whether the proportion of subjects who picked the 320×240 spatial resolution *vs.* the 192×144 spatial resolution was significantly different than chance at each bit rate (10-60 kbps in increments of 10 kbps). Recall that both videos were *displayed* at the same spatial resolution (320×240).

At 10 kbps, both subject groups overwhelmingly preferred the video quality of the lower 192×144 spatial resolution over the 320×240 spatial resolution ($\chi^2_{1,N=95}$=97.347, p<.0001). At transmission bit rates of 20 kbps and higher, both subject groups preferred the video quality of the 320×240 spatial resolution ($\chi^2_{1,N=95}$=68.40, p<.0001).

Video Comprehension

Respondents were asked to rate their perceived ease/difficulty of comprehending each of the twelve videos on a 7-point Likert scale. Recall that the wording of this question alternated *between* respondents, but remained the same *within* each participant session.

Nonparametric analyses were used to analyze the 7-point Likert scale responses for rating the perceived ease/difficulty of comprehension. Since data gathered were ordinal and dichotomous, a Friedman test [40] was used to analyze the main effect of bit rate and spatial resolution on comprehension. Separate Wilcoxon tests [113] with Bonferroni procedure were performed to investigate the effect of spatial resolution *within* each bit rate.

The Friedman test indicated a significant main effect of spatial resolution on video comprehension ($\chi^2_{1,N=53}$=8.33, p<.01). The Friedman test also indicated a significant main effect of bit rate on video comprehension ($\chi^2_{5,N=53}$=146.15, p<.0001).

Wilcoxon tests with Bonferroni procedure were performed *within* each bit rate to identify the effect of spatial resolution on comprehension. Of the 53 respondents, 24 were asked to rate the difficulty of comprehension and 29 were asked to rate the ease of comprehension. The results of the Wilcoxon test for the perceived ease/difficulty of comprehension are presented separately, below.

Rating Difficulty of Comprehension

Recall that about half of the respondents saw a 7-point Likert scale concerning the *difficulty* of comprehension, ranging from 1 (strongly disagree), *i.e.*, less difficult to comprehend, to 7 (strongly agree), *i.e.*, more difficult to comprehend. Table 1 shows the mean Likert scale response for the difficulty of comprehending the ASL video transmitted at each bit rate and spatial resolution and displayed at 320×240 pixels.

Figure 12 is a double *y*-axis plot of the mean Likert responses and the negative PSNR values for each bit rate and spatial resolution. Notice that the PSNR values are *negative*, where *lower* values correspond to *higher video quality*.

Comprehension was significantly less difficult at 60 kbps for the 320×240 spatial resolution than the 192×144 spatial resolution (Z=35.0, *p*<.01). However, changing the spatial resolution within other bit rates did not indicate more difficulty in comprehension. For example, Table 1 and Figure 12 indicated a large difference of mean Likert scores at 40 kbps, but changing the spatial resolution within that bit rate was not significant in affecting the difficulty of comprehension (Z=48.5, *n.s.*). Figure 12 may suggest that there is a large difference between mean Likert scores at 40 kbps; however, Wilcoxon tests with Bonferroni procedure were performed indicating this different was not significant.

Table 1: Mean Likert Scale responses (1-7) for difficulty of comprehending video quality. Note *lower* Likert scores correspond to *less* perceived difficulty.

	Spatial Resolution			
	320×240		192×144	
Bit rate	Mean	Std. Error	Mean	Std. Error
10	6.00	0.28	**5.71**	0.24
20	**4.38**	0.35	4.54	0.29
30	3.83	0.33	**3.54**	0.32
40	**2.75**	0.33	3.79	0.33
50	**2.75**	0.33	3.42	0.31
60	**2.67**	0.30	3.41	0.35

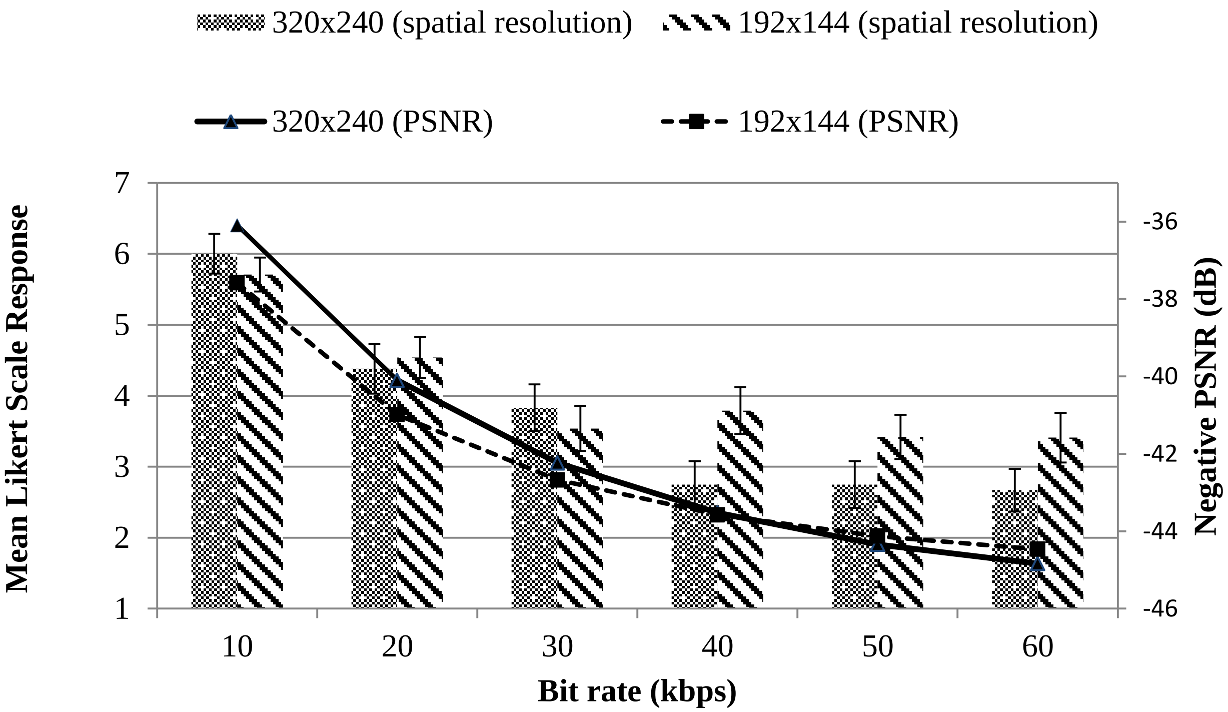


Figure 12: Double *y*-axis plot of 7-point Likert scale. Negative PSNR values of spatial resolutions and bit rates. Lower Likert scores correspond to less difficulty and lower PSNR values correspond to higher video quality. Notice a negative PSNR crossover point occurs at 40 kbps.

Rating Ease of Comprehension

Recall that about half the respondents saw a 7-point Likert scale concerning the *ease* of comprehension, ranging from 1 (strongly disagree), *i.e.*, less easy to comprehend, to 7 (strongly agree), *i.e.*, more easy to comprehend. Table 2 shows the mean Likert scale response for the ease of comprehending ASL video transmitted at each bit rate and spatial resolution and displayed at 320×240 pixels.

Figure 8 is a double *y*-axis plot of the mean Likert responses and the positive PSNR values for each bit rate and spatial resolution. Notice that the PSNR values are *positive*, where *higher* values correspond to *higher video quality*.

Transmitting at 320×240 spatial resolution rather than at a 192×144 spatial resolution at 50 and 60 kbps was significantly easier to comprehend (Z=100.0, *p*<.001 and Z=88.5, *p*<.001, respectively). This result is also shown in the PSNR curve in Figure 8; at 50 kbps and 60 kbps, the positive PSNR values were higher for the larger spatial resolution. However, changing the spatial resolution within other bit rates did not make the content easier to understand. Even though Table 2 and Figure 13 indicate a large difference of mean Likert score at 10 kbps, changing the spatial resolution within that bit rate was not significant in affecting comprehension (Z=45.5, *n.s.*).

Table 2: Mean Likert Scale responses (1-7) for ease of comprehending video quality. Note *higher* Likert scores correspond to *easier* perceived comprehension.

	Spatial Resolution			
	320×240		192×144	
Bit rate	Mean	Std. Error	Mean	Std. Error
10	2.90	0.31	**3.55**	0.28
20	**5.10**	0.29	4.72	0.29
30	5.34	0.26	**5.48**	0.26
40	**5.90**	0.25	5.41	0.23
50	**6.27**	0.19	5.48	0.22
60	**6.34**	0.14	5.62	0.20

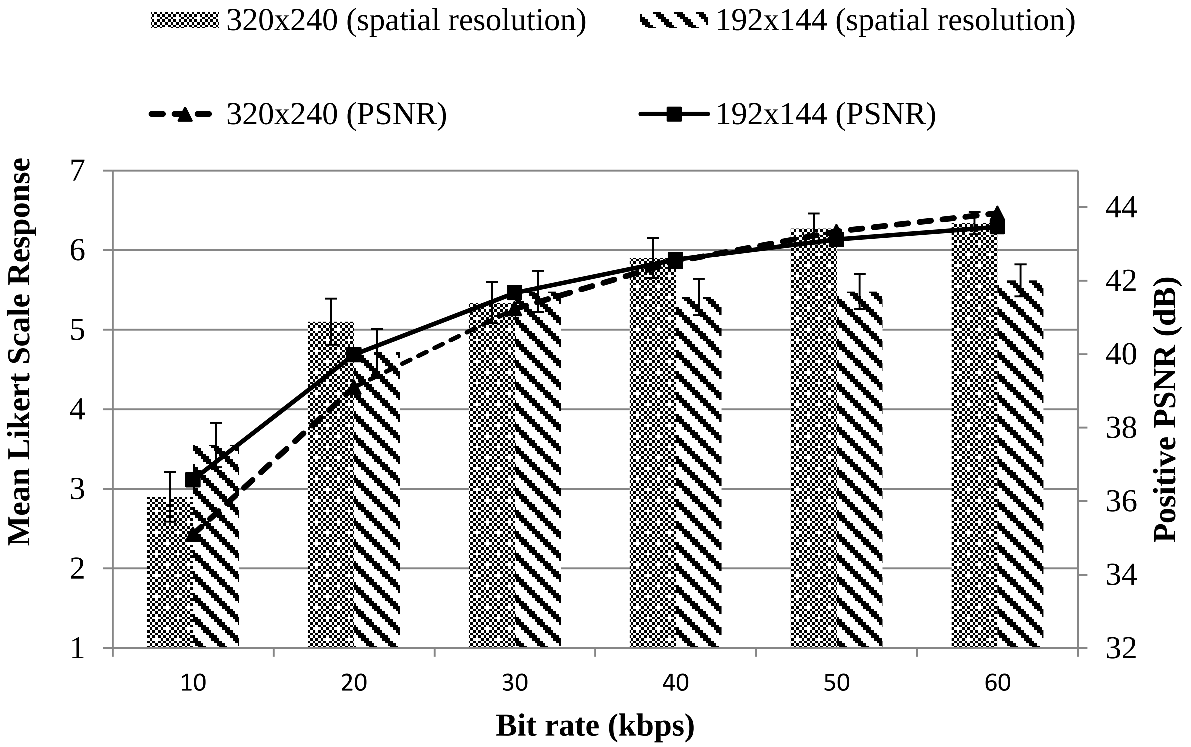


Figure 13: Double *y*-axis plot of 7-point Likert scale. Positive PSNR values of spatial resolution and bit rate. Higher Likert scores correspond to more ease and higher PSNR values correspond to higher video quality. Notice a positive PSNR crossover point occurs at 40 kbps.

5.3 Discussion

Video preferences from part 1 were compared to PSNR measurements, which reinforced the claim that PSNR may not accurately reflect subjective video quality. The PSNR values suggested that bit rates at 40 kbps and lower spatial resolution of 192×144 pixels had higher objective quality than the 320×240 spatial resolution; however, subjective user preferences revealed that at 20 kbps and higher, the larger spatial resolution was preferred. This finding is not unexpected since PSNR does not account for compression artifacts (blockiness and Gibbs's phenomena [42]) that can be highly distracting for users. Also, visual inspection of each pair of videos showed that at bit rates 20 kbps and higher, enlarging the smaller spatial resolution to display at 320×240 pixels caused the video to appear more blurry than when simply transmitting the larger spatial resolution.

One might expect that the same bit rates and spatial resolutions indicated as preferred in part 1 would similarly influence content comprehension; that is, that respondents would indicate greater ease (or less difficulty) of comprehension when shown video at the 320×240 spatial resolution at bit rates of 20 kbps and higher. However, transmitting either spatial resolution at 10-50 kbps had no effect on making comprehension more difficult. At 60 kbps only, respondents expressed that transmitting the larger spatial resolution made the content significantly less difficult to comprehend. This result was the same among the respondents who were asked to rate the ease (rather than the difficulty) of comprehension. Neither of the two spatial resolutions, at bit rates of 10 to 40 kbps, made comprehending the video easier. However, at 50 and 60 kbps, respondents did indicate that transmitting the larger spatial resolution made comprehension easier. When comparing these findings to the PSNR curves (Figure 12 and Figure 13), PSNR measurements may accurately reflect the perceived ease/difficulty at which respondents rated comprehension of ASL video. The PSNR curves showed a threshold where at 50 kbps and higher, transmitting the larger spatial resolution produces better video quality than transmitting and enlarging the smaller spatial resolution. The results of the survey agree with this and also indicate that at 50 kbps and higher, video comprehension was made easier.

These results suggest that PSNR may be a reliable measure for ASL video intelligibility and can further assist in selecting the spatial resolution and bit rate for mobile video telephony. When possible, selecting the smaller spatial resolution at the PSNR crossover point provides intelligible video while keeping computational complexity and cost of video transmission low. From these study results, a recommendation of transmitting mobile sign language video at 40 kbps at 192×144 spatial resolution would be sufficient to hold an intelligible conversation while saving limited computing resources.

Chapter 6 Web Study: Perceived Video Intelligibility

Having investigated perceived video intelligibility when varying the bit rate and spatial resolution and how intelligibility compares to PSNR, I now turn to investigating perceived video intelligibility when video is transmitted at low frame rates and bit rates. Mobile sign language video conversations can become unintelligible due to high video transmission rates causing network congestion and delayed video. In an effort to understand how much sign language video quality can be sacrificed, the perceived lower limits of intelligible sign language video transmitted at four low frame rates (1, 5, 10, and 15 frames per second [fps]) and four low fixed bit rates (15, 30, 60, and 120 kilobits per second [kbps]) are evaluated in a new national web survey. The goal of this study is to demonstrate that relaxing the recommended international video transmission rate, 25 fps at 100 kbps or higher, would still provide intelligible content while considering network resources and bandwidth consumption.

Study Design

Chapter 4 formally describes the motivation and implementation of web study structure. The HSIM (described in Chapter 3) informs the design of this new web study evaluating how much frame rate and bit rate can be reduced before intelligibility is compromised in mobile sign language video communication. Owing to the need to ensure all receivers' minds are fully capable of comprehension, participants were screened for ASL fluency. Thereafter, I can attribute any differences in comprehension to differences in intelligibility and not knowledge.

This web study evaluated sign language video intelligibility transmitted at four low frame rates (1, 5, 10, and 15 fps) and four low bit rates (15, 30, 60, and 120 kbps) in a full-factorial design. The spatial resolution was held constant at 320×240 pixels because the video stimuli was recorded at that spatial resolution. The web study was selected over a laboratory study because parameter settings could be evaluated with more participants from across the nation. The survey consisted of three parts and took 12-26 minutes to complete. Upon survey completion, participants had an opportunity to enter their email for a chance to win one of four $75 gift cards. Their e-mail was not associated with their anonymous and confidential responses. A male native ASL signer/consultant signed 16 short ASL sentences specifically chosen to include various amounts of finger spelling and descriptive lexicons. The English sentences and the ASL glosses are listed in Appendix B.

6.1 Results

The web survey received 300 hits, with 99 respondents completing the survey, all of whom self-reported fluency in ASL. Results were eliminated from those who responded with the same answers for all 16 videos, such as selecting all 1s or all 7s. Data were analyzed from 77 respondents (48 women). Their age ranged from 18-72 years old (median=40 years, *SD*=12.73 years). Of the 77 respondents: 56 were deaf (38- native ASL signers, 11 of 38 have deaf parents), 54 indicated ASL as their daily language, and the number of years they have spoken ASL ranged from 5-59 years (median=28 years, *SD*=12.73). All but 7 respondents owned a smartphone and sent text messages; 65 indicated they use video chat; and 53 use video relay services.

Perceived Intelligibility

Results will be reported in terms of intelligibility even though comprehension questions were asked. As outlined in the HSIM, video intelligibility can be inferred from comprehension questions provided that the receivers' knowledge stores are fully adequate to understand the received signals. Nonparametric analyses were used to analyze the Likert responses since the data were ordinal and not normally distributed. Analysis was performed using the nonparametric *Aligned Rank Transform* [115] procedure that enables the use of ANOVA after alignment and ranking, while preserving interaction effects.

Frame Rate Main Effect

Frame rate was found to have a significant main effect on video intelligibility (F(3,1139)=636.99, *p*<.0001). Post-hoc contrast tests with Holm's sequential Bonferroni procedure [51] were performed for 1 fps *vs.* 5 fps; 5 *vs.* 10 fps; 5 *vs.* 15 fps; and 10 fps *vs.* 15 fps. Table 3 and Figure 14 list the mean Likert score for question 1, where higher scores correspond to higher agreement with the ease of perceived understanding of video content. As expected, videos displayed at 5 fps when compared to 1 fps received higher mean Likert scores for video intelligibility (F(1,1139)=921.07, *p*<.0001). Videos displayed at 10 fps when compared to 5 fps received higher mean Likert scores for video intelligibility (F(1,1139)=111.13, *p*<.0001). However, when comparing 10 fps *vs.* 15 fps, videos displayed at 10 fps were found to have a higher mean Likert score for intelligible content (F(1,1139)=77.22, *p*<.0001). As Figure 14 shows, videos displayed at 10 fps (averaged across four bit rates) received higher mean Likert scores than all other frame rates. An unexpected finding was that videos were not perceived to be more intelligible at 5 fps *vs.* 15 fps (F(1, 1139)=3.11, *n.s.*). One would expect that a higher frame rate would yield higher intelligibility for a temporal language since the ITU-T recommends 25 fps for intelligible sign language video.

Bit Rate Main Effect

Changing the bit rate was found to have a significant main effect on ASL video intelligibility (F(3,1139)=145.53, *p*<.0001). Post-hoc contrast tests with Holm's sequential

Bonferroni procedure [51] were performed for 15 kbps *vs.* 30 kbps; 30 kbps *vs.* 60 kbps; and 60 kbps *vs.* 120 kbps. Unsurprisingly, increasing the bit rate from 15 kbps to 30 kbps to 60 kbps was found to significantly improve ASL video intelligibility (F(1,1139)=82.75, *p*<.0001). However, videos displayed at 60 kbps *vs.* 120 kbps were not found to be significantly different in terms of intelligibility (F(1,1139)=4.62, *n.s.*).

Frame Rate × Bit Rate Interaction

There was also a significant frame rate × bit rate interaction (F(9,1139)=23.40, *p*<.0001*).* Upon closer inspection, videos transmitted at 10 fps, independent of bit rate, received the highest mean Likert scores for ease of understanding video quality as shown in Table 3 and Figure 14. Additionally, videos displayed at 60 kbps *vs.* 120 kbps were not found significantly different in terms of intelligibility, which is reflected by similar mean Likert scores suggesting that 60 kbps is a high enough bit rate to transmit intelligible video. Also, video transmitted at 15 fps have more artifacts since fewer bits are allocated to each frame. Videos displayed at 1 fps received the lowest mean Likert score, suggesting that 1 fps is too low to support intelligible sign language video.

Table 3: Mean Likert score responses for ease of understanding video quality. Note higher Likert scores correspond to higher perceived intelligibility.

	Bit rate (kbps)							
	15		30		60		120	
frame rate (fps)	Mean Likert	std. error	Mean Likert	std. error	Mean Likert	std. error	Mean Likert	std. error
1	2.14	0.14	1.13	0.07	1.75	0.11	1.90	0.10
5	3.01	0.16	4.43	0.15	4.95	0.14	4.75	0.13
10	**4.04**	0.16	**4.74**	0.13	**5.66**	0.13	**5.91**	0.14
15	3.51	0.17	3.97	0.15	5.13	0.15	5.25	0.14

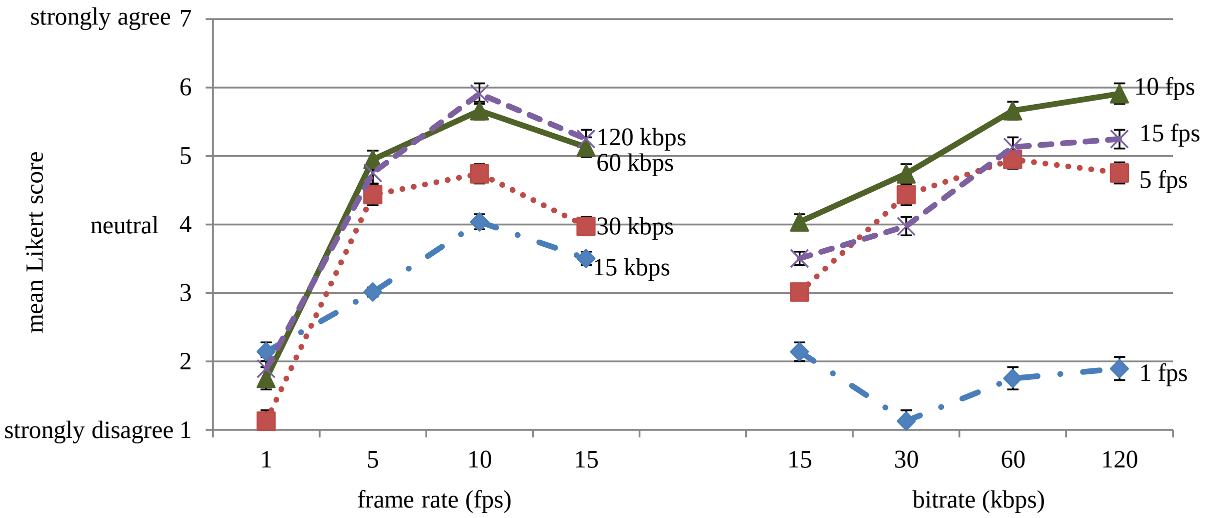


Figure 14: Plot of 7-point Likert ratings for participants' ease of understanding the video for each frame rate and bit rate averaged over all participants. Error bars represent ±1 standard error.

Comprehension Question Response-Time

The time participants took to respond to the comprehension questions was unobtrusively logged and the logged time started when the question appeared on the screen and ended when the answer was submitted. Thirteen of 16 comprehension questions were answered correctly with 95% accuracy or higher. Findings are reported on correctly answered comprehension questions across frame rates (averaged over all four bit rates) and across bit rates (averaged over all four frame rates). Table 4 lists the mean time and standard deviation for respondents who answered the comprehension question correctly.

The fastest mean response-times for correctly answering the comprehension questions for both frame rate (averaged over all four bit rates) and bit rate (averaged over all four frame rates) were found to receive the highest mean Likert scores for perceived video intelligibility. These results are demonstrated by the strong negative correlation between mean response-time and mean Likert scores for frame rate (averaged overall all four bit rates) (r=-0.66); and mean response-time and mean Likert scores for bit rate (averaged overall all four frame rates) (r=-0.82). These results suggest that higher perceived video intelligibility leads to faster content comprehension; this particular relationship is explored more thoroughly in Chapter 7. Figure 15 is a double *y*-axis plot showing mean Likert score rating perceived video intelligibility *vs.* mean response-times for correctly answering the comprehension questions for both frame rate (averaged over all four bit rates) and bit rate (averaged over all four frame rates).

Table 4: Mean Likert score (higher values are better) and mean response-time (in seconds) for correctly answered comprehension questions for both frame rate (averaged over all four bit rates rates) and bit rate (averaged over all four frame rates). Bold values indicate highest mean Likert scores and fastest times to submit answer.

Frame rate (fps)	Mean Likert Score	std. error	Mean Response Time (sec)	SD	Bit rate (kbps)	Mean Likert Score	std. error	Mean Response Time (sec)	SD
1	1.77	0.10	6.34	5.19	15	3.18	0.16	5.97	3.18
5	4.29	0.15	6.07	3.74	30	3.61	0.13	5.81	5.28
10	**5.09**	0.14	**4.19**	1.74	60	4.37	0.13	5.03	2.62
15	4.46	0.15	4.51	2.17	120	**4.45**	0.13	**4.11**	1.89

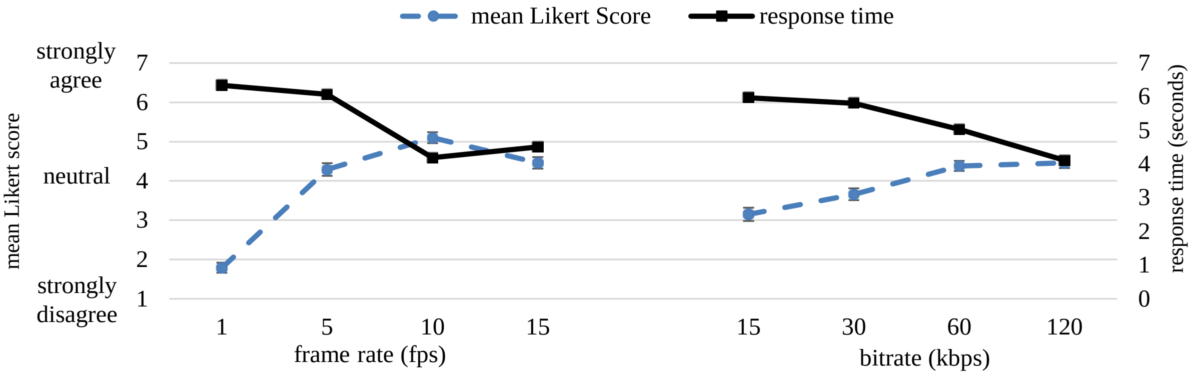


Figure 15: Double y-axis plot of a 7-point Likert scale rating participants' ease of understanding the video and mean response-time (seconds) for correctly answered comprehension questions for both frame rate (averaged over all four bit rates rates) and bit rate (averaged over all four frame rates). Higher Likert scores correspond to higher perceived intelligibility.

