Human-Centered Approach Evaluating Mobile Sign Language Video Communication

Jessica J. Tran[1], Eve A. Riskin[1], Richard E. Ladner[2], Jacob O. Wobbrock[3]

[1]Electrical Engineering
DUB Group
University of Washington
Seattle, WA 98195 USA
{jjtran, riskin}@uw.edu

[2]Computer Science & Engineering
DUB Group
University of Washington
Seattle, WA 98195 USA
ladner@cs.washington.edu

[3]The Information School
DUB Group
University of Washington
Seattle, WA 98195 USA
wobbrock@uw.edu

ABSTRACT

Mobile video is becoming a mainstream method of communication. Deaf and hard-of-hearing people benefit the most because mobile video enables real-time sign language communication. However, mobile video quality can become unintelligible due to high video transmission rates causing network congestion and delayed video. My dissertation research focuses on making mobile sign language video more accessible and affordable without relying on higher cellular network capacity while extending cellphone battery life. I am investigating how much frame rate and bitrate of sign language video can be reduced before compromising video intelligibility. Web and laboratory studies are conducted to evaluate perceived intelligibility of video transmitted at low frame rates and bitrates. I also propose the *Human Signal Intelligibility Model* (HSIM) addressing the lack of a universal model to base video intelligibility evaluations.

Keywords

Intelligibility, comprehension, American Sign Language, bitrate, frame rate, video compression, web survey, communication model, Deaf community.

1. INTRODUCTION

Mobile video communication is becoming integrated into daily use; however, both cellular network capacity and battery life are major limiting factors for mainstream adoption. Transmitting real-time video requires large amounts of bandwidth and many cellular networks do not provide unlimited data plans, or are throttling down network speeds in response to high data consumption rates. Deaf and hard-of-hearing people benefit the most from using real-time mobile video communication because they communicate in sign language. American Sign Language (ASL) is signed in the United States (U.S.) and is a visual language with unique grammar and syntax independent of spoken languages. People who choose to communicate using mobile video consume network bandwidth faster than do average data users.

Currently, the international recommended standard to transmit intelligible sign language video is 25 frames per second (fps) at 100 kilobits per second (kbps) or higher [11]. However, cellular phone companies do not subsidize the extra cost of mobile video communication used by deaf and hard-of-hearing people. My dissertation research contributes to the MobileASL (American Sign Language) project [9] by applying electrical engineering, computer science, and human-computer interaction methods to make mobile sign language video communication more accessible and affordable to deaf and hard-of-hearing people.

Figure 1: Two participants holding a real-time sign language conversation using the MobileASL software-providing real-time video at extremely low bandwidths (30 kilobits per second at 15 frames per second).

The goal of my dissertation research is to use video compression algorithms to reduce bandwidth consumption and increase battery duration for mobile sign language video communication. My research will answer how much frame rate and bitrate of video quality can be reduced before intelligibility is compromised. These findings will make mobile video communication more accessible while providing intelligible content, reducing bandwidth consumption, and extending cell phone battery life. Finally, I present the *Human Signal Intelligibility Model* [14] addressing the lack of a universal model to base video intelligibility evaluations.

2. RELATED WORK

Providing real-time, two-way sign language video conversations at extremely low bandwidths has been the primary focus of the MobileASL project. In 2008, an experimental smart phone application, also called MobileASL, was created transmitting sign language video at 30 kilobits per second (kbps) at 8-12 frames per second

(fps) [9]. In the summer of 2010, a 3 week pilot field study was conducted evaluating everyday use of MobileASL among deaf and hard-of-hearing teenagers [7]. Figure 1 is an image of two participants signing to each other using MobileASL. Participants were given an HTC TyTNII cell phone with MobileASL installed and asked to communicate with each other as often as possible. On average 0-2 video calls were made per day and participants preferred using MobileASL over text messaging. Participants also found the physical phone too clunky and outdated with a short battery life lasting on average 2-3 hours after a fully charged battery. These and other findings further motivate the need for longer battery life for mainstream adoption of mobile video communication.