6.2 Discussion

6.2.1 Frame Rate and Bit Rate

I anticipated finding frame rate and bit rate pairs where video quality begins to affect intelligibility too negatively or diminishing returns begin. Unsurprisingly, respondents overwhelmingly ranked video displayed at 1 fps to have the lowest mean Likert scores for ease of understanding the video content. One fps was selected to achieve a sufficiently low frame rate so that we “bottomed out” on intelligibility. Prior work investigating the impact of frame rate on perceived video quality acknowledged not selecting a low enough frame rate to explore [4,16]. Although transmitting video at 1 fps is not ideal for ASL conversations, it was observed that transmitting video at 1 fps and 15 kbps, which is the lowest bit rate, received the highest mean Likert score across all bit rates at 1 fps. This finding corroborates my earlier finding [25] that people perceived the least amount of negative effects when the lowest frame rate and bit rate settings were applied.

Diminishing returns for videos displayed at 60 kbps and 120 kbps independent of frame rate were discovered. Figure 14 shows how the mean Likert scores for 60 kbps and 120 kbps, when averaged over all four frame rates, had similar Likert scores and were not found significantly different in terms of intelligibility (F(1,1139)=0.47, *n.s.*). These findings suggest 60 kbps is high enough to provide intelligible video conversations.

Another important finding was that video transmitted at 10 fps received a higher mean Likert score than video transmitted at 15 fps across all bit rates. One would think that ASL, which is a temporal visual language, would require video communication to be transmitted at higher frame rates; however, this may not be the case at low bit rates. The preference of viewing

ASL video at 10 fps over 15 fps was also discovered in earlier ASL video communication research conducted by Cavender *et al*. [17]. However, their findings only reported a slight but significant main effect that frame rate influenced video intelligibility. My results strongly affirm that ASL video intelligibility peaks at 10 fps across all bit rates. At a fixed low bit rate, more bits are allocated per frame at 10 fps *vs.* 15 fps, and this difference is noticeable enough to result in higher perceived intelligibility. These findings suggest that relaxing the recommended frame rate and bit rate to 10 fps at 60 kbps will provide intelligible video conversations while reducing total bandwidth consumption to 25% of what the current recommended standards of 25 fps at 100 kbps or higher consume.

6.2.2 Comprehension Question Response-Time

The strong negative correlation between mean Likert scores for rating perceived video intelligibility and mean response-times for correctly answering comprehension questions for both frame rate (averaged over all four bit rates) and bit rate (averaged over all four frame rates) suggests higher video transmission rates lead to faster comprehension of video content. There are limitations to these preliminary findings since comprehension difficulty level was not controlled for. Some videos used in this study may have been easier to comprehend than others due to varied amounts of finger spelling and descriptive lexicons used. Chapter 7 describes a new web study that investigates this relationship more thoroughly. Nevertheless, I observed that respondents answered comprehension questions more quickly when viewing ASL video with higher perceived intelligibility, suggesting that measuring response-time may serve as an indicator for measuring video intelligibility.

6.2.3 Signing Speed

The signing speed used in the video stimuli may have contributed to the non-significant intelligibility improvement of video transmitted at 5 fps *vs.* 15 fps. These findings suggest that 5 fps would be sufficient for intelligible video communication.

6.3 Summary

It was discovered that intelligibility was affected too negatively at 1 fps at 15 kbps, and increasing transmission rates beyond 10 fps at 60 kbps provided negligible gains. Regardless, these findings suggest that the recommended ITU-T sign language transmission rates can be relaxed to 10 fps/60 kbps while preserving intelligible ASL video and reducing bandwidth and network load.

Chapter 7 Web Study: Response-Time and Mental Effort Relationship

Chapter 6 described preliminary results suggesting a relationship between response-time and mental effort. This chapter investigates more rigorously whether a relationship between response-time and mental effort exists and how it could be used to inform video intelligibility evaluations. Often evaluations of video intelligibility do not account for the mental effort required by viewer to understand the content. Measuring the mental effort required of viewers would give better insight to the lower limits at which mobile sign language video can be transmitted. In psychology, response-time has been used to measure cognitive load and working memory in task completion, where longer response-time corresponds to higher cognitive load [2,3]. Chapter 2 discussed prior work that explored this relationship. Establishing a similar relationship between mental effort (as measured by response-time) and video intelligibility may provide a meaningful method to evaluate video intelligibility.

I introduce the Intelligibility Response-Time Method (IRTM), a new method using response-time as an indicator of mental effort, along with perceived video intelligibility and comprehension accuracy, to investigate the lower limits at which sign language video can be transmitted, as shown in Figure 16. The IRTM is tested in a new web study evaluating perceived intelligibility of sign language video transmitted at four low frame rates (5, 10, 15, and 30 frames per second) and four low bit rates (15, 30, and 60, 120 kilobits per second). These combinations were compared to the ITU-T standard.

This chapter presents (1) the development of the Intelligibility Response-Time Method (IRTM); (2) empirical findings from a web study using the IRTM in evaluating perceived intelligibility of video transmitted at four low frame rates and four low bit rates; and (3) recommendations for the lowest frame rates and bit rates that do not negatively impact perceived video intelligibility or increase mental effort.

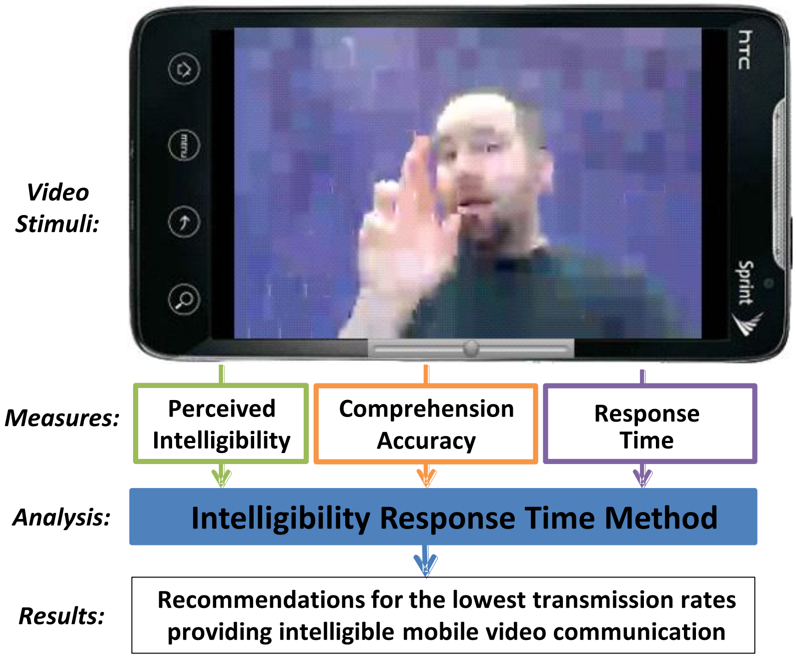

Video Stimuli:
htc
Sprint
Measures:
Perceived Intelligibility
Comprehension Accuracy
Response Time
Analysis:
Intelligibility Response Time Method
Results:
Recommendations for the lowest transmission rates providing intelligible mobile video communication

Figure 16: Flow diagram example of the Intelligibility Response-Time Method used for video intelligibility analysis.

7.1 Existing Workload Evaluations

There are other subjective tools to assess perceived workload, such as the NASA Task Load Index (NASA-TLX) [46], the Subjective Workload Assessment Technique (SWAT) [86], and the Workload Profile (WP) [104]; however, these workload evaluations were created to evaluate subjective workload between various human-machine environments such as aircraft cockpits, command, control, and communication workstations. These evaluations are often time consuming and require multiple questions per task, which is not conducive to establishing a relationship between metal effort (captured by response-time) and video intelligibility in a web study. For this web study, one question was used to evaluate mental effort instead of multiple questions per stimuli.

The NASA-TLX evaluates workload across six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration. Responses are measured on a Likert scale ranging from high, medium, and low, with 21 gradations on the scale. SWAT is another subjective workload assessment technique which requires respondents to give a 3-point rating (low, medium, high) to time pressure, mental effort exerted, and stress level corresponding to a stimuli presented. Finally, WP assesses subjective mental workload by having respondents experience all tasks and then indicate the proportion of attention allocated to each task. Although all of these methods assess workload, I establish a simpler technique utilizing the relationship between response-time and mental effort to better inform video intelligibility evaluations. The amount of effort exerted by viewers is often overlooked and may provide meaningful insight to video intelligibility evaluations.

Building Upon the HSIM

Comprehension question response-time can be attributed to many factors within a study such as the study structure, video stimuli, and comprehension questions used. By accounting for and controlling specific factors within the study design, I investigate whether a relationship between comprehension question response-time and perceived video intelligibility exists, and if so, leverage this relationship to improve video intelligibility evaluations. Chapter 3 described the Human Signal Intelligibility Model, a conceptual model differentiating signal intelligibility and signal comprehension for evaluation, which is used in the creation of the IRTM, described next.

7.2 Intelligibility Response-Time Method

Using the HSIM's definition of intelligibility, I introduce the IRTM, which uses response-time as a measure of mental effort in evaluating video intelligibility. During the first part of an IRTM study, participants watch a video and rate the perceived ease of understanding the video on a 7-point Likert scale, with 7-very easy and 1-very difficult. Second, respondents answer a comprehension question pertaining to the video as quickly as possible. The time to answer the comprehension question is unobtrusively recorded, which represents the mental effort exerted. In cognitive psychology [78], response-time is measured from the moment the stimulus is presented to the first action made by the participant. In an IRTM study, response-time is measured from the time the comprehension question is first presented to the time the comprehension question answer is *submitted*, which includes memory recall and answer selection by the respondent. The IRTM's measurement of response-time does not stop after the first selection made by the respondent because multiple selections may occur before her answer is submitted. The additional time used by the respondent to select an answer may reflect an increase in mental effort to deem the video intelligible, as investigated in the web study described below. The IRTM's definition of response-time and how it is measured will be used throughout the remainder of this dissertation.

For each video shown, three data points were gathered: (1) Likert score representing perceived intelligibility of video; (2) comprehension question response; and (3) response-time to answer the comprehension question as quickly as possible. It is important to present comprehension questions with similar levels of difficulty so that variance in response-times will result from changes in video intelligibility and not question difficulty. Below, the web study design outlines how comprehension question difficulty is accounted for.

To draw a relationship between perceived video intelligibility and response-time, there are two inclusion criteria for which data are analyzed. First, all Likert scores and response-times associated with correctly answered comprehension questions are included for analysis. Second, Likert scores and response-times associated with comprehension questions that were *incorrectly* answered *and* received Likert scores indicating somewhat difficult, difficult, and very difficult (Likert ratings 1-3 on a 7-point scale) to perceived ease of understand the video are also included. Data from the latter inclusion criteria are necessary for analysis because the goal is to identify that the lower limits of video intelligibility have been surpassed. Figure 17 is an example of the data selected for analysis using the IRTM. This subset of data will be used to create an objective score of subjective intelligibility derived from the relationship between question response-time and perceived video intelligibility. Finally, the IRTM results will aid in the selection of the lowest transmission parameters that optimize intelligibility and mental effort.

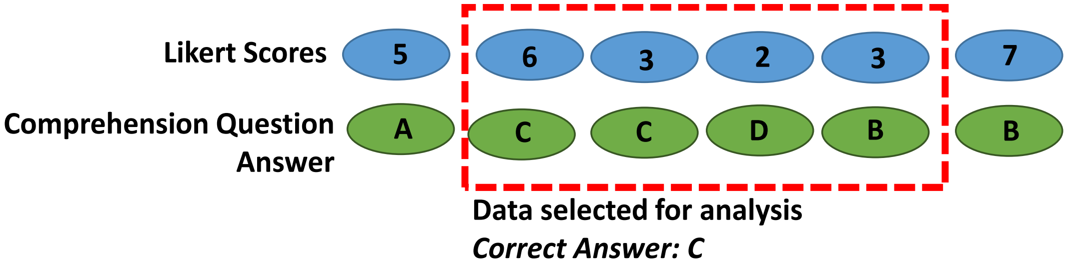


Figure 17: Example of data inclusion using the Intelligibility Response-Time Method. Note a 7-point Likert scale is used, where 7 is very easy and 1 is very difficult for how easy the video was to understand.

Study Design

The HSIM and IRTM informed the web study design and Chapter 4 formally describes the web study structure. The IRTM is tested from the data collected from a new web study evaluating perceived video intelligibility transmitted at four low frame rates (5, 10, 15, 30 frames per second) and four low bit rates (15, 30, 60, 120 kilobits per second). Using the IRTM, I (1) determine how response-time correlates with perceived intelligibility, and (2) compare findings from the lower frame rates and bit rates to the ITU-T standard. In this web study, 30 fps and 120 kbps were selected as the ITU-T standard to compare against the lower frame rates and bit rates in a full-factorial design. I aim to demonstrate that the ITU-T recommended standards for sign language video can be relaxed without negatively impacting video intelligibility or severely increasing mental effort.

7.2.1 Logging Response-Time

Participants were prompted to answer each question "as quickly as possible." The amount of time respondents took to answer each comprehension question was unobtrusively logged. The start time began when the question appeared on the screen and the stop time occurred once the answer was submitted, as specified in using the IRTM.

In this study, participants watched 16 different videos at each frame rate and bit rate combination (4 frame rates and 4 bit rates). There were 256 possible content, frame rate, and bit rate combinations for respondents to view. Each participant was randomly assigned to view each video with a randomly assigned frame rate and bit rate pair without replacement. To ensure that ample data were collected for the various settings, all 256 content setting combinations were shown every 16 participants.

7.3 Results

The results for comprehension question accuracy, perceived video intelligibility, and comprehension question response-times are presented below.

Demographics

This web survey received 275 hits, with 74 respondents (43 women) completing the survey, all of whom self-reported fluency in ASL. Their age ranged from 20-67 years old (median=37 yrs, *SD*=13.07 yrs). Of the 74 respondents: 56 were deaf (39 of 56 are native ASL signers, 10 of 39 native ASL signers have deaf parents, and 25 of 39 native ASL signers have parents who natively sign ASL). Forty-nine respondents indicated ASL as their daily language, 10 indicated speaking English and ASL, and the remaining 18 respondents communicated in

English. The number of years signing ASL ranged from 4-62 years (median=24 yrs, *SD*=13.83 yrs). All but 10 respondents use video phones and video chat applications, with FaceTime (52 respondents) and Skype (33 respondents) listed as the most popular applications. Finally, using the IRTM gave a subset of data for analysis, resulting in 1162 data points for both Likert scores and comprehension question response-times.

Separate analysis was initially performed for native ASL signers and English speaking ASL signers to determine whether results differed among groups. It was discovered that results among both groups produced the same findings; therefore, that data were combined for analysis and reported.

7.3.1 Comprehension Question Accuracy

A one-sample Chi-Square test of proportions was performed on the percentage of comprehension questions answered correctly to determine whether frame rate or bit rate affected comprehension question accuracy. Frame rate (when averaged over bit rates) was not found to impact comprehension question accuracy ($\chi^2_{(3,N=1162)}$=6.21, *n.s.*). However, bit rate (when averaged over frame rate) was found to impact comprehension question accuracy ($\chi^2_{(3,N=1162)}$=43.34, *p*<.0001.) Mainly, comprehension accuracy increased as bit rate increased. This result is clearly expected since more bits are allocated to each frame. Table 5 shows the percentage of correctly answered comprehension questions across frame rate and bit rate.

Table 5: Percentage of correctly answered comprehension questions across frame rate and bit rate. Accuracy increases with bit rate but not with frame rate.

	frame rate (fps)			
bit rate (kbps)	5	10	15	30
15	83.58	91.78	89.71	84.29
30	91.43	90.28	97.10	91.55
60	93.85	98.48	100.00	97.22
120	97.06	98.57	100.00	100.00

7.3.2 Perceived Intelligibility

Recall that higher Likert scores correspond to higher ease of perceived understanding of video content. Nonparametric analyses were used to analyze the Likert responses since the data were ordinal and not normally distributed. Analysis was performed using the nonparametric *Aligned Rank Transform* [115] procedure that enables the use of ANOVA after alignment and ranking, while preserving interaction effects. Table 6 and Table 7 lists the mean Likert scores for frame rate (averaged over bit rate) and bit rate (averaged over frame rate) rating the perceived ease of understanding the video, respectively.

Frame Rate Main Effect

Frame rate was found to have a significant main effect on perceived video intelligibility (F(3,1007)=43.81, *p*<.0001). Post-hoc contrast tests with Holm's sequential Bonferroni procedure [51] were performed for 5 fps *vs.* 10, 15, and 30 fps, respectively. There was an increase in perceived video intelligibility for video displayed at 10 fps over 5 fps (F(1,1007)=80.75, *p*<.0001); no significant difference between video displayed at 10 fps vs. 15 fps (F(1,1007)=2.07, *n.s.*); and decrease in perceived video intelligibility for video displayed at 15 fps vs. 30 fps (F(1,1007)=70.21, *p*<.0001). The decrease in perceived video intelligibility between 15 fps and 30 fps supports the notion of an “intelligibility ceiling effect” where increasing the frame rate (while holding the bit rate constant) does not improve or even reduces perceived video intelligibility beyond a certain point.

Bit Rate Rate Main Effect

Bit rate also had a main effect on perceived video intelligibility (F(3,1007)=249.01, *p*<.0001), which indicates that increasing the bit rate (averaged over frame rate) increased perceived ease of understanding video content. Post-hoc contrast tests with Holm’s sequential Bonferroni procedure were performed for 15 kbps *vs.* 30 kbps; 30 kbps *vs.* 60 kbps; and 60 kbps *vs.* 120 kbps. Unsurprisingly, as the bit rate increased, the higher bit rate was always found to significantly improve perceived ASL video intelligibility (F(9,1007)=158.2, *p*<.0001).

Frame Rate × Bit Rate Interaction

There was also a significant frame rate × bit rate interaction (F(9,1006)=9.35, *p*<.0001*).* Upon closer inspection, specific frame rate and bit rate combinations influenced perceived intelligibility ratings more than others. At 60 kbps, video transmitted at 10 fps received the highest mean Likert scores for perceived video intelligibility. Additionally, at 15 fps, videos displayed at 30 kbps *vs.* 60 kbps were not found to be significantly different for perceived intelligibility.

7.4 Comprehension Question Response-Time

Mental effort was measured by comprehension response-time. Recall, comprehension question response-time started when the question appeared on the screen and ended when the answer was submitted. A log-transform was performed on this temporal data before performing a repeated measures ANOVA due to its lognormal distribution, which is typical of temporal measures.

Frame Rate Main Effect

Frame rate did not have a main effect on comprehension question response-time (F(3,1007)=1.45, *n.s.*). The non-significant result of frame rate on response-time may indicate that the same amount of mental effort is needed to understand the content, regardless of the frame rate at which information is presented. Table 6 and Figure 18 list the mean response-time for respondents answering the comprehension questions for frame rate averaged over bit rate.

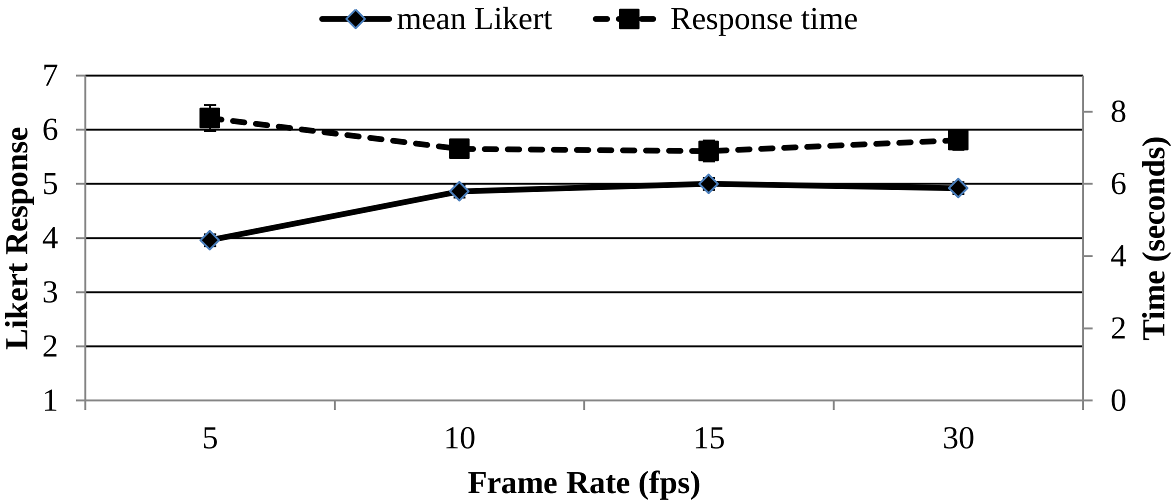


Figure 18: Plot of the mean Likert scores (higher values are better) and mean response-times (in seconds) for comprehension questions for frame rate (averaged over bit rates rates).

Table 6: Mean Likert scores (higher values are better) and mean response-times (in seconds) for comprehension questions for frame rate (averaged over bit rates rates). Bold values indicate highest mean Likert scores and fastest times to submit answer.

Frame Rate^	Mean Likert	std. error	Mean Response-time (secs)	std. error
5	3.96	0.11	7.82	0.36
10	4.86	0.11	6.97	0.23
15	**5.00**	0.11	**6.91**	0.29
30	4.92	0.11	7.21	0.27

^averaged over bit rates; *averaged over frame rates

Bit Rate Main Effect

Bit rate had a main effect on comprehension question response-time (F(3,1007)=3.95, *p*<.01). This result was expected since bit rate also significantly affected comprehension question accuracy. Post-hoc contrast tests with Holm's sequential Bonferroni procedure were performed for 15 vs. 30, 30 vs. 60, 60 vs. 120, and 15 vs. 120 kbps. Comprehension question response-time was significantly reduced when video was shown at 120 kbps vs. 15 kbps (F(1,1069)=8.50, *p*<.01). However, comparing video displayed at 15 kbps vs. 30 kbps and 15 kbps vs. 60 kbps (when averaging over frame rates) did not result in faster response-times (F(1,1071)=2.14, *n.s.*). Finally, there was not a significant frame rate × bit rate interaction (F(9,1007)=0.67, *n.s.*) for response-time. Table 7 and Figure 19 lists the mean response-time for respondents answering the comprehension questions for bit rate averaged over frame rate.

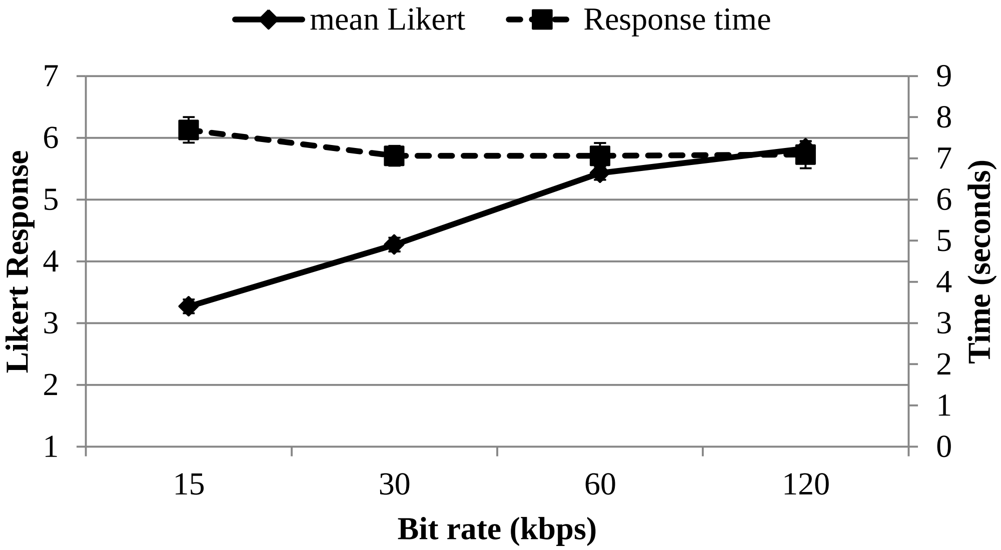


Figure 19: Plot of the mean Likert scores (higher values are better) and mean response-times (in seconds) for comprehension questions for bit rate (averaged over frame rates).

Table 7: Mean Likert scores (higher values are better) and mean response-times (in seconds) for comprehension questions for bit rate (averaged over frame rates). Bold values indicate highest mean Likert scores and fastest times to submit answer.

Bit rate*	Mean Likert	std. error	Mean Response-time (secs)	std. error
15	3.27	0.11	7.69	0.31
30	4.27	0.11	**7.06**	0.24
60	5.43	0.09	**7.06**	0.31
120	**5.83**	0.09	7.09	0.33

*averaged over frame rates

7.5 Correlation between Intelligibility and Response-Time

Frame rate was only found to significantly impact perceived intelligibility and not comprehension question accuracy or response-time. Bit rate had a significant impact on perceived intelligibility, comprehension question accuracy, and comprehension question response-time. Therefore, the following analysis focuses on establishing a correlation between perceived video intelligibility and comprehension question response-time for bit rate. A strong negative correlation (Pearson r = -.96) was found between mean Likert scores rating perceived video intelligibility and mean response-time for bit rate (averaged over frame rates). Below is the resulting regression equation for the predicted video intelligibility (I) given response-time (RT):

$$I = -2.93RT + 24.57$$

Table 8 lists the predicted intelligibility score using the regression equation.

Table 8: Predicted Intelligibility Score compared to Mean Likert score for perceived video intelligibility using regression equation where lower response-time and higher Likert and Intelligibility scores are better.

Bit rate (kbps)	Response-time (secs)	Mean Likert score	Predicted Intelligibility Score (I)	Absolute % difference between predicted and actual scores
15	7.33	3.24	3.10	4.3
30	6.81	4.28	4.62	7.9
60	**6.49**	5.43	**5.56**	**2.4**
120	6.53	5.84	5.43	7.0

7.6 Discussion

These results reaffirm that an “intelligibility ceiling effect” exists where increasing the frame rate above 10 fps (averaged across bit rates) does not provide increased perceived intelligibility. These findings also corroborate the findings from earlier work presented Chapter 6 and Cavender *et al.*’s [17] work, which is meaningful because the videos used were counterbalanced across all frame rates and bit rates, which was not done in prior work. The intelligibility ceiling effect occurs because respondents are most likely observing blurrier frames as the frame rate is increased while the bit rate is held constant. This study also reveals that bit rate has more of an effect on perceived video intelligibility, comprehension question accuracy, and response-time than does frame rate. Perhaps the slow speed of signing in the web study videos limited the discovery of a possible greater impact of frame rate on response-time.

Using the IRTM, a negative correlation between perceived intelligibility and mental effort (as measured by response-time) was established. A “mental effort floor effect” was discovered, where increasing the bit rate above 60 kbps did not produce a higher predicted intelligibility score, as shown in

Table 8. The benefit of using the IRTM and calculating the predicted intelligibility score allows for a quicker method in parameter selection. Instead of looking at subjective scores, which requires many users, or objective measures, which may not be subjectively meaningful, response-time-to-comprehension-questions offers a simpler method to evaluate perceived video intelligibility.

7.7 Frame Rate and Bit Rate Recommendations

The IRTM revealed many possible frame rate and bit rate combinations that provide intelligible video. From this study, it is recommended that video be transmitted at 10 fps and 60 kbps. The frame rate recommendation is resulting from the intelligibility ceiling effect where increasing the frame rate above 10 did not increase video intelligibility when the bit rate is held constant. The bit rate recommendation comes from the "mental floor effect," where increasing the bit rate above 60 kbps did not produce a higher predicted intelligibility score.

7.8 Summary

The IRTM was created to utilize response-time as a measure of mental effort required of viewers to evaluate videos. The IRTM was used in data analysis of this web survey evaluating four low frame rates (5, 10, 15, 30 fps) and four low bit rates (15, 30, 60, 120 kbps) in comparison to the ITU-T standard. It successfully demonstrated a negative correlation between perceived intelligibility and mental effort (as measured by response-time) for video transmitted at a constant bit rate while varying the frame rate. Mainly, response-time decreased as perceived intelligibility increased. Secondly, the IRTM informed selection of the lowest transmission rates that did not negatively impact video intelligibility. Finally, the findings from this study recommend a frame rate and bit rate pair that is lower than the recommended ITU-T standard.

These findings suggest that the ITU-T standard can be relaxed, especially when considering limited resources such as total bandwidth, network congestion, and mobile phone battery life.

Chapter 8 Laboratory Study: Effects of Lower Video Transmission Rates

Up until now, multiple web studies have been created to evaluate perceived video intelligibility of mobile sign language video transmitted at frame rates, bit rates, and spatial resolutions lower than the recommended ITU-T standards (at least 25 fps and 100 kbps) for reducing total bandwidth consumption and increasing battery life. Findings from the web studies (Chapter 6 and Chapter 7) suggest an "intelligibility ceiling effect," where increasing the frame rate above 10 fps and bit rate above 60 kbps does not significantly improve perceived video intelligibility. The subjective responses to rating perceived video intelligibility suggest that intelligible video transmitted at these lower transmission rates can facilitate intelligible conversations on mobile devices. The next step in this dissertation research is demonstrating that intelligible conversations can occur among fluent ASL signers using an experimental smartphone application with the lowered frame rate and bit rate settings implemented.

In a continued effort to reduce total bandwidth consumption and extend battery life for mobile sign language video telephony, I conducted a laboratory study, where fluent ASL signers in pairs held free-form conversations over an experimental smartphone app transmitting real-time video at (5 fps, 25 kbps), (10 fps, 50 kbps), (15 fps, 75 kbps), and (30 fps, 150 kbps). The objectives of this study were: (1) to identify the minimum video quality settings allowable for intelligible sign language communication; (2) to learn what adaptation techniques participants use to compensate for the lowered transmission rates; (3) to objectively measure user perceived intelligibility of video content used in mobile sign language conversations; and (4) to quantify how much battery life is extended. Results from the laboratory study also demonstrate that intelligible conversations can occur at transmission rates lower than the ITU-T standard.

8.1 Technology Used

Mobile Phone

The Samsung Galaxy S3 smartphone was used to run an open source video chat software app called IMSDroid[1], whose encoder was modified to transmit video at 5, 10, 15, and 30 fps. The bit rate averaged 5 kb/frame, resulting in the bit rate increasing as the frame rate increased, namely 25, 50, 75, and 150 kbps, respectively. The spatial resolution of the video transmitted was at 320×240 pixels displayed horizontally on the phone to maximize the screen size. Prior to the selection of the Samsung Galaxy S3 phone, the Sprint EVO, Samsung Galaxy S2, Samsung Galaxy S4, HTC One, and Google Nexus Phone 4 were investigated as alternatives, but each of these phones' encoders failed to allow for the lowered frame rates. Only the Samsung Galaxy S3 encoder was compatible with the IMSDroid frame rate modifications and thus, the Galaxy S3 was selected for the laboratory study.

IMSDroid

IMSDroid is an open source video conferencing application running on Doubango [55], a 3GPP IMS/LTE (IP Multimedia Subsystem) framework for embedded systems. IMSDroid is a Java-based front-end to Doubango, which is open source VoIP client that references implementation to the Doubango framework. IMSDroid has a GUI interface allowing for both audio and video calls with the robustness of selecting different video encoder. Doubango is the backend framework running 3GPP IMS/LTE which can run many different types of protocols like SIP/SDP, HTTP/HTTPS, and DNS. In this study, the Session Initiation Protocol (SIP) was selected for the VoIP.

[1] http://doubango.org/. Accessed on May 9, 2012.

Asterisk Server

An Asterisk [9] server was set up as the communication server for the laboratory study. Asterisk is an open source framework that supports the server side of facilitating Voice over Internet Protocol (VoIP) video communication, where we used the Session Initiation Protocol. A specific configuration file was modified to regulate the bit rate at which video was transmitted, specifically averaging 5 kb/frame. Asterisk uses User Datagram Protocol, which is suitable for fast efficient transmission of data for video conversations.

Unobtrusive Logging

Network traces were conducted on the Asterisk server monitoring the frame rate and bit rate at which video was transmitted for each video call. The battery drain of each phone was also unobtrusively logged on the mobile device using an open source mobile application called AndroSensor [5]. AndroSensor logged the battery life percentage every 30 seconds.

8.2 Participants

Social media and email listservs were used to recruit fluent ASL signers to participate in the study. Participant inclusion criteria included: (1) deaf and/or hard-of-hearing people for whom ASL is the primary language; (2) hearing people who fluently sign ASL (over 5 years of signing experience); and (3) people 18 years old or older. Participants received a $25 gift card upon completing the 75-minute laboratory study. Those who responded to the e-mail were either paired with a random person to sign with or brought a friend fluent in ASL. Demographic questions asked in the laboratory study (described below) were used to further ensure language fluency.

The laboratory study had 20 participants (11 women), all of whom fluently signed ASL. Their age ranged from 26-74 years old (median=48.5 years, SD=13.5 years). Of the 20 participants, 18 were deaf (2 of 18 wore hearing aids) and 2 were Children of Deaf Adults with full hearing. Eight participants were randomly assigned to their signing partner (4 sessions) and the other participants were paired with a friend (6 sessions). Thirteen participants indicated that ASL was their daily language, and the number of years they had spoken ASL ranged from 26-74 years (mean=47 years, SD=13 years). All but one participant owned a smartphone and everyone had sent text messages; 19 participants indicated they use video chat; and 17 use video relay services.

8.3 Study Design

8.3.1 Apparatus

Participants sat on the same side of a table with a black drape behind them. They were separated by a board. Two phones were propped up with a business card holder and placed, one each in front of the participants. Participants were told to adjust the location of the phone for comfortable conversation. Figure 20 is a photo of the experimental setup.

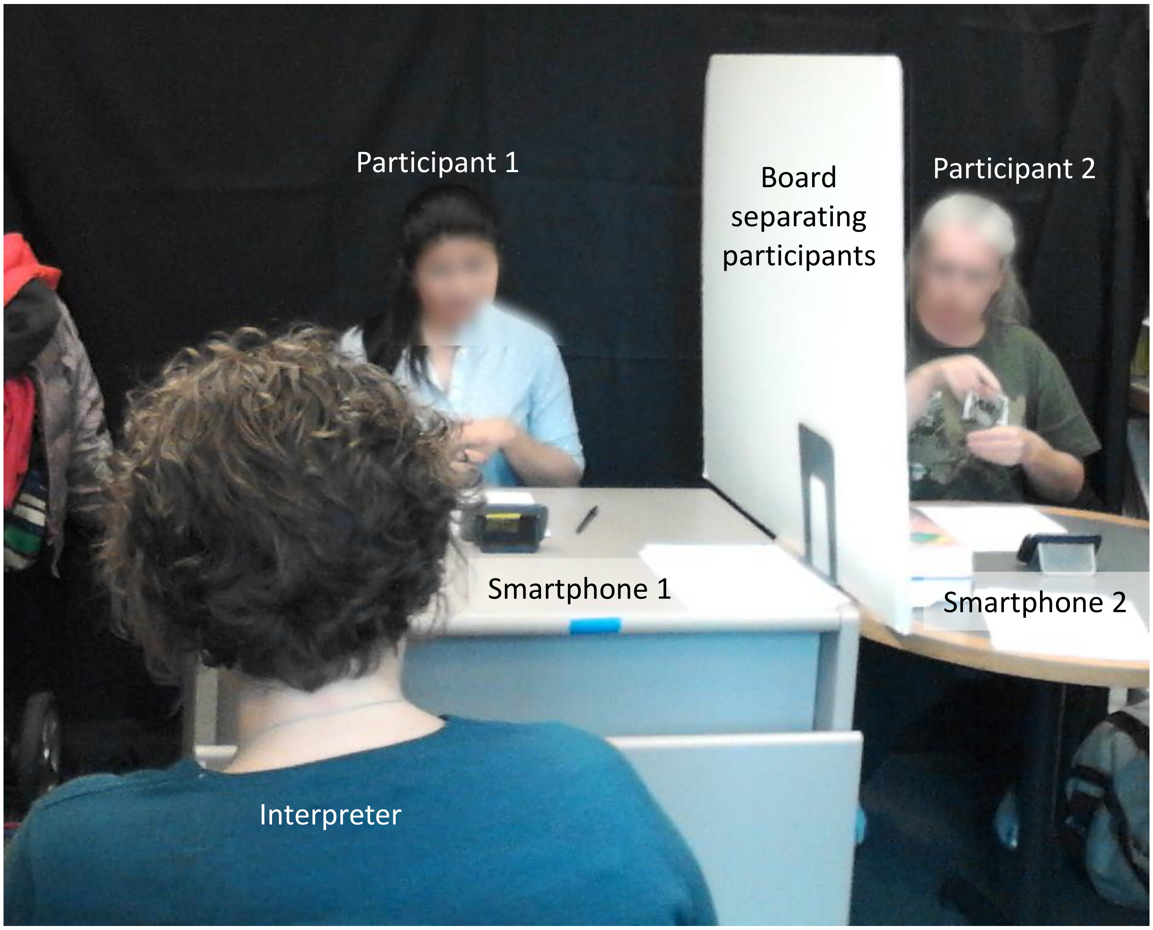


Figure 20: Experimental setup with two participants separated by a board. A certified ASL interpreter was always present.

8.3.2 Conversation Task

Participants were instructed to hold five, 5-minute free-form conversations over the smartphones. The first conversation was a practice round for participants to familiarize themselves with the phone and available signing space. Participants were instructed to talk about whatever they liked, but for each subsequent conversation, they were asked to discuss a different topic than the conversation before. After each session, participants filled out a paper questionnaire, described below. All participants were video recorded during the study. The smartphone did not record conversations. A randomized Latin Square was used to assign the order in which video frame rate was used on IMSDroid. Participants were not told how the video quality was altered, only that they were using different versions of the smartphone app. A certified ASL interpreter was present during all study sessions and facilitated communication between the study participants and myself, who conducted the studies.

8.3.3 Subjective Measures

Participants were asked to fill out a subjective questionnaire after each 5-minute conversation. The questions are listed below and respondents circled the response that best answered the question.

- Question 1: How easy was it to understand the video?
  (7-point Likert scale ranging from very easy to very difficult)

- Question 2: Rate the video quality for sign language.
  (7-point Likert scale ranging from excellent to poor)

- Question 3: Rate the video quality for fingerspelling.
  (7-point Likert scale ranging from excellent to poor)

- Question 4: Rate the video quality for lip reading.
  (7-point Likert scale ranging from excellent to poor);

- Question 5: During the conversation, indicate how often you had to guess what the other signer was signing.
  (0% never, 25% sometimes, but not often, 50% half the time; 75% most of the time, and 100% all of the time).

After all trials were completed, participants filled out a demographic questionnaire which included questions such as, “how long have you been signing ASL?’; “what language do you prefer to sign with family?”; and, “do you own a smartphone?” Lastly, participants were asked exit interview questions regarding their overall experience while signing over the different frame rates and bit rates. Examples of questions asked included, “did you notice changes in video quality?"; "at any time were you frustrated with the video quality provided?"; and, "would you use the lower video quality if you knew you could save battery life?"

8.3.4 Objective Measures

A conversation with low intelligibility may contain a lot of requests for repetitions, called "repair requests" [110], and "conversational breakdowns," where a signer may sign the equivalent of, "I didn't understand what you said," or give up. Also, the rate of signing may decrease with the lowered frame rate. Therefore, we analyzed the rate of fingerspelling. Fingerspelling occurs when a signer spells out the name of something, which is usually for titles, proper names, and technical words. Signs that are lexicalized "loan signs," which are common words that have become the stylized fingerspelling, are not counted in our fingerspelling measure.

The objective measures were the number of repair requests, average number of turns associated with repair requests, number of conversational breakdowns, and speed of fingerspelling. These measures were calculated from the videotaped sessions with the assistance of a certified ASL interpreter. A repair request in a signing conversation may include signing "what?" or "again." For each repair request, the number of turns was counted until the concept was understood. Conversational breakdowns were counted as the number of times the participant signed the equivalent of "I can't see you" due to the video being blurry, choppy, or frozen. An unresolved repair request was also counted as a conversational breakdown. Finally, the speed of fingerspelling was measured as the time it took to sign each letter of the word, divided by the number of characters in that word minus 1, producing the characters per second.

8.4 Results

8.4.1 Perceived Intelligibility

Nonparametric analyses were used to analyze each question, which captured responses on 7-point Likert scales. Since data gathered were ordinal and dichotomous responses, a Friedman test [40] was used to analyze the main effect of bit rate and spatial resolution on comprehension. Separate pairwise Wilcoxon tests [113] with Holm's Sequential Bonferroni procedure [51] were performed to investigate the effect of frame rate. Results will be reported for each question.

Question 1 asked participants to rate how easy it was to understand the video from 7-very easy to 1-very difficult. The Friedman test did not indicate a significant main effect of frame rate on perceived video intelligibility ($\chi^2_{3,N=20}$=5.08, *n.s.*).

Question 2 asked participants to rate the video quality for sign language communication from 7-excellent to 1-poor. The Friedman test indicated a significant main effect of frame rate on perceived video quality ($\chi^2_{3,N=20}$=11.01, *p*<.05). Wilcoxon tests with Holm's Sequential Bonferroni procedure were performed to identify the effect of frame rate on perceived video quality. Increasing the frame rate from 5 fps vs. 10 fps, 15 fps, and 30 fps, respectively, was found to increase perceived video quality ($\chi^2_{3,N=20}$=46.5, *p*<.05). However, comparing perceived video quality between 10 fps vs. 15 fps vs. 30 fps was not found to significantly increase perceived video quality ($\chi^2_{3,N=20}$=9.0, *n.s.*).

Question 3 asked participants to rate the video quality for fingerspelling from 7-excellent to 1-poor. The Friedman test indicated a significant main effect of frame rate on perceived video quality for fingerspelling ($\chi^2_{3,N=19}$=8.11, *p*<.05). Wilcoxon tests with Bonferroni procedure were performed to identify the effect of frame rate on perceived video quality for fingerspelling. Increasing the frame rate from 5 fps vs. 10 fps, 15 fps, and 30 fps, respectively, was found to increase perceived video quality ($\chi^2_{3,N=20}$=35.5, *p*<.05). However, comparing perceived video quality between 10 fps vs. 15 fps vs. 30 fps was not found to significantly increase perceived video quality for fingerspelling ($\chi^2_{3,N=20}$=10.0, *n.s.*).

Only half of the participants indicated that they lip read during signing. Therefore, analysis for question 4, which asked participants to rate the perceived video quality for lip reading from 7-excellent to 1-poor, was performed for 10 participants. The Friedman test did not indicate a significant main effect of frame rate on perceived video intelligibility for lip reading ($\chi^2_{3,N=10}$=2.92, *n.s.*).

Question 5 asked participants to rate how often they had to guess what the signer was signing during their conversation (0% never, 25% sometimes, but not often, 50% half the time; 75% most of the time, and 100% all of the time). The Friedman test indicated a significant main effect of frame rate on the rate at which participants had to guess what their signing partner was signing ($\chi^2_{3,N=20}$=29.75, *p*<.0001). Wilcoxon tests with Bonferroni procedure were performed to identify the effect of frame rate on participants guessing what the other signer was signing. Increasing the frame rate from 5 fps vs. 10 fps, 15 fps, and 30 fps, respectively, was found to decrease how often a participant had to guess what the other signer was signing ($\chi^2_{3,N=20}$=52.5, *p*<.001). However, comparing how often a signer had to guess what their partner was signing for video transmitted between 10 fps vs. 15 fps vs. 30 fps was not found to significantly reduce how often they guessed what the other person was signing ($\chi^2_{3,N=20}$=6.0, *n.s.*).

8.4.2 Objective Measures

All sessions were video recorded to be objectively analyzed in post-analysis with a certified ASL interpreter. Each conversation was analyzed to identify and count instances of (1) repair requests during a conversation; (2) conversational breakdowns; and (3) speed of fingerspelling (reported as characters per second).

A Friedman test was performed for each objective measure to determine how varying the frame rate affected it. Frame rate was found to significantly impact the number of repair requests ($\chi^2_{3,N=10}$=11.0, p<.05) and the number of conversation breakdowns made during a conversation ($\chi^2_{3,N=10}$19.8, p<.001); however, varying the frame rate was not found to statistically significantly impact the speed of fingerspelling ($\chi^2_{3,N=10}$=2.48, *n.s.*). Table 9 lists the number of instances of fingerspelling and the average characters signed per second at each frame rate.

Table 9: Count of the number of fingerspelled words and the average, max, min, and standard deviation of the number of characters signed per second.

frame rate/bit rate (fps/kbps)	5/25	10/50	15/75	30/150
Total count of finger spelled words (over all sessions)	153	191	166	180
average characters/sec	4.08	4.16	4.03	4.29
SD of characters/sec	1.99	2.03	1.45	1.97

As Table 9 demonstrates, the average number of characters per second did not change as the frame rate increased, even though participants perceived changes in video quality. Perhaps participants adapted quickly to the temporal video quality or used alternative methods, which are discussed further below.

Sign language conversations held over video transmitted at 5 fps received the most repair requests and conversational breakdowns, as expected. Video transmitted at 10, 15, and 30 fps did not have any instances of repair requests or conversational breakdowns across all sessions. Figure 21 lists the number of repair requests and conversational breakdowns that occurred for each session for 5 fps/25 kbps.

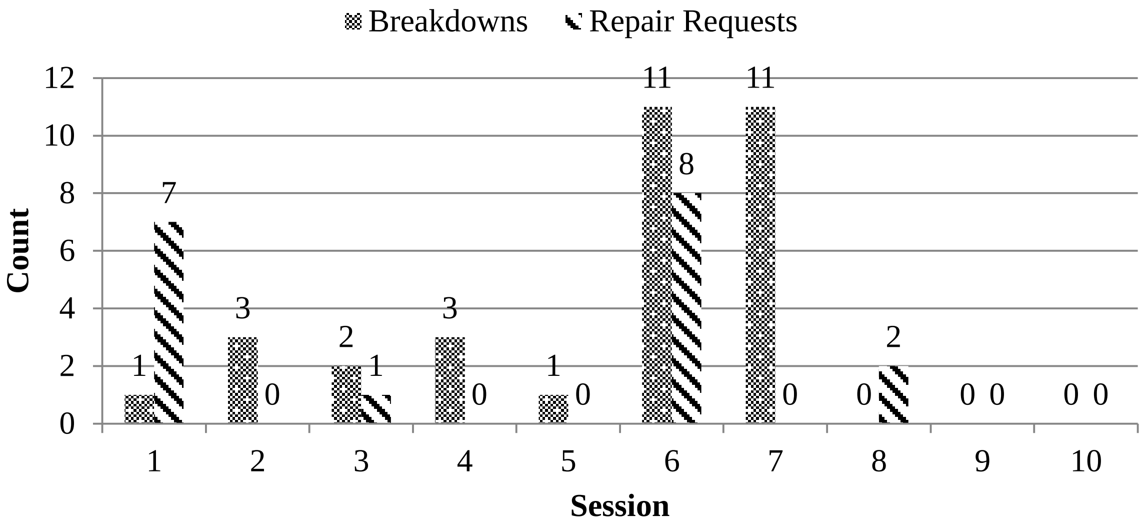


Figure 21: Count of conversational breakdowns and repair requests that occurred for each session when video was transmitted at 5 fps/25 kbps.

Figure 21 shows that sessions 6 and 7 received the highest counts for conversational breakdowns with 11 total breakdowns occurring in a 5 minute conversation. Participants in sessions 4, 5, 6, 7, 8, and 9 were friends while the other sessions had participants paired with strangers.

8.5 Exit Interviews

During the exit interviews, participants were asked to indicate which version of the video app they preferred to use. There were four recurring themes that arose during the exit interviews, which were: (1) noticeably lower quality of video transmitted at 5 fps; (2) desire for larger screens; (3) different adaptation techniques used to compensate for lower video quality; and (4) comparison of video quality used in the experimental app to commercially available apps.

8.5.1 5 FPS Video Quality

All participants voiced their observations that video transmitted at 5 fps was noticeably more “choppy” or “frozen” than other versions of the app that they used. When asked what they liked or disliked about signing at 5 fps, many participants said they “would not want to use the video at all.” P3 signed that she really could not express herself like she normally would (when signing to someone in-person) because of the lower video quality. P13 and P14 said they chose to have a “lighter conversation,” *i.e.*, not talk about anything that required a lot of background information to be signed first. They were unsure how often they would need to repeat themselves so they wanted to keep the conversation short.

Many participants signed that they would not use mobile video communication at 5 fps, even though the video quality provided intelligible content. When asked if they would “give up” signing to each other at video transmitted at 5 fps, participants expressed that they probably would turn to texting to clarify what they wanted to say since texting is more reliable than mobile video at 5 fps. P17 and P18 said they would rather text message instead of sign over video transmitted at 5 fps. When asked why, they said because more energy was needed to repeat themselves over video, while texting required only one message. P17 did acknowledge that texting was asynchronous, but believed texting was more reliable than current mobile video apps. P18 followed up by saying she didn’t use mobile video chat on her phone, so texting was her solution for mobile communication.

8.5.2 Desire for Larger Screens

During the exit interviews, many participants spoke about the form factor of the device, specifically desire for larger screen sizes. P13 and P14 made comments that they preferred to sign over a larger device with a bigger screen similar to the screens available on the iPad or Samsung Galaxy Note. P14 expressed she did not feel like she could express everything she wanted to say because of the confined signing space. Also, the angle at which video was shown made it more difficult to understand her signing partner. Mainly, the hands were closer to the screen, but the signer's head appeared to look like a "pin head" because of the camera angle. P14 also said that lip reading was hard to do because of the "pin head" appearance of her signing partner.

8.5.3 Adaptation Techniques

When participants were asked what adaptation techniques they used to compensate for the lower video quality, a majority of the participants said they deliberately fingerspelled more slowly than their regular signing speed. They also had to ask their signing partner to repeat what was signed and slow down whatever they were signing. Some participants also said doing this often disrupted what they were trying to say, which caused some frustration for both the signer and receiver. Interestingly, as results from section 8.4.2 showed, participants did not actually sign more slowly when the frame rate varied (mean characters per second: 4.97 at 5 fps vs. 5.22 at 30 fps), as listed in Table 9, even though they were perceived to sign more slowly.

When participants were asked which version of the video app they preferred to use, many participants indicated they preferred signing over video transmitted at 15 and 30 fps; however, many participants indicated that they could not tell the difference between video transmitted at 15 fps and 30 fps. When asked about video transmitted at 10 fps, participants did say it was better than video transmitted at 5 fps, but not as good as video transmitted at 15 or 30 fps.

8.5.4 Comparisons to Commercial Video Apps

In many of the laboratory sessions, participants compared the video quality they were using to commercially available apps like Skype and FaceTime. Those participants who referred to FaceTime said that FaceTime's video quality was clearer and smoother. This particular comment was expected since FaceTime transmits video at 30 fps at 1-3 Mbps at 960×640 screen resolution [76]. In one of the sessions, P7 and P8 were signing over video transmitted at 15 fps and began to discuss how IMSDroid's video quality compared to FaceTime:

> *P7: How does this compare to FaceTime?*
>
> *P8: FaceTime is more clear, but this is fine… your hands are a little more blurry [using this app]. I understand you fine though.*
>
> *P7: Am I signing too fast?*
>
> *P8: No, you're signing fine.*
>
> *P7: Well...I'm signing normal, just trying to test the limitations. Is the finger spelling clear?*
>
> *P8: Yeah, I can see you fine.*
>
> *P7: So when I spelled 'ameba'*
>
> *P8: Yes, ameba*
>
> *P7: Did you see all the signs or did you just catch the 'b' 'a'?*
>
> *P8: ...I saw the full spelling, but deaf understand what you're saying anyhow. We're used to doing that.*

This snippet of conversation is an example of how people who are deaf naturally interpolate what they view to understand the overall message of a conversation. For instance, when words are fingerspelled, all the letters of the word may not have been viewed by the receiver, but the word can be discerned from the context of the conversation.