Transmitting video in real-time is computationally intensive leading to a quickly drained battery. Video quality is affected by cellular network congestion leading to video delay and reduced video quality. Applying video compression lowers video transmission rates; however, video quality and intelligibility are sacrificed. Video quality is often objectively measured using peak signal-to-noise ratio (PSNR), which measures quality of image reconstruction after lossy compression. However, PSNR does not reflect content intelligibility, which is most important for meaningful sign language communication. Intelligibility is often subjectively measured using perception or comprehension of content. Researchers have attempted to link higher objective quality to greater intelligibility [4, 5]. For example, Nemethova *et al.* [10] created a different rule-based algorithm that adapts the PSNR curve to mean opinion scores (MOS) by scaling, clipping, and smoothing PSNR results. Feghali *et al.* [6] 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.

3. DISSERTATION RESEARCH

3.1 HUMAN SIGNAL INTELLIGIBILITY MODEL

To date, there is not a standard method to evaluate video intelligibility, or a good communication model to use as a benchmark for evaluation. Often, intelligibility and comprehension are loosely defined and used interchangeably in video quality evaluations. Existing communication models focus on the communication channel itself [12] without considering the environment or the human sender and receiver. Models that have attempted to do so have been poorly defined and do not clearly identify the components of video intelligibility and comprehensibility [1,2]. I argue that intelligibility does not imply comprehension, but comprehension could imply intelligibility depending on certain conditions. Intelligibility depends on signal quality, specifically how the signal was captured, transmitted, received, and perceived by the receiver, including the environmental conditions affecting these steps. Comprehension relies on signal quality *and* the receiver having prerequisite

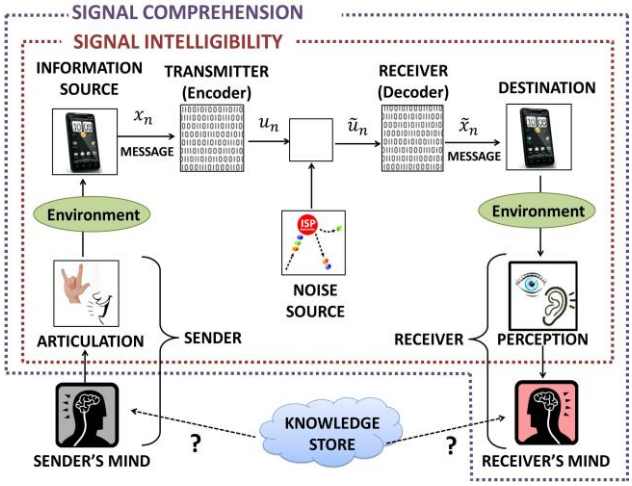

Figure 2: Block diagram of the *Human Signal Intelligibility Model*. Note that the components comprising signal intelligibility are a subset of signal comprehension, which is signal intelligibility *plus* the receiver's mind.

knowledge to process the information. Comprehension can infer intelligibility once common language knowledge is established between the sender and the receiver.

Using this insight to distinguish intelligibility from comprehension, I have developed an analytical model, the *Human Signal Intelligibility Model (HSIM)* [14], to base evaluation and quantify the effects of video compression on video intelligibility. The HSIM informs web and laboratory studies evaluating how much frame rate and bitrate can be reduced before sign language intelligibility is compromised.

3.2 WEB SURVEY DESIGN

The web study investigates perceived intelligibility of ASL video sentences transmitted at four low frame rates (1, 5, 10, 15 fps) and four low bitrates (15, 30, 60, 120 kbps), in a full factorial design, that is representative of what would be displayed on mobile devices. A preliminary web study is necessary before the laboratory study because more parameter settings were evaluated. The web study findings are influencing the frame rate and bitrate settings that will be implemented in the laboratory study.

The survey consisted of three parts and took 12-26 minutes per respondent to complete. 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. Part 3 asked demographic questions. Each video was shown once, *without* the option to repeat or enlarge the video, and then removed from the screen and replaced by two questions shown one at a time. Figure 3 is a screen shot of one video from the web survey.

After each video, participants rate how easy the video was to understand. Although comprehension is measured, participants are screened to ensure they are fluent in ASL and therefore comprehension is controlled for, allowing me

Figure 3: Screen shot of one video from web survey evaluating intelligibility of sign language video displayed at 15 frames per second at 30 kilobits per second.