8.6 Battery Drain

The battery drain was unobtrusively logged using an open source app called AndroSensor, which ran in the background and logged the percentage battery drain every 30 seconds for each 5 minute conversation. Data were collected from the phones after each session for later analysis.

The rate at which the battery percentage depleted was calculated for each 5 minute video call. We verified that the battery drain was linear, which allowed us to use linear regression to model the data. The estimated average battery duration for each frame rate was calculated for every conversation and shown in Figure 4. As anticipated, the higher the frame rate at which video was transmitted, the higher the rate at which the battery drained. We found that the Samsung Galaxy S3 has an average battery life of 1000 minutes in standby mode and an average battery life of 750 minutes if IMSDroid was "active" but not transmitting video.

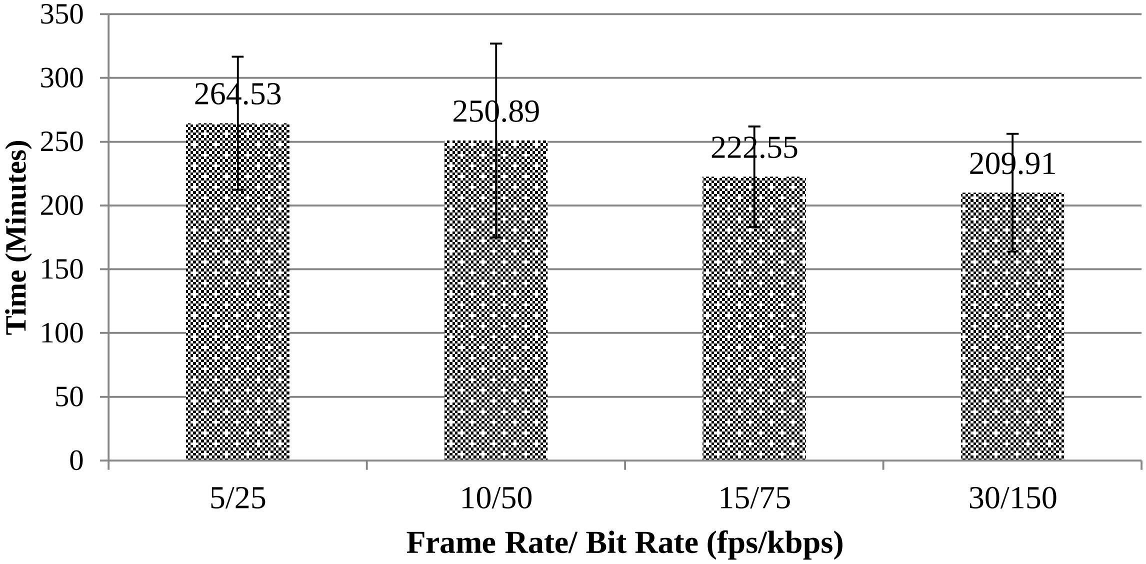


Figure 22: Estimated average battery life (in minutes) for sign language video transmitted on IMSDroid at each frame rate/ bit rate.

8.6.1 Bandwidth Consumption

Network traces were performed on the Asterisk server to monitor the average rate at which data was transmitted. Bit rate control is an active area of research [21, 36, 84, 95] and was not the focus of this study. Table 10 lists the average bit rate at which video was transmitted for each frame rate. The bit rate was controlled by the Asterisk server and the network traces confirmed that the frame rate dictated the bit rate at which video was transmitted.

Table 10: Average, min, max, and SD of the bit rate when varying the frame rate as captured by the network traces.

frame rate (fps)	average bit rate (kbps)	Min bit rate (kbps)	Max bit rate (kbps)	SD (kbps)
5	23.89	20.87	32.19	3.38
10	50.00	39.78	67.76	8.67
15	73.04	64.43	91.25	8.67
30	129.89	114.78	147.38	9.91

8.7 Discussion

Participants were successful at holding intelligible conversations across all frame rates. All participants did notice and complain about the lower quality of video transmitted at 5 fps; however, participants' rate of fingerspelling did not decrease, even though they perceived their signing speed to be slower. Video transmitted at 5 fps had more instances of conversational breakdowns and repair requests. Sessions 6 and 7 received the most counts for conversational breakdowns (11 instances); the frequencies at which breakdowns occurred were low across other sessions. Closer inspection of the conversations in which the breakdowns and repair requests occurred revealed that the topic of conversation was very detailed and required more explanation. For example, P11 and P12 from session 6 were talking about a trip to Iceland. P12 asked if P11 was going to see the Aurora Borealis. It took multiple attempts by P11 asking the question to clarify what P12 was asking. The frame rate at which the video was signing was 10 fps. The conversational breakdown could have resulted from the conversation topic and not because of the video transmission rate.

8.7.1 Signing Adaptation Techniques

Signers are versatile when it comes to adapting their signing to the technology they use to communicate. The context of a conversation, signs used, loan signs (signs that represent an English word that has developed a unique movement), and fingerspelling words all assist in filling in missing information [14]. Signers may be naturally taking advantage of the "word superiority effect" where people are more successful recognizing letters presented within words than just isolated letters [13]. This may explain why the rate of fingerspelling did not vary across the frame rates.

During objective analysis of the video conversations, there were instances in which a participant would begin to finger-spell a word; however, she did not spell every letter within that word. For example, a participant was talking about the different seasons, but when she fingerspelled "season," she only signed "s" and "n" of the word. The receiver of the message was still able to infer the word. The receiver may also have been able to infer the word from the context of the message. Often the context of a conversation can aid in understanding a word that was not seen during the conversation [85].

8.7.2 Willingness to Use Lower Video Quality

When asked if they were willing to use a low video quality to hold conversations, all participants said they would be willing to use the mobile technology if there were a guarantee that video would be transmitted at 15 fps or 30 fps. However, video transmitted at lower frame rates would only be used for very short conversations, such as asking a quick question. When given the option between texting and mobile video chatting, participants said they always would prefer to sign over video; however, if the person they are communicating with does not sign, texting is considered necessary.

8.7.3 Technology Position Adjustments

Participants were allowed to adjust the mobile device to a position that felt comfortable. Some of the participants adjusted the phone to increase the angle at which it was displayed or raised the phone to increase their signing space. Figure 23(a) shows the original position of the phone placed in front of the participants. Figure 23(b) shows how a participant placed a pen behind the phone to increase the angle at which he viewed the phone. Figure 23 (c) and (d) are two different examples of how participants requested to use stacks of books located in the room to raise the smartphone's position.

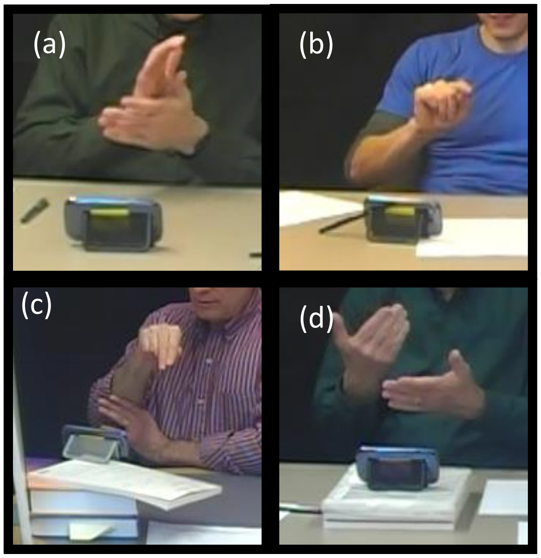


Figure 23: Four examples of how participants adjusted the phone position. (a) Original phone setup using a business card holder. (b) Phone propped up with a pen. (c) Increased height and viewing angle. (d) Increased height from table.

8.7.4 Recommendations

As anticipated, reducing the frame rate at which sign language video is transmitted increases the average battery life of IMSDroid. From the laboratory results, it is recommended that conversational video be transmitted at 10 fps to best balance resource consumption, video intelligibility, and user preferences. Transmitting video at 10, 15, and 30 fps received, on average, the same subjective responses from participants when asked to rate how easy it was to understand the video; rate the video for picture quality, fingerspelling, and lip-reading; and how often the signer had to guess what the other person was signing. While the battery life lasted the longest when video was transmitted at 5 fps, video transmitted at 5 fps also received the most counts for repair requests and conversational breakdowns. Finally, in the exit interviews, participants voiced their dissatisfaction of communicating at video transmitted at 5 fps because of the choppy video quality. Although some participants were able to tell that there was a difference between video transmitted at 10 fps vs. 15 fps vs. 30 fps in the exit interviews, both the subjective and objective results support that video can be transmitted at 10 fps, which is the lowest threshold at which intelligible sign language conversations can be comfortably held.

8.8 Summary

The ITU-T standard recommends that video should be transmitted at least at 25 fps and 100 kbps for intelligible conversations. My laboratory study clearly demonstrates that there is a lower limit at which intelligible mobile sign language video can be transmitted. My findings suggest that video transmitted at 10 fps with a bit rate averaging 50 kbps can facilitate intelligible sign language conversations, and can extend battery life by almost 20% compared to transmitting at 30 fps and 150 kbps.

The findings from this study provide the motivation for the creation of video technology specifically designed for use during emergencies and natural disasters, where the full cellular network infrastructure may become unavailable. In 2005, it was estimated that 50% of the total phone lines and wireless subscribers lost access to phone service for multiple days after Hurricane Katrina hit land [87]. In the laboratory study, people were still successful at holding intelligible conversations at 5 fps (averaging 23.89 kbps) even though participants did not prefer communicating at those video transmission rates. Having the capability to transmit emergency videos, even at these low transmission rates, would be useful to relay important information.

Chapter 9 Field Study: Using the MobileASL Application in the Wild

Both web and laboratory studies have limitations when it comes to simulating the environment in which real-time mobile video communication occurs. The benefit of web studies when evaluating video intelligibility is the ability to research hundreds of people. Results from the web studies have consistently demonstrated the intelligibility ceiling effect holds true, where increasing the frame rate above 10 fps when the bit rate is held constant does not significantly increase perceived video intelligibility. The web studies have also demonstrated that response-time can be used as another measure of video intelligibility, especially when creating studies based on the Intelligibility Response-Time Method.

All the findings from the web studies aided in synthesizing the specific video transmission parameters that were investigated in the laboratory study. Specifically, I investigated varying the frame rate in which video is transmitted; holding the spatial resolution constant; and allowing the bit rate to vary with the frame rates investigated. The laboratory results demonstrated that intelligible conversations were successfully held at frame rates as low as 5 fps and bit rates averaging at 25 kbps; however, participants experienced more conversational breakdowns and repair requests. Therefore, the recommended lowest video transmission rates at which intelligible real-time conversations can occur are frame rates of 10 fps and bit rates averaging 50 kbps.

A final component of this dissertation investigates how real-time mobile sign language video is used in the wild. In 2010, my colleagues and I conducted a three week pilot field study using the MobileASL app for the HTC TyTNII cell phone to investigate how MobileASL is used in everyday communication. During this study, FaceTime was not readily available for everyday use.

9.1 Technology Used

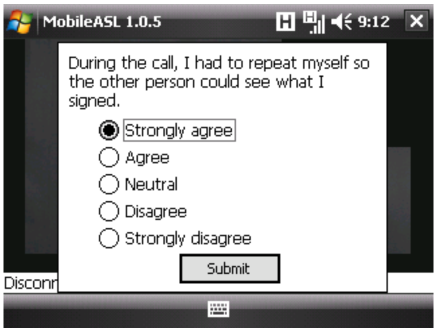
Chapter 2, section 2.4.1 described the implementation of the MobileASL software application for the Windows Mobile 6.1 phone. For this study, the MobileASL software was selected for evaluation because the software had the capability to control the frame rate and bit rate at which video was transmitted (averaging 10-12 fps at 30 kbps). The HTC TyTNII cell phone was used and pre-loaded with MobileASL with access to an unlimited AT&T data plan.

Unobtrusive Logging

Information about the phones' usage was unobtrusively logged in the background. Specifically, the battery life usage, number of video calls made, changes in IP address, and how long MobileASL was turn actively 'on' were recorded.

Experience Sampling

In-the-moment use of MobileASL was gathered using experience sampling [60], where a brief multiple-choice question appeared on the phone screen after a behavior trigger occurred. Experience sampling is a research method asking participants to provide a short response due to a behavioral trigger, time event, or contextual cue [3]. When certain events occurred in the MobileASL app, a multiple-choice question would appear on the screen and ask the user about the event that just occurred. Figure 24 is a screenshot of an experience sampling question.



During the call, I had to repeat myself so the other person could see what I signed.

- ● Strongly agree
- ○ Agree
- ○ Neutral
- ○ Disagree
- ○ Strongly disagree

Figure 24: Example of an experience sampling question appearing in MobileASL.

There were six triggers that caused an experience sampling question to appear: after a call; after declining a call; after a missed call; after a short call (< 30 seconds); after a call that used privacy; and after a change in IP address. There was a five question “quota” per day of experience sampling questions asked. If the quota was not met the previous day, the current day’s total would be added on from the prior day.

Participant Recruitment

At the time of the study, the University of Washington held a summer academy for advancing deaf and hard-of-hearing students in computing. The students from this program were ideal candidates because they were fluent in ASL or Pidgin Signed English (PSE). Also, I wanted to recruit participants who were tech savvy, owned mobile devices, and would be willing and interested in communicating with other participants using the MobileASL app. To recruit participants for the study, I gave an hour presentation about the MobileASL project, my research goals, and finally, invited the students to participate in a three week field study to evaluate the mobile application.

9.2 Study Procedure

The goal of the study was to learn how MobileASL is used in everyday life and how it influences mobile communication. The study was three weeks in duration and consisted of a pre-deployment questionnaire, which included asking background and demographic questions of the participant such as “what language do you prefer to communicate with friends, family, on a daily basis?”; “what year did you start texting?”; and “have you used a videophone?” There was also a weekly online survey asking participants about their overall use of MobileASL for that week. Two interviews were held: the first two days after MobileASL was handed out, and the second two days before the study ended. The purpose of the first interview was to learn about their current use of technology to communicate and overall impression using MobileASL. The purpose of the second interview was to learn about their usage of and satisfaction with MobileASL. A focus group was held at the end of field study to learn about overall impressions using MobileASL. All the questions used in the field study are listed in Appendix D.

Task

Participants were instructed to make as many calls as they could per day during their free time and when they were not in class. Recall that this field study was independent of the summer academy and participation was voluntary. Participants were not required to make a minimum number of video calls per day.

9.3 Results

Demographic Information

There were 11 participants (3 women) whose ages ranged from 16-23 years old (median=17, SD=2.01). Of the 11 participants, their preferred language to communicate on a daily basis consists of ASL (3), Pidgin Signed English (3), and English (5). Seven of 11 participants self-reported that they were deaf or hard-of-hearing. All participants own a mobile phone and text messaged. Only 5 participants were familiar with video phones and have used them to communicate with others.

9.3.1 Unobtrusive Logging

Over 300 phone attempted video calls were made using MobileASL. Calls tended to be short with varied duration (mean=105.1 sec, SD=158.6 sec). It was evident that the novelty of mobile video communication contributed to the initial high volume of calls made the first two days. Each participant made on average 0-2 calls a day, except the first day of the study, where each participant made on average 30 calls, and the second day, on average 7 calls. Figure 25 lists the total number of calls made each day after removing days 1 and 2.

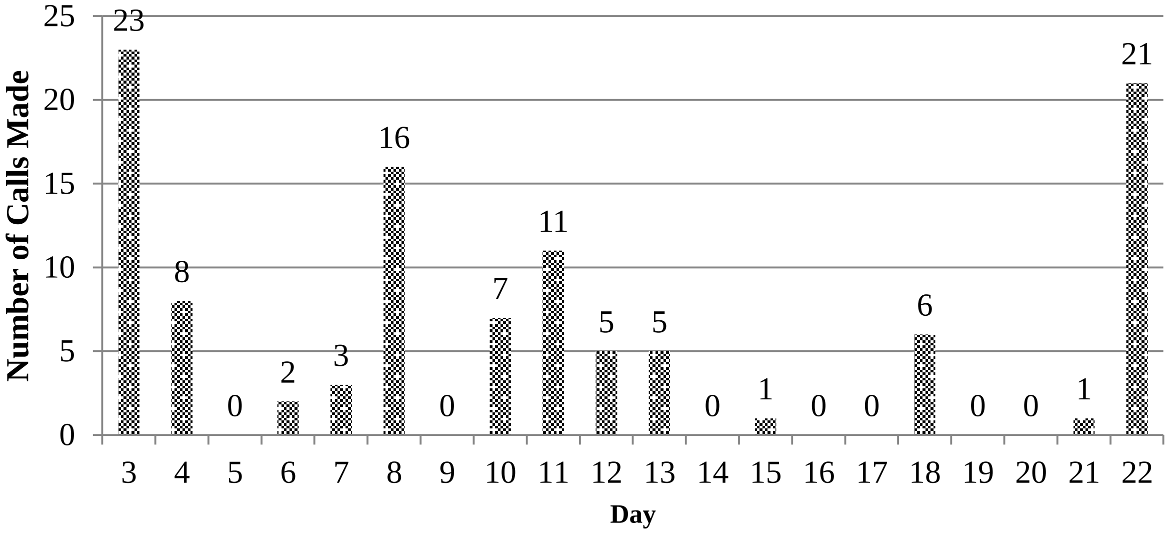


Figure 25: Total number of calls made each day excluding day 1 and 2.

9.4 Experience Sampling

There were not many experience sampling responses even though a quota was placed for the number of experience sampling questions to appear each day. Since participants did not make many phone calls each day, many of the experience sampling triggers did not occur. However, the experience sampling question “Which best describes where you are right now?” received many responses. Responses from this experience sampling question may suggest that participants utilized the mobility aspect of MobileASL since many video calls were made at school, in a public place or business, and other locations. Figure 26 demonstrates the distribution of responses for this particular experience sampling question.

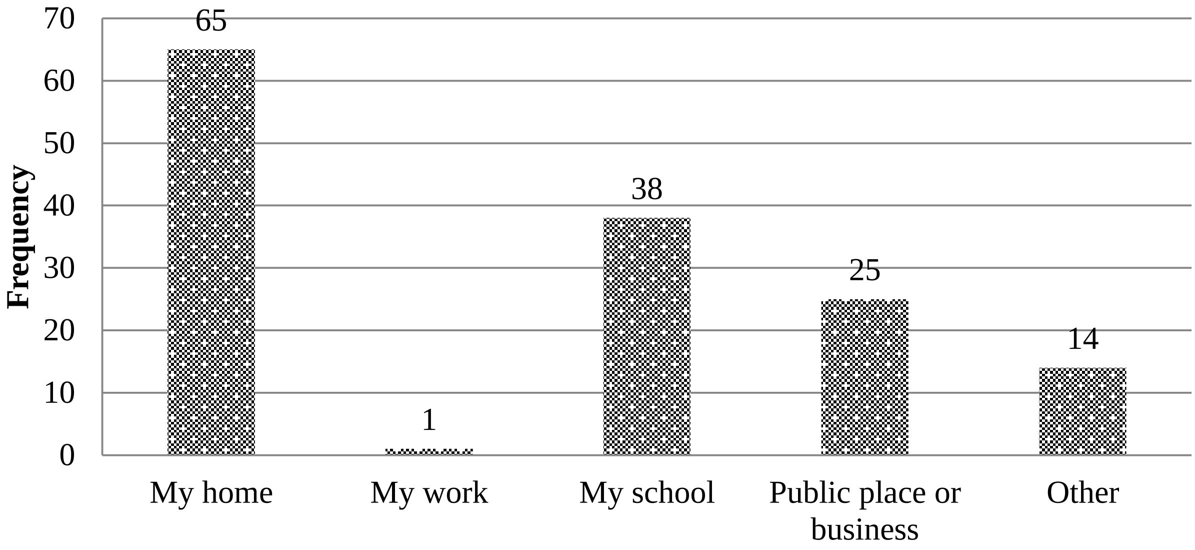


Figure 26: Distribution of responses for the experience sampling question, "Which best describes where you are right now?"

9.5 Interviews

Participants expressed overall positive experiences using MobileASL and in many ways found the technology preferable to existing stationary videoconferencing technologies (computer video chat programs and videophones) or texting. In the interviews, participants were asked what they liked and disliked about communication methods they already used—texting and stationary videoconferencing. Participants described texting as quick, easy to use and nearly always available. However, participants said they found it easy to misunderstand text messages. Stationary videoconferencing was said to provide more cues for communication because it is visual and interactive. However, although it allows real-time sign language conversation, participants pointed out the need to be in a specific place to use it.

When participants were asked about what they liked and disliked about MobileASL, they reported liking having a visual aspect to their mobile communication; not only were they able to see each other's expressions and reactions, but they were able to show what they were talking about to the other person. For example, when two participants became separated from each other while shopping, they used MobileASL to show each other landmarks and eventually reconnected.

9.5.1 Focus Group

A 45 minute focus group was held at the end of the study with all of the participants to learn more about their overall experience using MobileASL. Participants were open to comment on any aspect of using MobileASL and their thoughts on mobile video communication. It was evident that the form factor of the HTC TyTNII mobile phone and the short battery life influenced how often mobile video calls were attempted. Many participants compared the TyTNII phone to their personal mobile devices. For instance, people disliked that they needed to carry two phones to make a phone call. They did like that the TyTNII had the capability of being "flipped out" so the phone could be set on a table without needing extra equipment to prop up the phone.

The phones' battery life influenced how frequent calls were made. Many participants expressed disliking the short battery life because it limited their ability to call others. One participant said (English text transcribed from interview conducted in ASL):

> *The problem is that when I turn the program off in order to save the battery, then no one can call. It's almost like I would have to keep turning it on just to check and see if anyone else is on, and if not, turn it back on again. If someone else happens to be on at the same time, I can take advantage of the opportunity to chat…but that's just luck.*

MobileASL was considered very useful when needing to gather information in a moment’s notice and texting was considered not fast enough. For example, a group of participants became lost while riding the bus to the mall. Using MobileASL, they called another participant to ask for directions. Participants said that MobileASL was much better than texting in these cases because it would take a long time to describe the situation using text and to wait for a reply. With MobileASL, participants were able to immediately receive and convey information.

When asked about the potential of mobile video communication becoming a more mainstream method for them to communicate with others, many participants said there is a lot of potential, even with using MobileASL; however, the technology needs to support communication with more people instead of people who are just using MobileASL or other commercially available applications.

9.6 Summary

This field study demonstrated the potential of MobileASL facilitating real-time, two-way sign language communication when video is transmitted at extremely low frame rates and bit rates. It was clear that battery life limited the ability for users to sign to each other; however, participants were still successful in holding intelligible conversations and gathering information in real-time. Collecting information about non-laboratory use of mobile video communication further demonstrated the lower limits in which mobile video can be transmitted while maintaining intelligible conversations. Findings from the field study further motivated the need for longer battery life for successful mobile video communication. It was also evident that the purpose of the video call and whom the signer was contacting had a major influence in the frequency of use. Improvements to mobile video communication may include context awareness, which is discussed further in Chapter 12, section 12.3.1.

Chapter 10 Power Saving Algorithms

Reducing the transmission rates of sign language video is only half the solution to extending battery life. Smartphone batteries have evolved over the past decade with early portable devices using older technologies like nickel-cadmium (NiCD or NiCad) to today's most popular battery chemistry of lithium ion. However, even with the growth of battery technology, there is still no Moore's Law [70] for batteries. Mainly, the limiting factor is the fundamental chemistry in a battery's workings. Ions transferring charge within batteries are large, which in turn take up space, along with anodes, cathodes, and electrolytes [90]. Moore's Law holds for computer processors due to the lithography technology used to fabricate chips. Smaller features can be made on processors as lithography improves. However, battery life will continue to be a limiting factor for prolonged mobile video communication.

This chapter presents the investigation of alternative power saving algorithms that utilize features of sign language to extend battery life without negatively impacting mobile video intelligibility. First, two alternative power saving algorithms are introduced, variable spatial resolution (VSR) and the combination of VSR with variable frame rate (VFR). Next the implementation of VFR, VSR, and VFR+VSR on the experimental software, MobileASL, is addressed followed by battery usage analysis. Finally, a second battery savings investigation is presented using IMSDroid, an open source smartphone application, while controlling both the frame rate and bit rate at which real-time sign language video is transmitted.

10.1 Variable Frame Rate

Prior research conducted by Cherniavsky *et al.* [23] on the MobileASL project introduced *variable frame rate (VFR)*. VFR reduces the temporal resolution of the transmitted video based on activity recognition [22] to save computational resources. Basically, when a person is classified as signing, the frame rate is set to 10-12 fps. When a signer is identified as not-signing, the frame rate is reduced to 1 fps, which produces a choppy video quality. Figure 27 is a pictorial example of the VFR algorithm implemented.

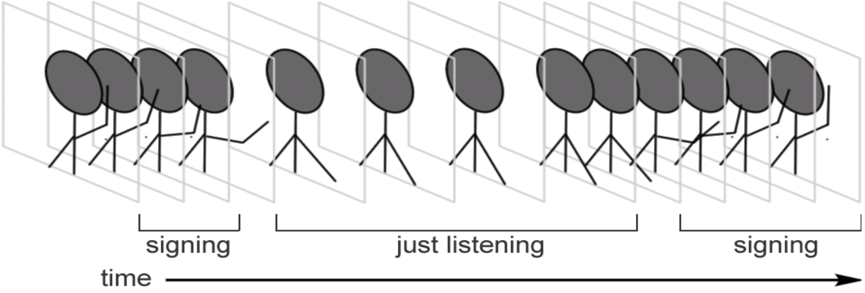


Figure 27: Depiction of variable frame rate algorithm. The frame rate decreases when the signer is not-signing, resulting in "choppy" video quality [23]**.**

Frames that are classified as signing may contain a lot of activity such as fast movement in the hands or face; this results in large inter-frame pixel differences. Frames that are classified as not-signing have small pixel differences due to little change in the background or movement in the hands and face by the user. If the difference between each frame is above a certain threshold, then frames are classified as signing; otherwise they are classified as not-signing. When a not-signing frame is identified, VFR reduces the frame rate from 10-12 fps down to 1 fps.

Cherniavsky *et al.* evaluated VFR in a laboratory setting with 15 participants fluent in ASL. The goal was to measure comprehensibility of conversations while the VFR algorithm was applied during the not-signing sections of a conversation. The objective measurements of this study included number of requests for repetition (repair requests) [103], number of turns associated with repair requests, number of conversational breakdowns, and the speed of finger spelling. Their findings revealed that when VFR was on, participants felt they had to guess at what was being said more frequently than when VFR was off. Applying the VFR algorithm also resulted in more repair requests, took more turns to correct the request, and resulted in more conversational breakdowns. These results prompted us to find alternative power saving algorithms to extend the battery life.

The successful investigation of altering the temporal resolution (VFR) to prolong battery life inspired the investigation of two alternative power saving algorithms: variable spatial resolution (VSR) and the application of VFR and VSR together.

10.2 Variable Spatial Resolution

The VSR algorithm is based on downsampling the width and height of each transmitted frame by a factor of 2, as shown in the block diagram Figure 28.

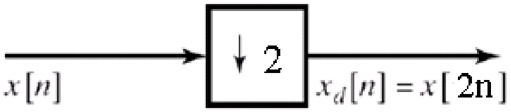


Figure 28: Downsampler Block Diagram.

VSR was implemented in the MobileASL software which transmits video in the YUV 420 format. YUV 420 is the color space takes human perception into consideration. The Y component is the luminance or brightness of a pixel and UV is the chrominance, or color component. YUV 420 specifies that for every four luminance components, there is one chrominance component representing the color of those four pixels. Figure 29 demonstrates the YUV 420 representation for each pixel. The implementation of VSR resulted in averaging four consecutive chrominance values for the downsampled luminance component.

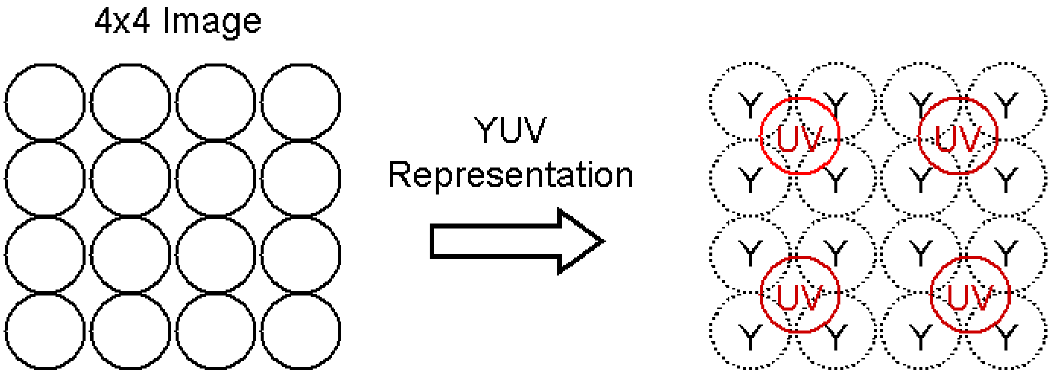


Figure 29: Example 4×4 image and YUV 420 representation.

The difference between VFR and VSR is that the former reduced the spatial resolution, while the later reduces the temporal resolution during not-signing sections of video. When using the VSR algorithm, the frame rate and bit rate are held constant, while the frame size is downsampled before reaching the encoder, transmitted, and enlarged at the receiving end, back to QCIF (176×144). This process produces a perceived blurry video quality. Figure 30 is a pictorial representation of VSR algorithm.

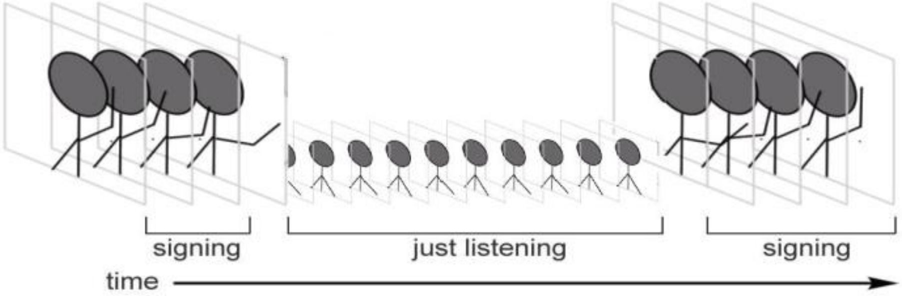


Figure 30: Depiction of variable spatial resolution algorithm. The not-signing frames are downsample to 1/4 the original spatial resolution and displayed at the same frame size, resulting in blurry video quality.

Figure 31 is an example of the degree of video quality degradation when VSR is applied to a not-signing frame.

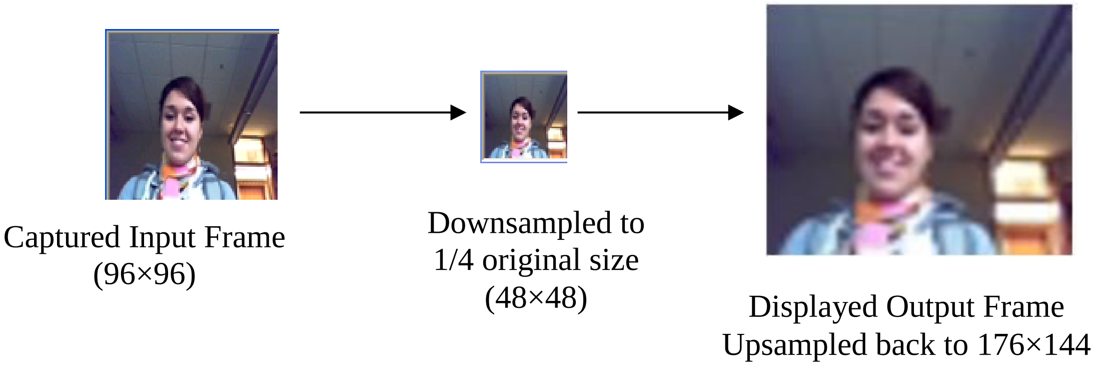


Figure 31: Example of a not-signing frame downsampled to 1/4 of original size, which produces a blurry video quality.

Figure 32 is an example of the video quality when no power saving algorithms is applied. This is the default implementation of MobileASL.

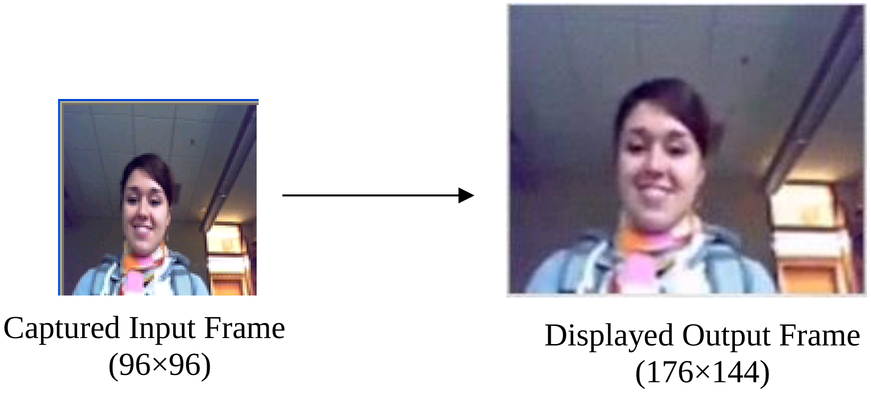


Figure 32: Example of a not-signing frame with default implementation of MobileASL.

10.3 Combination of Variable Frame Rate and Variable Spatial Resolution

The second power saving algorithm is the application of VFR and VSR together when not-signing frames are identified through activity recognition. Signing frames are not impacted when using VFR+VSR algorithms. When not-signing frames are identified, frames are downsampled to 1/4 of their original size before being sent to the phone's encoder *and* the frame rate is reduced to 1 fps. This amounts to transmitting 1/40 of the original amount of data needed when transmitting data at full resolution at the original frame rate of 10-12 fps. The phone receiving the transmitted frames will decode and enlarge the frames to QCIF for viewing. When VFR+VSR is used, viewers perceived both choppy and blurry video. Figure 33 is a pictorial representation of when VFR+VSR is applied.

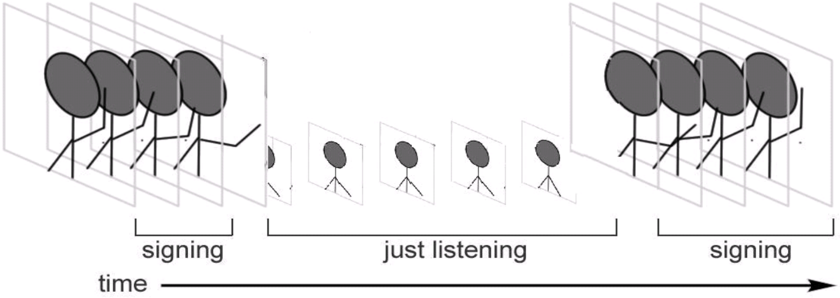


Figure 33: The implementation of VFR and VSR during not-signing portions of a conversation. The resulting output is a combination of blurry and choppy video.

10.4 Cell Phone Battery Power Study

The HTC TyTNII cell phone running MobileASL along with each of the power saving algorithms was used to quantify the extension of battery life. The maximum battery duration of the HTC TyTNII was investigated and compared to the battery life of the default (no power savings algorithm applied) MobileASL implementation. In this experiment, the maximum battery life occurs when the power saving algorithms are continuously implemented since fewer resources are used to encode and transmit video depending on the selected power saving algorithm. Therefore, I conducted a power study using the ideal case when the selected power saving algorithm is constantly implemented i.e. one signer is never signing.

The manufacturers of the HTC TyTNII cell phone specify that a full battery charge can hold 1350 mAh [33]. The minimum current drain for this particular cell phone to operate is 128 mA and the minimum average percent CPU usage is 22.4% to operate the Windows Mobile 6.1 operating system. With this knowledge, a simple formula was used to calculate the battery life of the cell phone:

(Eq. 1) Battery Life in Hours = 1350 mAh / X current drain (mA)

A comparison of battery drain, current consumption, and CPU usage was used to gain better insight of how the MobileASL application consumes the phone's resources. When conducting these experiments, the data were collected with both the MobileASL application and Windows Mobile 6.1 operating system running, unless otherwise noted.

Set Up

For the purpose of this power study, when two cell phones are "holding a conversation," this means that two cell phones were running MobileASL and transmitting data to one another. To simulate the not-signing portions of a conversation, two phones in conversation with each other were placed so that they faced a static object (e.g. the wall) for 30 minutes. A publicly available software tool [2] was used to monitor the battery consumption, current drain, and CPU usage of each cell phone during each experiment. There were four experiments conducted: (1) VFR only; (2) VSR only; (3) both VFR+VSR; and (4) no algorithms applied (control). It is important to note that before each experiment, each cell phone was fully charged to capacity to be consistent across all experiments. Also, since the HTC-TyTNII cellular phones use a lithium-ion battery, which degrades over time; the data collected from two different cell phones were averaged and used in analysis.

10.5 Battery Consumption

For each experiment, the voltage drop across the battery was logged every 5 seconds for 30 minutes. Regression analysis demonstrated that the battery drain was linear for each experiment; therefore, the battery drain data were extrapolated to determine when the battery discharged to 0%. Figure 34 shows the extrapolated average battery life of the HTC TyTNII duration for each power saving algorithm and for the default setting.

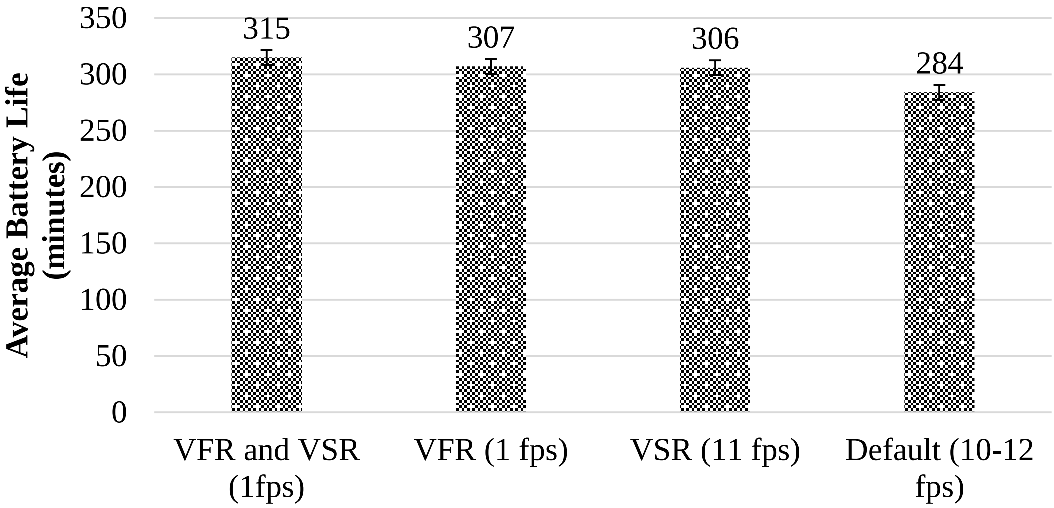


Figure 34: Average battery life (minutes) for each power saving algorithm.

Figure 34 demonstrates that the three battery saving algorithms extend the average battery life of the cell phone. This experiment determined that the average battery life of the cell phone when running the default setting is 284 minutes. The application of VFR, VSR and both VFR and VSR each extend the battery duration on average by 23, 22, and 31 minutes, respectively.

The VFR and VSR algorithms performed similarly with only a minute difference, while applying both methods exceeded the performance of each algorithm alone by 8 or 9 minutes. These experiments show that battery life can be extended when a power saving algorithm is implemented during the not-signing sections of a conversation.

10.6 Current Drain

In addition to measuring the battery drain, I also recorded the current drained from the cell phone’s battery. Similar to recording the battery drain, the value of the current drain was logged every five seconds for 30 minutes. From previously observing battery consumption, I anticipated that applying both VFR and VSR algorithms would consume the least amount of current and the default setting would consume the most current. The current drain for the VFR and VSR algorithms individually was anticipated to have similar current consumption. Figure 35 and Table 11 demonstrate the current drain for each experiment.

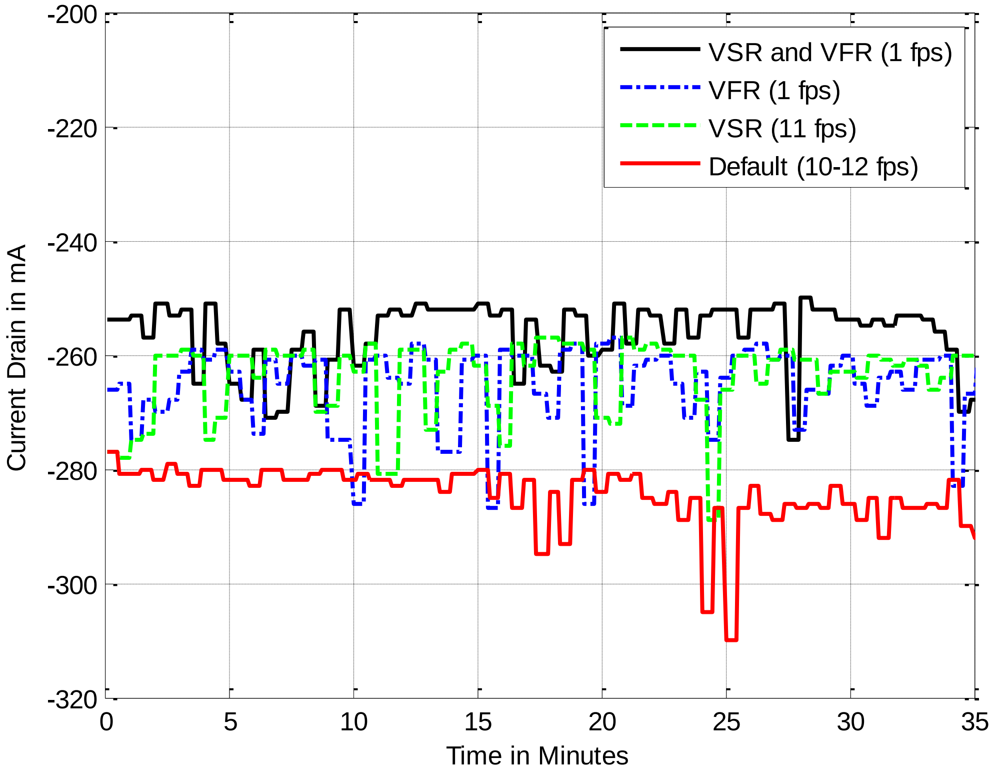


Figure 35: Measured current drain (mA) vs. time (minutes) for each encoding algorithm.

Table 11: Average current drain (mA) for each power saving algorithm.

Method	Average Current Drain (mA)
Default	284
VSR	264
VFR	265
VFR+VSR	257

Comparing Figure 34 to Figure 35 and Table 11 demonstrates how the average current drain and average battery duration are related. The default setting drains the most current, and also has the shortest battery life. The current drain for VSR and VFR is similar, which parallels the similarity of battery duration between these two algorithms. Finally, applying both VFR and VSR has the least current drain and the longest battery duration. (The measured current drain includes running both MobileASL and the Windows Mobile 6.1 operating system.)

When running these experiments, I wanted to check that the phones were holding close to 1350 mAh of charge. The estimated battery life in hours equation (eq 1) was used to confirm the calculated battery drain were accurate and the battery charges when starting the experiments were consistent with manufacturer specifications. Also, I also wanted to quantify how much current was just being consumed by MobileASL.

Table 12 is the compiled data for battery life; the current drain of the phone when it is just running the Windows 6.1 operating system; the current drain by the MobileASL application only; total current drain; and the total charge held by the battery.

Table 12: Comparison of battery life (minutes) and current drain (mA) of each experiment.

Method	Default	VSR	VFR	VFR and VSR
Battery Life^ (minutes)	284	306	307	315
Current Drain of Phone Only (mA)	128	128	128	128
Current Drain of MobileASL Only^ (mA)	156	136	137	129
Total Current Drain^ (mA)	284	264	265	257
Full Battery Charge^ (mAh)	1344.27	1346.40	1355.92	1349.25

^ Averaged over two phones

Recall that the battery of the HTC TyTNII cell phone is intended to hold 1350 mAh of charge and needs at least 128 mA to run the Windows 6.1 operating system. Therefore, I determined the current drain of the MobileASL application only by using the results of the average current drain for each experiment minus 128 mA.

Table 12 lists the average mAh for a full battery charge for the HTC TyTNII phones. On average, a full battery charge was 1349 ± 4.79 mAh, which confirms the formula for the relationship between battery life and current drain was correct, with an expectation of a full battery charge holding 1350 mAh. These results demonstrate that the phones used for these experiments were capable of being charged to their intended capacity.

10.7 CPU Usage

Measuring the battery duration and the current drain alone did not reveal why certain implementations consumed more current than the others. I suspected that since the VFR and

VSR algorithms were sending temporally or spatially reduced resolutions of the video, this reduced the amount of computation needed to process the video; however, more information was needed. Figure 36 shows the total CPU usage for each method. The total CPU usage is the sum of all the applications running on the phone (operating system and MobileASL) and how much of the phone's processor is being consumed.

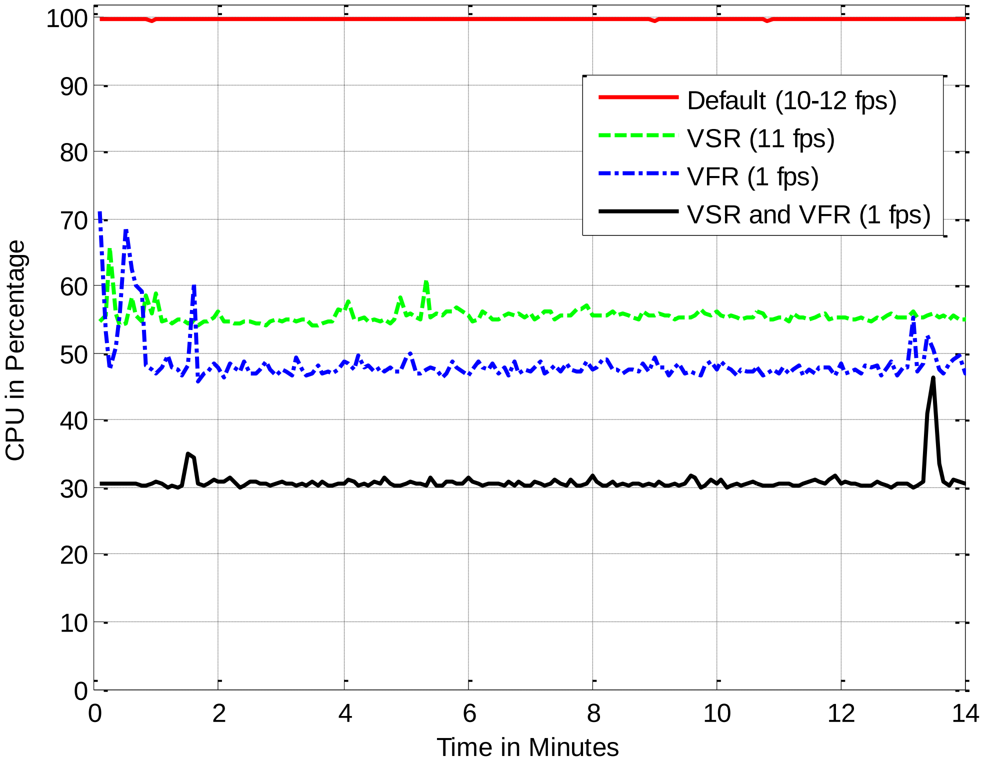


Figure 36: Total CPU usage (percentage) vs. time (minutes).

As Figure 36 shows, the default setting of the MobileASL application uses 99% of the CPU. This large CPU usage could be a possible explanation as to why the battery drains more quickly than applying the different power saving algorithms.

Igor and Cruck investigated how the HTC TyTNII cell phone consumes resources through isolating and measuring different components of the phone. Their results found the baseline CPU usage is 22.2% for the cell phone to be functional [33]. Using this knowledge as a reference, the CPU usage for the phone only (when running just the operating system) was measured as well as the average CPU usage of the MobileASL app are shown in Table 13.

Table 13: CPU usage (percentage) for phone only, MobileASL application only, and total system.

Method	Default	VSR	VFR	VFR and VSR
CPU Phone Only^(%)	23.8	23.7	22.0	22.0
CPU MobileASL Only^(%)	75.9	31.8	26.4	10.8
CPU Total System^ (%)	99.7	55.5	48.4	30.8

^Averaged over two phones

As Table 13 shows, the CPU usage of the phone only was measured at 23.8%, which is approximately the same rate found by Igor and Cruck. They found that the Windows Mobile 6.1 occupies 22.2% of the operating system. When comparing the CPU usage of the MobileASL application only, there is an impressive drop in the average CPU usage for applying VFR and VSR, with only 10.8% of the CPU used when not-signing frames are transmitted. The default setting still reflects a large CPU usage of 75.9%. The CPU usage of only VFR and only VSR shows a slight difference with consuming 26.4% and 31.8% respectively. When the battery duration and current drain were measured for VFR and VSR individually, both algorithms produced similar results. However, here the difference between the two algorithms is more apparent, with VSR consuming 5.4% more CPU than the VFR algorithm. VSR consuming more CPU could be due to the VSR algorithm constantly transmitting 11 fps during the signing and not-signing frames, while in the VFR algorithm, the frame rate is reduced to 1 fps. Finally, applying both VFR and VSR algorithms utilizes the least amount of CPU, which is consistent with the longer battery duration and least amount of current drain.

10.8 Summary of Battery Power Algorithms

The cell phone battery power study determined that individually, VFR and VSR algorithms both extend the battery life over the default MobileASL implementation by 22 minutes. The combined implementation of VFR and VSR extended battery life of 31 minutes over the default setting. Chapter 11 discusses a different web study evaluating video quality perception when these power saving algorithms are applied.

10.9 IMSDroid Battery Experiment

Utilizing the structure of sign language communication allowed for the implementation and evaluation of different power saving algorithms to extend smartphone battery life on the HTC TyTNII smartphone running on the Windows Mobile 6.1 platform. A limitation of this prior research was that implementation and evaluation of each power saving algorithm was specific to that phone. A benefit of using the HTC TyTNII was taking advantage of the smaller screen size and thus, the smaller frame size in which video was captured, transmitted, and received. The evolution of smartphone technology has resulted in larger screen sizes, increased processing power, and longer battery life. The next step in the evaluation of battery drain is quantifying the battery drain for video transmitted at frame rates and bit rates lower than recommended standards on more current mobile technology.

Using the technology implemented in the laboratory study (described in Chapter 8, section 8.1), the battery life was evaluated for transmitting video at four low frame rates (5, 10, 15, 30 fps), while the bit rate per frame was held constant (averaging 5 kb/frame). Unlike prior work, this battery power experiment evaluated battery drain when the transmission rate was constant during a two-way conversation.

Experiment Setup

The Samsung Galaxy S3 smartphone running Android 4.1 was used to evaluate battery life while transmitting sign language content in real-time over IMSDroid, an open source video chat app [55]. IMSDroid was modified to transmit video at each of the low frame rates. A free smartphone diagnostic app, called AndroSensor [5], was used to log the discharge of the battery life in the experiment.

AndroSensor ran in the background of IMSDroid and logged the battery life percentage in 5 second increments for 30 minutes. In a preliminary experiment, I discovered that transmitting video of a person signing consumes more battery life than transmitting a static image. Therefore, I decided to conduct all experiments with the smartphone facing a computer monitor where a person was signing on the screen. Figure 37 is a picture of this experimental setup.

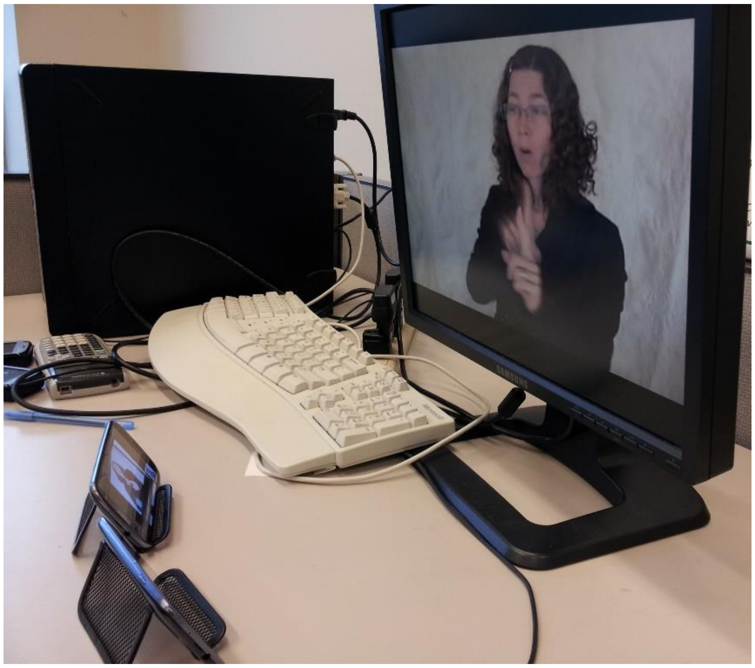


Figure 37: Experimental setup where two Samsung Galaxy S3 phones are facing a computer screen with a video of a woman signing in ASL.

A total of seven experiments were conducted: one for each of the frame rates of interest; IMSDroid 'on' and not transmitting data; and IMSDroid 'off'; and the Samsung Galaxy S3 phone on standby mode.

10.9.1 Results

As anticipated, increasing the frame rate at which sign language video was transmitted over the smartphones consumed the battery life more quickly, which agrees with the battery results presented in the laboratory study (Chapter 8). This is because more processing power is required to transmit video at higher frame rates. Regression analysis demonstrated that the battery drain was linear for each experiment; therefore the battery drain data were extrapolated to determine when the battery discharged to 0%. Figure 38 shows the extrapolated data for the average battery life of the Samsung Galaxy S3 for each frame rate.

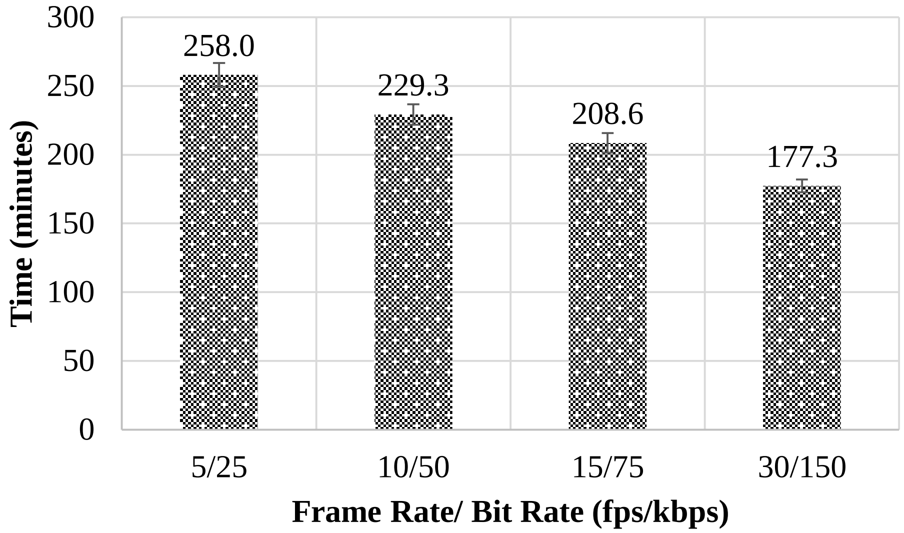


Figure 38: Average battery life for transmitting sign language video at each frame rate/bit rate.

From this experiment, it is estimated that the Samsung Galaxy S3 on standby, with IMSDroid turned off, has a battery life of 1000 minutes and IMSDroid turned on and not transmitting video has an estimated battery life of 750 minutes. The Samsung Galaxy S3 specifications listed that a fully battery charge could last up to 8 hours of talk time [101]. My results demonstrate that transmitting video on mobile devices is computationally intensive and depletes a full battery charge in 4 hours.

Chapter 8 described the laboratory study investigating the lower limits at which sign language video can be transmitted without sacrificing intelligibility. Part of the laboratory study included quantifying how much battery life can be extended when transmitting video at each rate during actual sign language conversations. The results from the battery analyses in the laboratory study corroborate the findings here: transmitting video at the lower frame rates also increases the battery life duration. These and other analyses performed in the laboratory study further support my thesis statement that mobile sign language video transmitted at frame rates and bit rates below recommended standards, does indeed extend battery life and reduce total bandwidth consumption, is still intelligible and can facilitate real-time mobile video communication.

Chapter 11 Web Study: Perception of Power Saving Algorithms

The power saving algorithms introduced in Chapter 10 inherently contributes visual distractions, such as perceived blurred and/or choppy video quality. To better understand the potential negative attributes of viewing sign language video with lower video quality, a web study was created to investigate how users of mobile video communication experience and feel about degradation of video quality in exchange for extended smartphone battery life. *Intelligibility of video content was not the focus of this web study.* Rather, the purpose was to investigate the perceived intelligibility of three power saving algorithms: variable frame rate (VFR), variable spatial resolution (VSR), and VFR+VSR.

Study Design

The study design was a 2×2 within-subjects factorial design. The two factors were the two encoding schemes (VFR, VSR), each with two levels, "on" or "off." Figure 39 depicts the combinations of each factor and its levels.

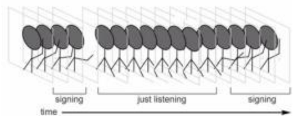
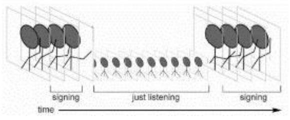
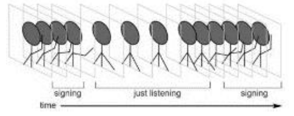
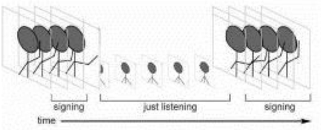
		VSR (factor 2)	
		OFF	ON
VFR (factor 1)	OFF		
	ON		

Figure 39: Combinations of factors and levels within the web study.

Each respondent was randomly assigned to view one of three videos of a person signing in ASL. The content of the video was a one-sided conversation with an equal amount of signing and not-signing. The assigned video was shown four consecutive times, but each time a different power savings algorithm was applied (VFR, VSR, VFR and VSR, or none). The respondents did not know which encoding algorithm was applied; they were only told that there may be changes to the video quality, but not when, where, how, or how much.

After each video, four statements were presented to understand the users' perception of the video.

The four statements were:

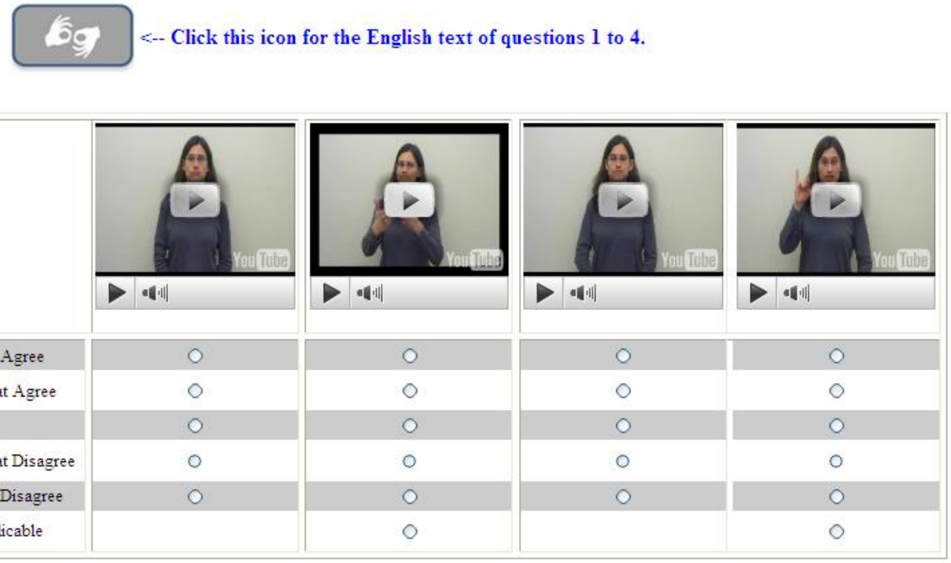
Q1) I notice portions of this video were choppy.

Q2) The choppy portions of the video are distracting.

Q3) I notice portions of this video are blurry.

Q4) The blurry portions of the video are distracting.

The same four statements were presented after each video. A 5-point Likert scale was used to gather respondent feedback after each video was shown. The degrees of the 5-point Likert scale in descending vertical order were: *strongly agree*, *somewhat agree*, *neutral*, *somewhat disagree*, *strongly disagree*. For Q2 and Q4, a *not applicable* option was provided, since these answers depended on those to Q1 and Q3, respectively. (A respondent cannot agree or disagree that the perceived choppiness or blurriness of a video was found to be distracting if choppiness or blurriness was not noticed in the first place.) The study concluded with a demographic questionnaire.



<-- Click this icon for the English text of questions 1 to 4.

Strongly Agree	○	○	○	○
Somewhat Agree	○	○	○	○
Neutral	○	○	○	○
Somewhat Disagree	○	○	○	○
Strongly Disagree	○	○	○	○
Not Applicable		○		○

Figure 40: Screen shot of ASL video interpretation of survey questions and 5-point Likert scale.

11.1 Results

The web study investigated the effects of VFR and/or VSR on video quality perception. There were 148 respondents fluent in ASL (80 men, 65 women, and 3 who did not specify). Their ages ranged from 18-75 years old and all but four respondents were deaf. All but sixteen respondents indicated that they own a cell phone and use it to text message. Finally, all but eleven respondents indicated that they use video phones and use video relay services.

A nonparametric factorial analysis was used to analyze the 5-point Likert scale responses for the four questions presented after each video in the web study. An Aligned Rank Transform [49] was performed since data were not normally distributed, were ordinal in nature, and were bounded by the scale endpoints. A repeated measures ANOVA was performed on the aligned ranks. In the analysis of Q2, 430 of 596 data points were used which represented responses from respondents who marked 3-5 (neutral-strongly agree) in Q1, or who did not indicate Q2 was "N/A." In the analysis of Q4, 445 of 596 data points were used which represented responses from respondents who marked 3-5 (neutral-strongly agree) in Q3, or who did not indicate Q4 was "N/A." Therefore, Q2 and Q4 only analyzed the responses from respondents who *did* notice choppy or blurry video by marking 3-5 (neutral-strongly agree) for Q1 and Q3, respectively. Table 14 displays the mean values of applying VFR or VSR for Q1-Q4.

Perceived Choppiness

Q1 asked respondents if they noticed choppy sections of video when VFR and/or VSR were turned on or off. Q2 followed by asking if choppy video sections were distracting. When VFR was on, respondents unsurprisingly felt the video was choppier than when VFR was off (F(1,591)=80.94, *p*<.001). This result can also be seen in Table 14 where the mean value for Q1 increased when VFR was turned from off to on.

Table 14: Mean values of applying VFR or VSR for Q1-Q4.

	VFR	mean	standard error	VSR	mean	standard error
Q1	off	3.13	.09	off	3.57	.09
	on	4.12	.07	on	3.68	.08
Q2*	off	3.87	.08	off	4.07	.07
	on	4.25	.06	on	4.11	.07
Q3	off	3.50	.09	off	3.15	.08
	on	3.90	.07	on	4.25	.06
Q4*	off	4.11	.07	off	3.95	.08
	on	4.19	.07	on	4.30	.06

*Mean calculated from respondents who marked 3-5 (neutral-strongly agree) from the previous question and did not mark "N/A."

As expected, having VSR on or off had no effect on the perceived choppiness of the video (F(1,591)=3.58, *n.s.*). However, there was a significant VFR×VSR interaction (F(1,591)=8.09, *p*<.01). An important finding, as Figure 41 demonstrates, is that when VFR was on, the use of VSR significantly lowered the perceived choppiness of the video (F(1,591)=23.48, *p*<.001).

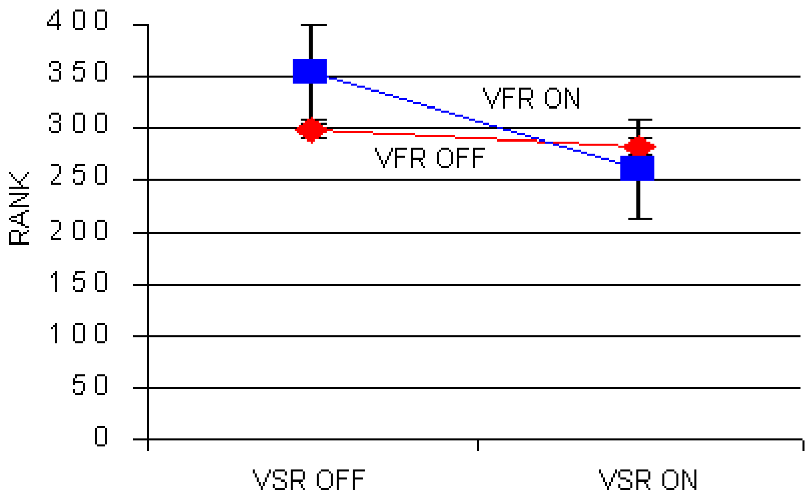


Figure 41: VFR×VSR Interaction for Q1. Note the Y-axis is the rank used by the nonparametric analysis procedure. Lower values indicate less perceived choppiness.

For Q2, respondents who marked 3-5 (neutral-strongly agree) in Q1, or who did not indicate that Q2 was "N/A," was found that when VFR was on, they felt that the choppiness *was* distracting (F(1,425.3)=18.10, *p*<.001). Similar to the results found in Q1, whether VSR was on or off had no effect on respondents feeling that choppiness was distracting (F(1,425)=3.86, *n.s.*). There was no VFR×VSR interaction (F(1,425)=1.65, *n.s.*).

Perceived Blurriness

Q3 asked respondents if they noticed blurry sections of video when VFR and/or VSR were turned on or off. Q4 then asked if blurry video sections were distracting. Expectedly, when VSR was on, respondents noticed the video was blurrier than when VSR was off (F(1,591)=131.57, *p*<.001). Unexpectedly, respondents felt that when VFR was on, the video also appeared more blurry than when VFR was off (F(1,591)=21.95, *p*<.001). There was a significant VFR×VSR interaction (F(1,591)=18.99, *p*<.001). As Figure 42 shows, when VSR was off, whether VFR was on or off did not matter for perceived blurriness (F(1,591)=2.20, *n.s.*).

But when VSR was on, the use of VFR significantly lowered the perceived blurriness of the video (F(1,591)=21.90, p<.001).

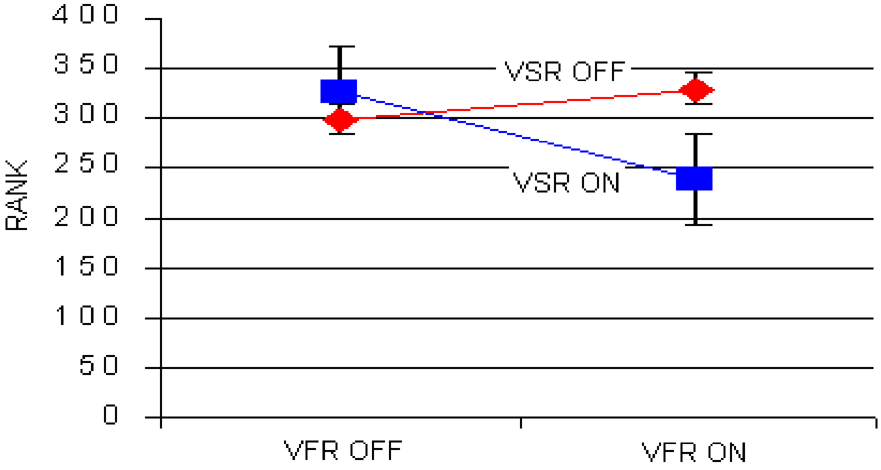


Figure 42: VFR×VSR Interaction for Q3. Note the Y-axis is the rank used by the nonparametric analysis procedure. Lower values indicate less perceived blurriness.

For Q4, respondents who marked 3-5 (neutral-strongly agree) on Q3, or who did not indicate that Q4 was "N/A," it was unexpectedly found that when VFR was on they perceived an increase in blurriness of the video (F(1,440.2)=7.91, *p*<.01). Not surprisingly, as Table 14 shows, when VSR was on, respondents felt that the blurriness was more distracting than when VSR was off (F(1,440)=26.26, *p*<.001). Finally, there was a significant VFR×VSR interaction (F(1,440.1)=5.71, *p*<.05). As Figure 43 demonstrates, when VSR was off, whether VFR was on or off did not contribute to perceived blurriness to cause distractions (F(1,440.2)=0.34, *n.s.*) despite switching VFR off to on contributing to perceived blurriness in Q3. But when VSR was on, the use of VFR significantly reduced the distracting nature of perceived blurriness of the video (F(1,440)=9.38, p<.05).

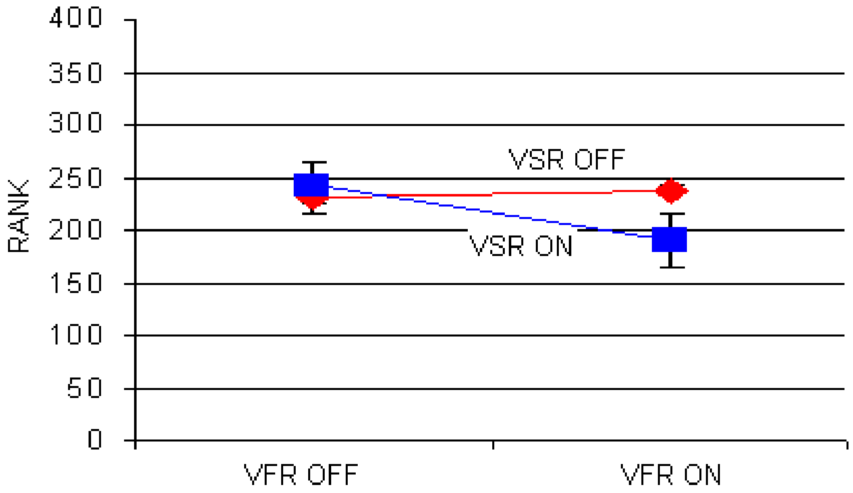


Figure 43: VFR*VSR Interaction for Q4. Note the Y-axis is the rank used by the nonparametric analysis procedure. Lower values indicate less perceived distraction due to blurriness.

11.2 Discussion

Although one would expect to find that VFR produces perceived video choppiness and VSR produces perceived video blurriness, which was found, it was also discovered that when *both* VFR and VSR are used, they largely ameliorate the choppiness and blurriness perceived, *i.e.*, they each improve the use of the other. A reason for this improvement could be that the blurriness caused by VSR “smoothes out” the choppy effect caused by VFR. This smoothing effect has been found in prior work to improve perception of shaky video quality when video compression is introduced [18]. It has also been found that shaky video with low temporal movement, like a home cooking show, does not degrade perceptual quality as does shaky video containing high action motion like a sports game [28]. Therefore, the findings concerning the significant VFR×VSR interactions for Q1 and Q3 indicate that VFR and VSR may work together to produce a smoothing effect. For Q3 and Q4 it was surprising to find that applying VFR increased respondents’ perception of blurriness, since that algorithm does not objectively contribute to blurry video quality. This could be a result of respondents noticing a change in video quality due to VFR and applying what they see to answer Q3 and Q4.

11.3 Summary

The web study evaluated video perception of VFR, VSR, and VFR+VSR power saving algorithms. Chapter 11 demonstrated that video transmitted when both VFR+VSR are applied saves the most battery power. The web study evaluating video perception of VFR, VSR, and VFR+VSR power saving algorithms revealed that respondents demonstrated a decrease in perceived blurriness and choppiness caused by each algorithm alone. This result demonstrates that manipulating both the temporal *and* spatial resolution of a video to save battery power is a good approach.

Chapter 12 Conclusion and Future Work

The purpose of this dissertation was to investigate the lower limits in which mobile sign language video can be transmitted, for the purpose of saving bandwidth and battery life, without sacrificing intelligibility. By taking a human-centered approach to evaluating video compression, this dissertation demonstrates the following thesis: that mobile sign language video transmitted at frame rates and bit rates below recommended standards (Chapter 5-Chapter 9), does save bandwidth and battery life (Chapter 8, Chapter 10, Chapter 11), while maintaining intelligibility (Chapter 6 and Chapter 7) and still facilitates real-time mobile video communication (Chapter 8 and Chapter 9). In addition, Chapter 2 motivates that mobile sign language video communication has the potential to be more accessible and affordable if the current recommended video transmission standard of 25 frames per second at 100 kilobits per second (kbps), as prescribed in the International Telecommunication Standardization Sector (ITU-T) Q.26/16, were relaxed. This dissertation has explored a relaxed standard for sign language video transmission using lower frame rates, bit rates, and spatial resolutions to increase the accessibility and affordability of mobile video communication. Chapter 3 establishes the components comprising signal intelligibility, disentangling signal intelligibility from signal comprehension for evaluations, a distinction that has not been made with prior models. Finally, this dissertation demonstrates that a human-centered approach to evaluating video compression is more dynamic; takes a holistic approach to improving mobile sign language video communication while reducing resource consumption; and provides recommendations for change with industry standards for video transmission rates.

Mobile technology is rapidly evolving from the speed at which data is transferred to the device in which content is presented. Tradeoffs will continually need to be made between cost, quality of experience, and performance of devices. Ideally, consumers of video content prefer both high quality of experience and device performance at a low cost; while service providers strive to maintain their bottom line without providing more services than required. The future rates at which mobile video content are transmitted will increase; however, the total network bandwidth will be limited, even with the growing infrastructure over time. This work opens up new approaches and techniques to human-centered evaluations of mobile video content and improves upon regulated standards at which video content is transmitted while considering cost and accessibility.

Each web study builds upon the findings from prior studies to demonstrate the potential benefits for transmitting video below the current recommended standards for real-time mobile sign language video communication. In addition to presenting the design and implementation of each study, I have discussed the limitations of the findings. Now I will reflect on some of the main contributions and findings that have emerged from this research. I will also discuss future research directions that address some of the limitations of this work, indicating ways to build upon insights found from this dissertation to explore new problems. Finally, I restate the major contributions that this dissertation makes to the electrical engineering field, specifically in the area of digital signal processing, video compression, and HCI, and close with final remarks.

12.1 Reflections and Insights

In this section, I reflect on several insights that have been gained through this dissertation about (1) the importance of distinguishing video intelligibility from video quality and video comprehension; (2) linguistically accessible studies; (3) user-centered approaches to evaluating video intelligibility; and (4) lessons learned during my dissertation research.

12.1.1 Distinguishing Video Intelligibility from Quality and Comprehension

Evaluating video intelligibility is often not the primary focus when compression is applied to video. As part of this work, Chapter 3 presents the Human Signal Intelligibility model, a conceptual model distinguishing the components comprising intelligibility and comprehension for evaluation of videos. It was important to make a distinction between video quality, intelligibility, and comprehension because one of the goals was to provide intelligible sign language communication. Evaluating video intelligibility based on video quality or comprehension does not necessarily reflect the videos' ability to facilitate a conversation. For example, a video can be perceived to have high quality or receive a high PSNR score, but appear at 1 fps. Also, people can perceive changes in video quality before intelligibility of content is affected. Video comprehension evaluations may not necessarily reflect video intelligibility either. Often video comprehension evaluations require participants to repeat back what was seen, which does not imply true understanding of the material. Answering comprehension questions may not necessarily reflect video intelligibility because an incorrect comprehension question could result from misunderstanding of the question or from not having adequate ASL vocabulary, but not necessarily reflect that the entire video was unintelligible. Establishing the HSIM justifies my approach to evaluating signal intelligibility by measuring comprehension of video, with the caveat that components within the model are accounted for.

12.1.2 Linguistically Accessible Studies

Much of the success of the multiple web studies, laboratory study, and field study can be attributed to creating linguistically accessible instructions for the deaf and hard-of-hearing participants. Web studies are straightforward to create and often assumptions are made that the

English text would provide adequate information. Since ASL is a different language from English, it would be naïve to assume that the English text included in the studies is adequate. Although it may take extra time and resources to work with a certified ASL interpreter to create the ASL interpretation of the instructions, this extra step demonstrates to participants that the study design is intended to respect their culture and language.

Part of the benefit of creating web studies is the ability to recruit participants via e-mail, posting to social media, and snowball sampling, where existing study participants can share the study with their friends. Participants tend to appreciate the instructions in ASL as well. Below is an e-mail I received from one of the participants who is a certified ASL interpreter:

> *"I took the survey today and posted to both the ORID Facebook page and my wall. I wanted to thank you personally for the excellent quality of the video instructions. I assumed that Deaf participants would still have to read English instructions and was so happy you'd thought beyond that.*
>
> *I have a Deaf/Interpreting list of people I follow on Twitter so I'll broadcast it there too."*

12.1.3 User-Centered Approach to Evaluating Video Intelligibility

Realistic adoption of mobile video communication with video transmitted at the lower frame rate and bit rates I investigated is dependent on the users' willingness to use this technology in the "real world." A user-centered approach was taken in each evaluation to inform implementation of the mobile video application. Conducting web studies allowed feedback from hundreds of participants. A limitation of web studies was the inability to have participants sign to others over the lower video transmission rates. However, the findings from the web studies gave insights to the lower limits of video intelligibility. The intelligibility ceiling effect was demonstrated in multiple web studies conducted (Chapter 6 and Chapter 7) and corroborates findings from earlier work conducted within the MobileASL project.

The laboratory study allowed fluent ASL signers to experience and evaluate an experimental app transmitting video at the lower transmission rates. A limitation of the laboratory study is that participants are in an artificial setting and may not use the technology for as long as asked. However, it was clear from the study results that intelligible conversations can occur at the lower transmission rates. Also, the results agree with findings from the web studies.

12.2 Lessons Learned

Creating web surveys that are linguistically accessible to deaf and hard of hearing people gave this research more creditability within the deaf community. I was successful in recruiting over 100 respondents per web survey each of which took 12-20 minutes to complete. Over the years, it was clear that a subset of 50 participants who actively participated were invested in the success and the growth of the MobileASL software.

The Human Signal Intelligibility Model (Chapter 3) influenced all study designs and guided selection and identification of different components within each study to be held constant, like environmental factors and technology used. Establishing language fluency was another important component to account for; therefore, my approach was to allow participants to self-report ASL fluency, to encourage more participants, and use demographic questions to infer experience of ASL fluency. By making the distinction between signal intelligibility and signal comprehension, where the latter is defined as signal intelligibility *plus* human knowledge and the receiver's mind, I can confidently report findings on video intelligibility. Since all the study analyses were performed on data collected from fluent ASL respondents, I was not concerned with language proficiency influencing my results.

While working on this dissertation, I observed an evolution of smartphone technology and cellular services. Prior MobileASL studies conducted in 2006-2008 demonstrated that participants were eager to know when the MobileASL software would be available for use. Mobile Skype and FaceTime began to be available in 2010, which made real-time sign language video more widely available. Also, VRS companies began to produce their own apps facilitating mobile sign language communication. However, these apps suffered from network congestion, which resulted in poor video quality. Also, in 2010 there was limited access to 3G cellular service. In 2011, more smartphones were introduced into the US market with front-facing cameras, which increased the choices of hardware that could send video. MobileASL software became software-dependent and steps were taken to upgrade MobileASL from Windows Mobile 6.1 to the Android operating system in late 2012.

Porting software from Windows Mobile 6.1 to Android was not an easy task and took many years and person-hours to reverse-engineer the IMSDroid software to its working form used in the laboratory study (Chapter 8). An immediate challenge was that MobileASL was built using a specific ARM stack, which is not available on Android devices. Second, MobileASL was built using a home-grown communications architecture, which did not follow standard internet protocols like Session Initiation Protocol (SIP). IMSDroid was selected because of its existing functionality of video transmission and the open source code.

A huge lesson learned with making IMSDroid work is the importance of understanding the existing architecture of the technology before making major changes and adding new components. IMSDroid was originally created as a proof of concept, and a large challenge was compiling all the code and verifying that everything was working properly before making changes.

It is evident that the future of mobile data will be larger, faster, and provide higher quality of video content. The lower limits in which video can be transmitted may not be needed for well-established infrastructures; however, there is potential use of these findings in developing countries with little to no network access.

12.3 Future Work

There will be a continued need for investigating tradeoffs between video intelligibility and resource consumption when providing real-time mobile sign language communication. Several technical challenges remain so that higher video transmission rates can improve video intelligibility.

12.3.1 Context-Aware Video Quality Adaptation

The laboratory and field studies revealed that participants' choice to communicate over mobile video or texting was dependent on the nature of the conversation and to whom they were communicating. Current commercial video apps vary the video transmission rate based on bandwidth availability, while ignoring the external factors surrounding the conversation, such as the context of the conversation and the device facilitating the conversation. A more dynamic method to improve mobile video transmission rates is to create an algorithm that is context-aware to improve video quality. For example, during a video call, other external factors can be monitored such as: location of call; environmental factors such as sunlight and rain; remaining battery life and remaining data allotment for the month. These and other components outlined in the HSIM would aid in parameter selection. Part of this work will be to capture the different contexts in which conversations occur. A field study, in which participants are asked to communicate via texting and mobile video transmitted at the lower frame rates and bit rates recommended in this dissertation would allow researchers to understand context such as to whom the person was communicating and the nature of the conversation. A dynamic mobile video app that incorporates all of these components would allow better resource distribution and improvement on mobile video communication.

12.3.2 Region-of-Interest Improvements

The MobileASL software created for the Windows Mobile 6.1 platform, implemented a ROI-based video encoding system, where more bits were allocated to the face and hands than to the background. This dissertation focused on the baseline transmission rates at which to transmit video, without ROI-encoding. A future area of research would be developing new algorithms that would track the regions of interest most important to the signer and allocate more data to the ROIs.

12.3.3 Mobile Video Communication in Emergency Situations

Emergency response work can greatly benefit from the additional information provided with live video. Findings from this dissertation can be applied to transmitting live video broadcasted in emergency situations. A potential area of research would be identifying which transmission rates (frame rate, bit rate, and spatial resolution) provide enough intelligible content to aid emergency response workers. Part of this work would include understanding the situations faced by response workers on an accident site; identifying key interactions between response workers; and identifying how streaming video live could reflect situation-specific information.

12.4 Summary of Contributions

The different contributions made by this dissertation are broken down by themes:

Concepts/Theory

- The Human Signal Intelligibility Model, a new conceptual model that outlines the components comprising signal intelligibility and signal comprehension for the purpose of video intelligibility evaluations. (Chapter 3)

Method

- The Intelligibility Response-Time Method, a new method using response-time as an indicator of mental effort to further inform video intelligibility evaluations. (Chapter 7, section 7.2)
- Web study methodology for evaluating ASL intelligibility. (Chapter 4)

Study Results

- Empirical findings verifying an intelligibility ceiling effect for frame rate, where increasing the frame rate above 10 fps does not improve perceived video intelligibility when video is transmitted at a constant bit rate. (Chapter 6 and Chapter 7)
- Empirical findings verifying an intelligibility ceiling effect for bit rate, where increasing the bit rate above 60 kbps does not improve perceived video intelligibility. (Chapter 6 and Chapter 7)
- Empirical findings demonstrating a strong negative correlation between mental effort and response-time, where response-time can be used as an additional evaluator of video intelligibility. (Chapter 7, section 7.5)
- Empirical findings demonstrating that intelligible sign language conversations can be held at frame rates as low as 5 fps. (Chapter 8)

- Empirical findings demonstrating that the speed of finger spelling does not change with a lower frame rate. (Chapter 8, section 8.4.2)
- Empirical findings demonstrating what type of adaptation techniques are used to compensate for video transmitted at 5 fps. (Chapter 8, section 8.5.3)
- Empirical findings validating the bandwidth and power savings associated with reducing video transmission rates via frame rate, bit rate, and spatial resolution. (Chapter 8, Chapter 10, Chapter 11)
- Empirical findings demonstrating that objective and subjective evaluations alone are not the strongest indicators of video intelligibility. (Chapter 5)
- Empirical findings demonstrating that combining both variable frame rate and variable spatial resolution extends the most battery life while reducing the perceived negative effects introduced by each algorithm alone. (Chapter 11)
- Empirical findings demonstrating that mobile video communication can facilitate information gathering and responses more quickly than texting. (Chapter 9, section 9.5)

Technology

- Implementation of two new power saving algorithms in the MobileASL software that utilize aspects of sign language to reduce bandwidth and power consumption. (Chapter 10)
- Implementation of lowering the frame rates at which video is transmitted on an open source video application available for Android. (Chapter 8)
- Web study design structure for creating linguistically accessible web studies. (Chapter 4)

12.5 Final Remarks

This dissertation has demonstrated the following thesis:

Mobile sign language video transmitted at frame rates and bit rates below recommended standards (ITU-T vs. 10 fps/50 kbps), which saves bandwidth and battery life by about 30 minutes, is still intelligible and can facilitate real-time mobile video communication.

My main focus in this dissertation has been to investigate this claim by taking a human-centered approach in designing and implementing four web studies, a laboratory study, and a field study evaluating different video compression techniques by modifying the rate at which video is encoded via reduced frame rate, bit rate, and spatial resolution. I have also considered the intended end-users while evaluating the different ways to reduce resource consumption while maintaining intelligible content. Part of this work also addressed the lack of uniformity in the way that signal intelligibility and signal comprehension were operationalized for evaluation. This led to the creation of the *Human Signal Intelligibility Model* to distinguish the components comprising video intelligibility from video quality and video comprehension. Another part of this work demonstrated that the current recommended standards to transmit intelligible sign language communication can be more relaxed. Study results have demonstrated that intelligible mobile sign language conversations can occur at frame rates as low as 10 fps and bit rates as low as 50 kbps. Table 15 summarizes the different web and laboratory studies conducted in the MobileASL project and the resulting frame rate and bit rate at which the intelligibility ceiling effect and diminishing returns occurred, respectively.

During my investigation of this dissertation claim, I witnessed the evolution of smartphones becoming mainstream; data connectivity becoming more readily available; and mobile video communication becoming integrated into daily use. As the demand for mobile video content increases, whether that is with streaming media or video communication, tradeoffs have to be made between content intelligibility and resource consumption.

Table 15: Summary of MobileASL Project Studies with the Resulting Frame Rate and Bit Rate at which the Intelligibility Ceiling Effect and Diminishing Returns Occurred, respectively.

Study	Publication Venue	Frame Rate (fps) Intelligibility Ceiling Effect Occurred	Bit Rate (kbps) Diminishing Returns Occurred	Video Spatial Resolution
MobileASL: Intelligibility of Sign Language Video as Constrained by Mobile Phone Technology	ASSETS 2006	10	N/A	96×80
Evaluating Quality and Comprehension of Real-Time Sign Language Video on Mobile Phones	ASSETS 2011	N/A	40	192×144
A Web-Based Intelligibility Evaluation of Sign Language Video Transmitted at Low Frame Rates and Bitrates	ASSETS 2013	10	60	320×240
Response-Time as a Measure of Mental Effort in Evaluating Low Bandwidth Mobile Video Communication	ASSETS 2014-pending	10	60	320×240
Analyzing the Intelligibility of Real-Time Mobile Sign Language Video Transmitted Below Recommended Standards	ASSETS 2014-pending	10	50	320×240

References

[1] "iPhone 4 jailbreak unlocks 3G FaceTime calls: 2010. *http://www.informationweek.com/mobile/mobile-devices/iphone-4-jailbreak-unlocks-3g-facetime-calls/d/d-id/1091309?*

[2] acbTaskMan Version: 2009. *http://www.acbpocketsoft.com/Products/acbTaskMan/acbTaskMan-Overview-7.html*.

[3] Aimar, L., Merritt, L., Petit, E., Chen, M., Clay, J., Rullgrd, M. and Al., E. 2005. x264 - a free h264/AVC encoder.

[4] Akamatsu, C., Mayer, C. and Farrelly, S. 2006. An investigation of two-way text messaging use with deaf students at the secondary level. *Deaf Studies and Deaf Education*. 11, 1 (2006), 120–131.

[5] AndroSensor: 2013. *http://www.fivasim.com/androsensor.html*.

[6] Apple - QuickTime - Download: *http://www.apple.com/quicktime/download/*. Accessed: 2013-04-02.

[7] ARM Information Center: *http://infocenter.arm.com/help/index.jsp?topic=/com.arm.doc.ddi0419c/index.html*. Accessed: 2012-05-23.

[8] Arons, B. 1997. SpeechSkimmer: A system for interactively skimming recorded speech. *Proc. CHI* (1997), 3–38.

[9] Asterisk: 2014. *http://www.asterisk.org/*. Accessed: 2014-01-04.

[10] AT&T: *http://www.att.com/shop/wireless/data-plans.html#fbid=027qt05YFJ6*. Accessed: 2012-04-23.

[11] Bae, S., Pappas, T.N. and Juang, B. Spatial resolution and quantization noise tradeoffs for scalable image compression. *ICASSP* II–945–II–948.

[12] Barnlund, D. 1970. *A transactional model of communication*. Harper and Row.

[13] Baron, J. and Thurston, I. 1973. An analysis of the world-superiority effect. *Cognitive Psychology*. 4, 2 (1973), 207–228.

[14] Battison, R. 1978. *Lexical borrowing in American Sign Language*.

[15] Berlo, D.K. 1960. *The Process of Communication*. Holt, Rinehart, & Winston.

[16] Bornstein, H. 1978. Sign language in the education of the deaf. *Sign Language of the Deaf*. (1978), 333–361.

[17] Cavender, A., Ladner, R. and Riskin, E. 2006. MobileASL: Intelligibility of sign language video as constrained by mobile phone technology. *Proc. ASSETS* (2006), 71–78.

[18] Chang, Y., Carney, T., Klein, S., Messerschmitt, D. and Zakhor, A. 1998. Effects of temporal jitter on video quality: assessment using psychophysical methods. *Human Vision and Image Processing* (1998).

[19] Chen, B. 2013. AT&T allows FaceTime for limited data users. What about unlimited? *The New York Times*.

[20] Chen, J.Y.C. and Thropp, J.E. 2007. Review of low frame rate effects on human performance. *IEEE Trans. on Systems, Man, and Cybernetics - Part A: Systems and Humans*. 37, 6 (Nov. 2007), 1063–1076.

[21] Chen, Z. and Ngan, K. 2007. Recent advances in rate control for video coding. *Signal Processing: Image Communication*. 22, 1 (2007), 19–38.

[22] Cherniavsky, N., Chon, J., Wobbrock, J.O., Ladner, R. and Riskin, E. 2009. Activity Analysis Enabling Real-Time Video Communication on Mobile Phones for Deaf Users. *UIST* (2009).

[23] Cherniavsky, N., Chon, J., Wobbrock, J.O., Ladner, R. and Riskin, E. 2007. Variable Frame Rate for Low Power Mobile Sign Language Communication. *Proc. ASSETS* (Tempe, AZ, 2007), 163–170.

[24] Chon, J. 2011. *Real-time sign language video communication over cell phones*. University of Washington.

[25] Ciaramello, F., Cavender, A., Hemami, S., Riskin, E. and Ladner, R. 2006. Predicting intelligibility of compressed american sign language video with objective quality metrics. *Workshop on Video* (2006), 1–4.

[26] Ciaramello, F. and Hemami, S. 2011. A computational intelligibility model for assessment and compression of American sign language video. *IEEE Trans. IP*. 20, 11 (2011).

[27] Cicco, L., Mascolo, S. and Palmisano, V. 2008. Skype video responsiveness to bandwidth variations. *NOSSDAV* (2008).

[28] Claypool, M. and Tanner, J. 1999. The effects of jitter on the perceptual quality of video. *Multimedia* (1999).

[29] Coates, J. and Sutton-Spence, R. 2001. Turn-taking patterns in deaf conversations. *Social Linquistics*. (2001), 507–529.

[30] Collins, S. and Petronio, K. 1998. What happens in tactile ASL? *Pinky Extension and Eye Gaze: Language Use in Deaf Communities*. (1998), 18–37.

[31] Convo: 2011. *https://www.convorelay.com/*.

[32] Costs associated with using FaceTime: 2013. *http://www.ilounge.com/index.php/articles/comments/costs-associated-with-using-facetime/*.

[33] Crk, I., Albinali, F., Gniady, C. and Hartman, J. 2009. Understanding energy consumption of sensor enabled applications on mobile phones. *IEEE Engr. Med & Bio* (2009).

[34] Cumming, C. and Rodda, M. 1989. Advocacy, prejudice, and role modeling in the Deaf community. *Social Psychology*. 1, 129 (1989), 5–12.

[35] Design, V.S. Understanding image-interpolation techniques.

[36] Ding, W. and Liu, B. 1996. Rate control of MPEG video coding and recording by rate-quantization modeling. *IEEE Trans. Circuits Syst. Video Technol*. 6, 1 (1996), 12–20.

[37] Feghali, R., Speranza, F., Wang, D. and Vincent, A. 2007. Video quality metric for bit rate control via joint adjustment of quantization and frame rate. *IEEE Trans. on Broadcasting*. 53, 1 (Mar. 2007), 441–446.

[38] Fitzgerald, D. 2013. How much smartphone data do you really need. *The Wall Street Journal*.

[39] Foulds, R. a 2004. Biomechanical and perceptual constraints on the bandwidth requirements of sign language. *IEEE Trans. Neural Syst. Rehabil. Eng.* 12, 1 (Mar. 2004), 65–72.

[40] Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*. 32, 200 (1937), 675–701.

[41] Gatehouse, S. and Gordon, J. 1990. Response times to speech stimuli as measures of benefit from amplification. *Brit J Audiol*. 24, 1 (1990), 63–68.

[42] Gibbs's Phenomena: 2010. *http://cns.org/content/m10092/latest*.

[43] H.264: Advanced video coding for generic audiovisual services: *http://www.itu.int/rec/T-REC-H.264*. Accessed: 2012-05-24.

[44] Harkins, J., Wolff, A., Korres, E., Foulds, R. and Galuska, S. 1990. Intelligibility experiments with a feature extration system designed to simulate a low-bandwidth video telephone for deaf people. *RESNA* (1990), 92–95.

[45] Harrigan, K. 1995. The SPECIAL system: self-paced education with compressed interactive audio learning. *Journal of Research on Computing in Education*. 3, 27 (1995), 361–370.

[46] Hart, S.G. and Staveland, L.E. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload*. (1988).

[47] Hecker, M. and Stevens, K. 1966. Measurements of reaction time in intelligibility tests. *JASA*. 39, 6 (1966), 1188–1189.

[48] Heiman, G.W. and Tweney, R.D. 1981. Intelligibility and comprehension of time compressed sign language narratives. *Journal of Psycholinguistic Research*. 10, 1 (Jan. 1981), 3–15.

[49] Higgins, J.J. and Tashtoush, S. 1994. An aligned rank transform test for interaction. *Nonlinear World*. 1, 2 (1994), 201–2011.

[50] Hogg, N., Lornicky, C. and Weiner, S. 2008. Computer-mediated communication & the Gallaudet community: a preliminary report. *American Annals of the Deaf*. 153, 1 (2008), 89–96.

[51] Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 6, 2 (1979), 65–70.

[52] Hooper, S., Miller, C., Rose, S. and Veletsianos, G. 2007. The effects of digital video quality on learner comprehension in an American sign language assessment environment. *Sign Language Studies*. 8, 1 (2007), 42–58.

[53] How much 4G data do you really need: 2013. *http://www.techradar.com/us/news/phone-and-communications/mobile-phones/how-much-4g-data-do-you-really-need--1176594*.

[54] How much data will Skype use on my mobile phone?: 2013. *https://support.skype.com/en/faq/FA10853/how-much-data-will-skype-use-on-my-mobile-phone*.

[55] IMSDroid-High Quality Video SIP/IMS client for Google Android: *http://code.google.com/p/imsdroid/*. Accessed: 2012-05-23.

[56] ISO 2003. ISO 9921: Ergonomics-Assessment of speech communication.

[57] Johnson, B.F. and Caird, J.K. 1996. The effect of frame rate and video information redundancy on the perceptual learning of American sign language gestures.

[58] Koul, R. 2003. Synthetic speech perception in individuals with and without disabilities. *Augmentative and Alternative Communication*. 19, 1 (2003), 49–58.

[59] Lane, H. 1992. *The mask of benevolence: disabling the deaf community*. Alfred A. Knopf, Inc.

[60] Larson, R. and Csikszentmihalyi, M. 1983. The experience sampling method. *New Directions for Methodology of Social and Behavioral Science*. 15, (1983), 41–56.

[61] Lebeter, J. and Saunders, S. 2010. The effects of time compression on the comprehension of natural and synthetic speech. *Working Papers of the Linguistics Circle of the University of Victoria* (2010), 63–81.

[62] Lin, W. and Dong, L. 2006. Adaptive downsampling to improve image compression at low bit rates. *IEEE Trans. IP*. 15, (2006).

[63] Lucas, C. and Valli, C. 2000. *Linguistics of American Sign Language: an introduction*. Gallaudet University Press.

[64] Maher, J. 1996. *Seeing language in sign: the work of William C. Stokoe*. Gallaudet University Press.

[65] Manoranjan, M.D. and Robinson, J. a 2000. Practical low-cost visual communication using binary images for deaf sign language. *IEEE transactions on rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*. 8, 1 (Mar. 2000), 81–8.

[66] Masry, M. and Hemami, S. 2003. CVQE: A metric for continuous video quality evaluation at low rates. *SPIE Human Vision and Electronic Imaging* (Santa Clara, CA, 2003).

[67] Merriam-Webster 2003. *The Merriam-Webster Dictionary*.

[68] Mobile growth driving out unlimited data: *http://www.pcworld.com/businesscenter/article/242376/mobile_growth_driving_out_unlimited_data.html*. Accessed: 2011-11-23.

[69] MobileASL. University of Washington: 2012. *http://mobileasl.cs.washington.edu/*.

[70] Moore, G. 1965. Cramming more components onto integrated circuits. *Electronics*. 38, 8 (1965).

[71] Muir, L.J. and Richardson, I.E.G. 2002. Video telephony for the deaf. *Proceedings of the tenth ACM international conference on Multimedia - MULTIMEDIA '02* (New York, New York, USA, Dec. 2002), 650.

[72] Munoz-Baell, I. and Ruiz, T. 2000. Empowering the deaf. *Epidemiology and Community Health*. 1, 54 (2000), 40–44.

[73] Nemethova, A., Ries, M., Zavodsky, M. and Rupp, M. 2006. PSNR-based estimation of subjective time-variant video quality for mobiles. *Measurement of Audio and Video Quality in Networks*. (2006).

[74] Nguyen, V., Tan, Y. and Lin, W. Adaptive downsampling/upsampling for better video compression at low bit rate. *IEEE Int. Sym. CS*.

[75] Omoigui, N., He, L., Gupta, A., Grudin, J. and Sanocki, E. 1999. Time-compression. *Proc. CHI* (New York, New York, USA, May 1999), 136–143.

[76] Ou, G. 2010. Estimate of network bandwidth for iPhone 4 FaceTime. *Digital Society*.

[77] Ou, Y., Ma, Z. and Wang, Y. 2009. A novel quality metric for compressed video considering both frame rate and quantization artifacts. *Workshop on Video Processing and Quality Metrics for Consumer Electronics* (2009).

[78] Paas, F., Tuovinen, J.E., Tabbers, H. and Van Gerven, P.W.M. 2003. Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *J. Educ.* 38, 1 (Mar. 2003), 63–71.

[79] Padden, C. and Humphries, T. 2005. *Inside deaf culture*. Harvard University Press.

[80] Pearson, D. 1981. Visual communication system for the deaf. *IEEE Trans. Commun.* COM-29, 12 (1981), 1986–1992.

[81] Pratt, R. 1981. On the use of reaction time as a measure of intelligibility. *Brit. J. Audiol.* 12, 4 (1981), 253–255.

[82] Purple VRS on Your Devices: *http://www.purple.us/*.

[83] Reagan, T. 1995. A social culture understanding of deafness: American Sign Language and the culture of deaf people. *Intercultural Relations*. 19, 2 (1995), 239–251.

[84] Reed, E. and Lim, J. 2002. Optimal multidimensional bit-rate control for video communication. *IEEE Trans. Image Process*. 11, 8 (2002), 873–885.

[85] Reicher, G. 1969. Perceptual recognition as a function of meaningfulness of stimulus material. *Experimental Psychology*. 81, 2 (1969), 275–280.

[86] Reid, G. 1988. The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. *Human Mental Workload*. 52, (1988), 185–218.

[87] Reilly, G., Jrad, A., Nagarajan, R., Brown, T. and Conrad, S. 2006. Critical infrastructure analysis of telecom for natural disaters. *Telecom. Network Strategy and Planning* (2006), 1–6.

[88] Richardson, I. 2004. vocdex: H.264 tutorial white papers.

[89] Saks, A. and Hellström, G. 2006. Quality of conversation experience in sign language , lip - reading and text. *ITU-T Workshop on End-to-end QoE/QoS* (Geneva, 2006).

[90] Schlachter, F. 2009. No Moore's Law for batteries. *Proc. National Academy of Science of the USA*. 10, 14 (2009).

[91] Seitz, P. and Rakerd, B. 1997. Auditory stimulus intensity and reaction time in listeners with longstanding sensorineural hearing loss. *Ear Hearing*. 18, 6 (1997), 502–512.

[92] Shannon, C.E. 1948. A mathematical theory of communication. *The Bell System Technical Journal*. 27, 379-426 (Jan. 1948), 623–656.

[93] Skype Statistics: 2012. *http://www.statisticbrain.com/skype-statistics*.

[94] Skype: *http://www.skype.com/intl/en-us/home*. Accessed: 2012-04-23.

[95] Song, H. and Kuo, C. 2001. Rate control for low-bit-rate video via variable-encoding frame rates. *IEEE Trans. Circuits Syst. Video Technol.* 11, 4 (2001), 512–521.

[96] Sorenson Communications: *http://www.sorenson.com/*.

[97] Sosnowski, T. and Hsing, T. 1983. Toward the conveyance of deaf sign language over public telephone networks. *RESNA* (1983).

[98] Sperling, G. 1981. Video transmission of American Sign Language and finger spelling: present and projected bandwidth requirements. *IEEE Transactions on Communications*. 29, 12 (Dec. 1981), 1993–2002.

[99] Sperling, G., Landy, M., Cohen, Y. and Pavel, M. 1985. Intelligible encoding of ASL image sequences at extremely low information rates. *Computer Vision Graphics, and Image Processing*. 31, (1985), 335–391.

[100] Thu, H. and Ghanbari, M. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronic Letters*. 44, 13 (2008), 800–801.

[101] T-Mobile: *http://www.t-mobile.com/cell-phone-plans/individual.html#lshop_plans_1*.

[102] Tran, J.J., Kim, J., Chon, J., Riskin, E., Ladner, R. and Wobbrock, J.O. 2011. Evaluating quality and comprehension of real-time sign language video on mobile phones. *Proc. ASSETS* (2011), 115–122.

[103] Traum, D. and Hinkelman, E. 1992. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*. 8, (1992), 575–599.

[104] Tsang, P. and Velazquez, V. 1996. Diagnosticity and multidimensonal subjective workload ratings. *Ergonomics*. 39, 3 (1996), 358–381.

[105] Verizon Wireless: *http://www.verizonwireless.com/b2c/index.html*. Accessed: 2012-04-23.

[106] Video is fastest growing mobile data traffic source: 2013. *http://www.humanipo.com/news/36341/video-is-fastest-growing-mobile-data-traffic-source/*.

[107] Wang, R., Chien, M. and Chang, P. Adaptive down-samping video coding. *Proc. SPIE 7542*.

[108] Wang, Y. and Ou, Y. 2012. Modeling rate and perceptual quality of scalable video as functions of quantization and frame rate and its application in scalable video adaptation. *IEEE Trans. on Circuits and Systems for Video Technology* (2012), 671–682.

[109] Wang, Z., Bovik, A. and Lu, L. 2002. Why is image quality assessment so difficult? *ITASS* (2002), 3313–3316.

[110] Watson, A. and Sasse, M.A. 1998. Measuring perceived quality of speech and video in multimedia conferencing applications. *Multimedia* (1998), 55–60.

[111] Weber, E.. 1834. De pulsu, resorptione, auditu et tactu. *Anatationes anatomicae et physiologicae*. (1834).

[112] Wiegang, T., Schwarz, H., Joch, A., Kossentini, F. and Sullivan, G. 2003. Rate-constrained coder control and comparison of video coding standards. *IEEE Trans. CSVT*. 13, 7 (2003), 688–703.

[113] Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*. 1, 6 (1945), 80–83.

[114] Winkler, S. and Mohandas, P. 2008. The evolution of video quality measurement: from PSNR to hybrid metrics. *IEEE Trans. Broadcast.* 54, 3 (2008), 660–668.

[115] Wobbrock, J.O., Findlater, L., Gergie, D. and Higgins, J.J. 2011. The Aligned Rank Transform for nonparametric factorial analyses using only ANOVA procedures. *Proc. CHI* (Vancouver, BC, Canada, 2011), 143–146.

[116] Wright, R., Spanner, M. and Martin, M. 1981. Pilot experiments with a reaction time audiometer. *Brit J Audiol*. 15, 4 (1981), 275–281.

[117] Yadavalli, G., Hemami, S. and Masry, M. 2003. Frame rate preferences in low bit rate video. *IEEE Trans. IP* (2003), 441–444.

[118] Yang, K., Guest, C., El-Maleh, K. and Das, P. 2007. Perceptual temporal quality metric for compressed video. *IEEE Trans. on Multimedia*. 9, (2007), 1528–1535.

[119] ZVRS: *http://www.zvrs.com/z-series/*.

Appendix A-Press Coverage

1. Top EE Female Graduate Students Attend Grace Hopper Conference, Land Prestigious Positions. University of Washington EE Department News. April 2, 2014.

2. Mimi Gan. UW 360 UW Women Engineers. University of Washington TV. February 2014.

3. Lauren Clark. **Finding 'rising stars' in EECS**. MIT-hosted conference encourages the top women studying electrical engineering and computer science nationwide to become professors. MIT news, November 18, 2013.

4. Becky Chappel. **19 Companies Create Innovative Products with Google Technologies.** Google I/O Sandbox Case Studies, June 24, 2011.

5. Brian Callanan. **Breakthrough For Deaf, Hard Of Hearing: A Mobile Phone Call**. Q13 FOX News, August 25, 2010.

6. Jesse Emspak. **Bringing Mobile Phones To Those Who Can't Hear**. International Business Times, August 19, 2010.

7. Jean Enersen. **Deaf, hard-of-hearing UW students test sign language cell phone**. KING 5 News/HealthLink, August 18, 2010. *Click Link for Video*

8. Elisa Jaffe. **UW working on app to help deaf, hard of hearing**. KOMO 4 News, August 18, 2010.

9. Matt Hamblen. **Mobile sign language being developed at UW.** Computerworld, August 17, 2010.

10. Priya Ganapati. **Deaf Students Test Sign Language on Smartphones**. Wired: Gadget Lab blog, August 17, 2010.

11. Hannah Hickey. **Deaf, hard-of-hearing students do first test of sign language by cell phone**. University of Washington News, August 16, 2010.

Appendix B- Web study: Perceived Video Intelligibility

English sentences

P1) Yesterday evening at sunset, it was cold and raining outside.

P2) In the summer, my father and I like to go to the mountains to fish.

1) I use my cellphone to text message my friends often.

2) My daughter, Stephanie, rides the bus to school every day.

3) My mother likes dogs that don’t bark.

4) My brother, Jason plays football twice a week.

5) My favorite season is summer because it is always sunny and warm outside.

6) I enjoy watching TV and movies with captions.

7) Today is a beautiful day, the sun is shining and there are no clouds in the sky.

8) Every day I get up at 7 o’clock to eat breakfast and drink a big cup of coffee.

9) My favorite ice cream is chocolate.

10) For my birthday, my friend, David, gave me a digital camera.

11) During the weekend, I like to go bowling with my friends.

12) Next month I will be traveling to Hawaii on vacation.

13) When it’s sunny outside, I enjoy riding my motorcycle because I can drive fast.

14) My parents bought a new house two years ago.

15) My friend, Susie and her daughter went to the zoo and saw lions.

16) Earlier today I went to the grocery store and bought milk, eggs, and bread.

Glossed ASL sentences and English sentences used in web-study

Gloss ASL Sentences	**English Sentences**
MY SISTER SAY(said) SHE[point-right] ORDER(ordered) COFFEE HOWEVER(but) THE[t-twist] WAITER BRING(brought) TEA	My sister said she ordered coffee, but the waiter brought tea.
YESTERDAY MY BROTHER AND I GO(went) TO A[A-move-right] CONCERT[music-show?] #IT WAS COUNTRY[rural/farm] MUSIC	Yesterday, my brother and I went to a concert. It was country music.
TOMORROW IS A[A-move-right] BIG PARTY AT MY OFFICE MY COLLEAGUE(coworkers) INVITED ME 2 WEEKS AGO	Tomorrow is a big party at my office. My coworkers invited me 2 weeks ago.
PAST(last) FRIDAY MY WIFE DRIVE(drove) MY CAR TO WORK	Last Friday, my wife drove my car to work.
Many people, they go camping forest. Various states: 1st Colorado, 2nd Wyoming, 3rd California, 4th Washington	Many people go camping in the forest from various states: 1) Colorado, 2) Wyoming, 3) California, 4) Washington
2 students make computer program. It name #chess. Program plays game chess. Use #super computer	2 students make computer program named chess. Program that play's chess game use super computer
Rice important food for many people world. 3 months ago, cost increase double. Why? No rain.	Rice is an important food for many people in the world. 3 months ago, the cost doubled because no rain.
Last fall, my aunt #Sally She plan #Garage #sale. Know++? Set-up table outside house.	
Like: dancing, piano, watch #DVD. And love chatting.	

YESTERDAY I VISIT(visited) MY MOTHER AND FATHER THEY[on-right-side] ARE FINE[adjective]	Yesterday, I visited my mother and father. They are fine.
TODAY IT WAS SNOW(snowing) THE[t-twist] STUDENT(students) STAY(stayed) HOME TODAY FROM SCHOOL	Today it was snowing. The students stayed home from school today.
MY BROTHER SAY(said) HE[point-right] ORDER(ordered) PIZZA HOWEVER(but) PIZZA NEVER ARRIVE(arrived)	My brother said he ordered a pizza. But the pizza never arrived.
Today library me go library 3 book-book check-out me read.	Today, I went to the library and checked out 3 books to read.
During weekend hiking me love go hiking in mountains with friends.	During the weekend, I like to go hiking in the mountains with my friends.
YESTERDAY MY SISTER VISIT(visited) ME AT WORK[noun]	Yesterday, my sister visited me at work.
I LIKE #TO GO TO MOVIE(movies) AND GO TO PLAY(plays)[theatre-show]	I like to go to the movies and go to plays.
PAST(last) NIGHT I MEET(met) #CHARLIE I WOULD LIKE #TO CALL[verb,telephone-call] HE(him)[point-right] HOWEVER(but) HE[point-right] FORGET(forgot) #TO GIVE ME HIS[on-right-side] NUMBER	Last night, I met Charlie. I would like to call him, but he forgot to give me his number.

Appendix C- Web Study: Response-Time and Mental Effort Relationship

Survey Sentences and Questions Used

	English Sentences	Glossed ASL Sentences	Questions	Answers
1	My sister and I went to a coffee shop called the Coffee Bean. She ordered coffee, but the waiter brought tea.	Coffee Shop name Coffee Bean my sister we go finished. She order coffee wrong waiter gave tea	What did my sister order?	A) coffee B) water C) tea D) juice
2	Yesterday, my brother and I went to a concert located in a park. The concert was country music.	Yesterday concert where park my brother we go what music country	Where was the concert?	A) a stadium B) a concert hall C) the country D) a park
3	Tomorrow is a big party at my office. My coworkers invited me 2 weeks ago.	Tomorrow a big party. Where? My office. Two weeks ago my work people invited me.	When was I invited to the party?	A) yesterday B) 1 month ago C) 1 week ago D) 2 weeks ago
4	Last night there was a big windstorm. Luckily there was no power outage due to the storm.	Last night wind storm "big" none power outage why storm. Lucky!	Which of the following did the storm definitely not cause?	A) power outage B) fallen trees C) flooding D) hail
5	I own a bicycle, car, and truck. Yesterday, my wife drove the car to work and I rode my bike to work.	We have 3: bike car truck. Yesterday my wife car go work me work me ride bike yesterday	Which vehicle did my wife use to get to work?	A) bus B) bicycle C) truck D) car

6	I accomplished a lot yesterday. I visited the library, grocery store, and post office.	Yesterday finally went 3: library, food store, post office	Which of the following places did I not visit yesterday?	A) the bank B) the grocery store C) the the library D) the post office
6	My favorite ice cream flavor is strawberry when given the choice between mint, chocolate, strawberry, and vanilla.	4 ice cream: mint, chocolate, strawberry, vanilla. 3rd my favorite	Which of the following flavors of ice cream is my favorite?	A) vanilla B) chocolate C) strawberry D) mint
7	My sister, Lisa, is hearing and currently learning how to speak French . She fluently signs ASL and fluently speaks English and Spanish.	my sister Lisa hearing now learn French speak: 3: ASL skilled speak Spanish and English	Which language is my sister not fluent in?	A) French B) Spanish C) English D) ASL
8	My son, Charlie, who is 16, usually rides his bike to school, but today he rode the school bus.	my son Charlie-16 everyday ride bike school but today ride yellow bus	How old is my son, Charlie?	A) 12 B) 16 C) 14 D) 13
9	In the summer, my father and I like to go to the lake to catch fish. We compete to see who can catch the most fish.	summer my father me together fish where lake. 2-us compete see one catch fish	What do my father and I compete for while fishing?	A) the first fish B) the most fish C) the largest fish D) the strangest fish
10	During the weekend, I like to go bowling, watch movies, and read books. When my friends are available I like to go	weekend 3: like bowling, movies, and read books. My friends free we go with them bowling.	Which activity did I not mention?	A) watching movies B) bowling C) reading books

	bowling with them.			D) biking
11	Today I ate cereal for breakfast because I was in a hurry. When I have more time in the morning I usually eat eggs, toast, and drink coffee.	Today morning eat cereal why hurry. When me relax eat eggs, toast, drink coffee.	What did I eat today?	A) bacon B) eggs C) cereal D)toast
12	Winter is my favorite season because it snows outside. When it snows I can go skiing and snowboarding in the mountains.	favorite season winter. Why snow. When snow me ski snowboard where mountains	Which part of winter did I not mention?	A) snowshoeing B) the snow C) skiing D) snowboarding
13	Last night, I met Daniel, who is a coworker. I would like to contact him, but he forgot to give me his e-mail.	last night my work person me met Daniel me contact hum why he forgot give me his email	What did my coworker forget to give me?	A) e-mail B) name C) address D) phone number
14	Yesterday, I visited my mother, father, and their cat named Arrow. My parents are fine, but their cat was sick.	yesterday, me come see my mother dad cat-arrow my mom-dad fine but cat sick	Who was sick?	A) me B) my mother C) my father D) their cat
15	When given the choice to drink water, juice, or soda pop. I choose to drink water.	3: drink water, juice, pop, me pick water	Which drink did I not mention?	A) milk B) water C) juice D) pop
16	I like to go to the movies and my sister likes to go to plays.	movies me like go, plays my sister like	What does my sister like to go to?	A) concerts B) plays C) opera

				D) movies
17	My family went to the zoo and saw lions, flamingos, and gorillas. We didn’t see the pandas because there was a long line to see them.	zoo my family "group" saw 3: lions, #flamingo, gorillas not yet see panda why long line	Which animals did we not see at the zoo?	A) lions B) pandas C) flamingos D) gorillas
18	My daughter likes small dogs. Around big dogs she gets scared easily.	small dogs my daughter like but big dogs my daughter see scared	Who is afraid of large dogs?	A) my mother B) my aunt C) my daughter D) my son
19	Two weeks ago, my flight to New York was delayed because of heavy rain. I was lucky my flight wasn’t cancelled because other flights were cancelled due to dense fog.	two week go NY me fly delay why heavy rain airplane not cancelled lucky why many other flights cancelled why fog thick	Why was my flight delayed?	A) wind B) fog C) snow D) rain

Appendix D- Field Study: Using MobileASL Application in the Wild

Pre-Deployment Questionnaire

1. In a typical week, do you send or receive one or more text messages using a cell phone?

    Yes

    No (please skip ahead to #12)

2. In what year did you first use a cell phone to send or receive text messages?

    ____

3. What would you say is your familiarity with texting?

    1 2 3 4 5

    Unfamiliar Familiar

4. In your opinion, which term best describes the person you text most often? Please circle only one response.

    Deaf

    deaf

    hard-of-hearing

    hearing

5. Typically, who do you text most often? Please circle only one response.

    a. Girlfriend/boyfriend
    b. Family
    c. My close friends
    d. Other friends or acquaintances
    e. A business

    Other (Whom? ____)

6. Which best describes the purpose of your text messages? Please circle only one response.

    Chatting for fun

    Coordinating activities, plans, or times

    Checking in for safety

    Other (For what purpose? ____)

VIDEO CHAT

7. Which best describes your use of video chat technology? Please circle only one response. Examples of programs that contain video chat technology include, but are not limited to, AOL Instant Messenger (AIM), Skype, Google Chat, and Microsoft Instant Messenger.

   More than twice a day

   Once or twice a day

   Once every few days

   About once a week

   Less than once a week

   I don’t currently use video chat technology, but I have in the past (please answer questions a and b)

   a. In what year did you first use video chat technology? ____________
   b. In what year did you stop <u>making or receiving</u> one or more calls per week through video chat technology? _________ (please skip ahead to question #9)

   I have never used video chat technology. (please skip ahead to question #9)

   Other (How often? ____)

8. In what year did you first use video chat technology?

   ____

9. What would you say is your familiarity with video chat technology?

   12345

   Unfamiliar Familiar

10. Typically, when you use video chat technology, where is the technology located? Please circle only one response.

    Home

    Work

    School

    Other (Where is it located? ____)

11. Typically, how long are your conversations when you use video chat technology? Please circle only one response.

    Under 15 minutes

Between 15-30 minutes

Between 31-45 minutes

Between 46 minutes to one hour

Over one hour (How long? ____)

12. Typically, who do you chat with most often using video chat technology? Please circle only one response.
    a. Girlfriend/boyfriend
    b. Family
    c. My close friends
    d. Other friends or acquaintances
    e. A business

    Other (Whom? ____)

13. In your opinion, which term best describes the person you chat with most often using video chat technology? Please circle only one response.

    Deaf

    deaf

    hard-of-hearing

    hearing

14. Which best describes the purpose of your calls using video chat technology? Please circle only one response.

    Chatting for fun

    Coordinating activities, plans, or times

    Checking in for safety

    Other (For what purpose? ____)

VIDEO PHONE WITHOUT VRS

15. In a typical week, do you make or receive one or more calls using a video phone, without VRS?

    Yes

    No (please skip ahead to #18)

16. In what year did you first use a video phone without VRS?

    ____

17. What would you say is your familiarity with video phone technology?

12345

Unfamiliar Familiar

18. Typically, how often do you use a video phone, without VRS? Please circle only one response.

More than twice a day

Once or twice a day

Once every few days

About once a week

Less than once a week

Other (How often? ____)

19. Typically, when using a video phone, without VRS, where is the technology located? Please circle only one response.

Home

Work

School

Other (Where is it located? _______________________________________)

20. Typically, how long are your conversations when you use a video phone, without VRS? Please circle only one response.

Under 15 minutes

Between 15-30 minutes

Between 31-45 minutes

Between 46 minutes to one hour

Over one hour (How long? _______________________________________)

21. Typically, whom do <u>you speak with</u> most often using a video phone without VRS? Please circle only one response.

a. Girlfriend/boyfriend

b. Family

c. My close friends
d. Other friends or acquaintances
e. A business
Other (Whom? ____

22. In your opinion, which term best describes the person you speak with most often using a video phone without VRS? Please circle only one response.

Deaf

deaf

hard-of-hearing

hearing

23. Which best describes the purpose of your calls using a video phone without VRS? Please circle only one response.

Chatting for fun

Coordinating activities, plans, or times

Checking in for safety

Other (For what purpose? ____)

VIDEO PHONE – THROUGH VRS

24. In a typical week, do you make or receive one or more calls through a Video Relay Service (VRS)?

Yes

No (please skip ahead to #53)

25. In what year did you first make or receive a call through VRS? ____________

26. What would you say is your familiarity with VRS?

12345

Unfamiliar Familiar

27. Typically, how often do you make or receive a call through VRS? Please circle only one response.

More than twice a day

Once or twice a day

Once every few days

About once a week

Less than once a week

Other (How often? ____)

28. Typically, when you make or receive a call through VRS, where is the technology located? Please circle only one response.

Home

Work

School

Other (Where is it located? ____)

29. Typically, how long are your conversations when you make or receive a call through VRS? Please circle only one response.

Under 15 minutes

Between 15-30 minutes

Between 31-45 minutes

Between 46 minutes to one hour

Over one hour (How long? ____)

30. Typically, whom do you speak with through VRS? Please circle only one response.

a. Girlfriend/boyfriend
b. Family
c. My close friends
d. Other friends or acquaintances
e. A business

Other (Whom? ____)

31. In your opinion, which term best describes the person you speak with most often through VRS? Please circle only one response.

Deaf

deaf

hard-of-hearing

hearing

32. Which best describes the purpose of your calls through VRS? Please circle only one response.

Chatting for fun

Coordinating activities, plans, or times

Checking in for safety

Other (For what purpose? ____)

MOBILE

33. Have you used mobile video technology in the past? Examples include, but are not limited to, i711, MVP™ (from Hands On VRS®), ZVO Mobile, VPAD , and Viable Vision.

Yes

No (please skip ahead to #31)

34. In what year did you first use mobile video technology? ___________________________

35. What mobile video device(s) have you used? ______________________________________

BACKGROUND INFORMATION

To determine how people of different backgrounds respond to these questions, we'd like a few facts about you.

36. What is your gender?Please circle only one response.

Female

Male

Choose not to answer

37. What is your age?

____

38. With what language do you prefer to communicate?

ASL

English

Other (Please specify____)

39. What language do you prefer to use with family?

ASL

English

Other (Please specify____)

40. What language do you prefer to use with friends?

ASL

English

Other (Please specify____)

41. Are you fluent in ASL?

Yes

No

42. How many years have you spoken ASL?

____

40. From the choices below, how would you describe yourself?

a. Deaf

b. deaf

c. Hard of hearing

Interview Questions

In this interview, I'll be asking you why you prefer some methods of distance communication over others and what you value about those methods. We'll be talking about times when you can't meet someone face to face but you want to or need to communicate with them. In the end, I'll ask you what you expect of the study and you can ask any questions. I'll be recording the interview so that I have an accurate record of your responses, but you can ask me to stop recording. You can also decline to answer any questions. This should take about 20 minutes.

1. What do you like most about your mobile phone?
2. Warm up, relationship with the technology, general feelings towards it, perception of its role in everyday life, benefits
3. What do you like most about communicating with your mobile?
4. What do you dislike about communicating with your mobile?
5. How do you prefer to communicate with people at a distance?
6. When you decide to use that method, what are some of the reasons you've had for making that choice?
7. What do you **like** about that method? Why?
8. What do you **dislike** about that method? Why?
9. What **features** do you like best about **_______**?
10. What **features** of **________**do you dislike?
11. Can you recall the **best experience** you've had with this technology?
12. Can you recall the **worst experience** you've had with this technology?
    i. What makes someone a good (_) partner? What are your **expectations**? Why?
13. What role does it play in your **relationships**? Please give an example.
14. Overall, how would you **rate this technology** on a scale from 1 to 5?
15. In general, are you satisfied with your mobile phone? do you feel your mobile meets your distance communication needs?
16. What do you do if your cell phone battery is low?
17. What are your expectations of the study?
18. Why did you decide to participate in this research? How would you like to see this research used and shared?

Unobtrusive Logging Items

- **Calls**
  - Call type (incoming/outgoing)
  - If call was missed (if it was incoming) or was never answered (if it was outgoing) User name, IP address, Phone #, Device ID of person being called *(*User name, IP address, Phone #, and Device ID of person calling Call Duration Time call was requested
    - "requested" means someone pressed the "Sign" button so they could talk to someone.
  - Time call was started
    - "started" means the call actually started, with video transmission.
  - Average encoding FPS
  - Average decoding FPS
  - Packet loss #
  - Whether or not it was a VRS call
  - *Call declined*
- **Texts**
  - Text type (incoming/outgoing)
  - # of characters
  - If text is incoming
    - Phone # of sender
      - Since texts can be from both MobileASL and non-MobileASL phones, phone # is the only information we can collect about the other user.
    - Time when a text is received
  - If text is outgoing
    - Phone # of recipient
    - Time when a text is sent
- **Battery**
  - Time when phone gets plugged in
  - Time when phone gets unplugged (or the duration of the plug-in)
  - Is the phone plugged in by USB or to a wall-plug?
  - Time when battery level changes
  - percentage of battery life when call starts
  - percentage of battery life when call ends
- **Bit rate**
  - number of bits sent during a call
  - bit rate at start of call

    - bit rate at end of call
- **Connectivity**
    - Time of change
    - New IP Address
    - Type of connection (3G, WiFi, etc)
- **Program**
    - Time when MobileASL is started
        - When a phone goes online in the database, a trigger creates a new row in the Program Logging table with “time online” = current time
    - Time when phone goes offline
        - When a phone goes offline in the database (as detected by the last time online table), a trigger fills in the last blank “time offline” field in the Program Logging table with the current time.
- **Location**
    - Time when location changes by more than X feet

Experience Sampling Triggers and Questions

- **After a call (after_call) (3 possible questions + 1 more in the case of battery savings)**
    - Which phrase best describes the main purpose for the call? (a question for both callers)
        - Chatting for fun
        - Coordinating activities, plans, or times
        - Checking in for safety or whereabouts
        - Other
    - During the call, I had to repeat myself so the other person could understand what I signed.
        - Yes/No
        - (Branch Question) Reason for repeating
            - There was a delay in the other person seeing something I had signed
            - Video quality was blurry or choppy
            - Lighting was a problem
            - Other
    - Which best describes where you are right now?
        - My dorm
        - At school
        - Public place or business
        - On the bus
        - Other
- **After declining a call ^ (decline_call) 1 question**
    - Which best describes the reason you declined the call?
        - Declined by accident
        - Battery was too low
        - I didn’t want to be interrupted
        - I was concerned about privacy
            - (branch questions) I wanted to keep my conversation private
            - I didn’t want the person to know my location
            - I was self-conscious about my appearance
            - Other
        - Other
- **After acknowledging a missed call ^ (missed_call) 1 question**
    - Which best describes the reason you missed the call?
        - Missed by accident
        - Battery was too low
        - I didn’t want to be interrupted

- I was concerned about privacy
  - (branch questions) I wanted to keep my conversation private
  - I didn’t want the person to know my location
  - I was self-conscious about my appearance
  - Other
- Don’t remember
- Other

- **After closing out of reading a text ^ (receive_text) 3 questions**
  - We noticed you just read a text message. Which best describes with whom you were communicating?
    - Girlfriend/boyfriend
    - Family
    - My close friends
    - Other friends or acquaintances
    - A business
    - Other
  - We noticed you just read a text message. Which phrase best describes the main purpose of the text?
    - Chatting for fun
    - Coordinating activities, plans, or times
    - Checking in for safety or whereabouts
    - Other
  - Which best describes where you are right now?
    - My dorm
    - At school
    - Public place or business
    - On the bus
    - Other
- **After sending a text ^ (sent_text) 4 questions**
  - We noticed you just sent a text message. Which best describes with whom you were communicating?
    - Girlfriend/boyfriend
    - Family
    - My close friends
    - Other friends or acquaintances
    - A business
    - Other
  - We noticed you just sent a text message. Which phrase best describes the main purpose of the text?
    - Chatting for fun
    - Coordinating activities, plans, or times

        - Checking in for safety or whereabouts
        - Other
    - Which best describes why you chose to use text instead of MobileASL to communicate? (If possible, this question should be triggered every time we detect that they declined a call to the same number)
        - Texting is more private
        - Texting is faster
        - Texting is more reliable
        - Recipient doesn't use MobileASL
        - Other
    - Which best describes where you are right now?
        - My dorm
        - At school
        - Public place or business
        - On the bus
        - Other
- **After a call that used PC (pc_call) 1 question**
    - Which best describes with whom you were communicating? (move this)
        - Girlfriend/boyfriend
        - Family
        - My close friends
        - Other friends or acquaintances
        - Other
- **After a call that used VRS (vrs_call) 2 questions**
    - We noticed you just used Video Relay Service (VRS). Which best describes with whom you were communicating?
        - Girlfriend/boyfriend
        - Family
        - My close friends
        - Other friends or acquaintances
        - A business
    - We noticed you just used Video Relay Service (VRS). Which phrase best describes the main purpose of the VRS call?
        - Chatting for fun
        - Coordinating activities, plans, or times
        - Checking in for safety or whereabouts
        - Other
- Logging
    - If they use the privacy function
    - Changing from WiFi to 3G or whatever
    - Location (we just want to know if they were mobile)

Weekly Questionnaire

USEFULNESS

1. Typically, during the past 7 days, how would you rate the usefulness of MobileASL.

1 2 3 4 5 6 7

Not Applicable Not useful Very Useful

Please explain your answer.

2. On a scale from 1-7, where 1 is not useful and 7 is very useful, how would you rate the usefulness of MobileASL in the past 7 days? Please explain your answer.

3. This week, when I used a cell phone, I preferred using MobileASL to communicate.

1 2 3 4 5 6 7

Strongly Disagree Strongly Agree

Please explain your answer.

4. This week, when I used a cell phone, I preferred using MobileASL to communicate. Please rate your agreement with this statement on a scale from 1 to 7, where 1 is Strongly Disagree and 7 is Strongly Agree. Please explain your answer.