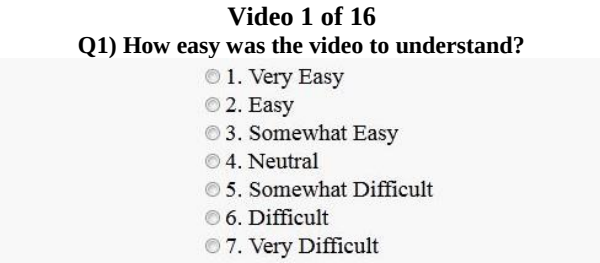

Figure 4: Example of question 1 shown in web survey.

Figure 5: Multiple choice comprehension question example.

to isolate intelligibility. Figure 4 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 5 is an example of a trivial comprehension question pertaining to the video shown. A four point multiple choice answer appeared with the corresponding images.

4. 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 was 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 indicated ASL as their native language and 11 of 38 people indicated having 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 own a smartphone and send text messages; 65 indicated they use video chat; and 53 use video relay services.

Table 1 and Figure 6 list the mean Likert score for question 1, where higher scores correspond to higher agreement with the ease of perceived understanding of video content. 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 to observe that intelligibility clearly suffered. Prior work investigating the impact of frame rate on perceived video quality acknowledged not selecting a low enough frame rate to explore in their study [3, 8]. Although transmitting video at 1 fps is not ideal for ASL conversations, we did notice that transmitting video at 1 fps and 15 kbps, which is the lowest bitrate, received the highest mean Likert score across all bitrates at 1 fps. This finding corroborates our earlier finding in [13] that people perceived the least amount of negative effects when the lowest frame rate and bitrate settings were applied.

We also discovered diminishing returns for videos displayed at 60 kbps and 120 kbps independent of frame rate. Figure 6 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.*). Our 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 bitrates. One would think that ASL, which is a temporal visual language, would require video communication to be transmitted at high frame rates; however, we discovered this is not the case at low bitrates. 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*. [3] However, their findings only reported a slight but significant main effect that frame rate influenced video intelligibility. Our results strongly affirm that ASL video intelligibility peaks at 10 fps across all bitrates. At a fixed low bitrate, more bits are allocated per frame at 10 fps *vs.* 15 fps, and this difference is noticeable enough to result in higher intelligibility. Our findings suggest that relaxing the recommended frame rate and bitrate to 10 fps at of 60 kbps may 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.

Table 1: Mean Likert score responses for ease of understanding video quality. Note *higher* Likert scores correspond to higher perceived intelligibility.

	Bitrate (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

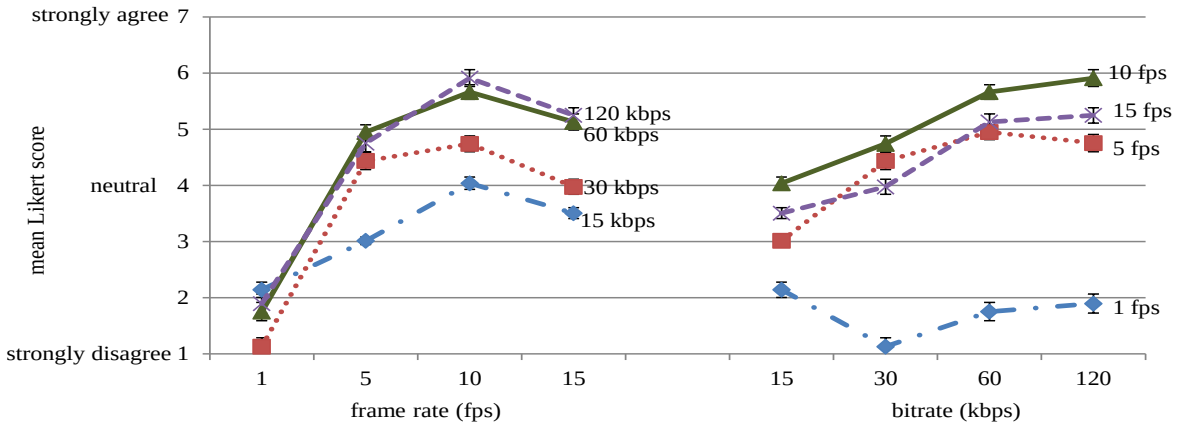

Figure 6: Plot of 7-point Likert ratings for participants' ease of understanding the video for each frame rate and bitrate averaged over all participants. Error bars represent ±1 standard error.

5. FUTURE WORK

The laboratory study will take a subset of parameter settings used in the web study to identify the minimum video quality settings needed for intelligible sign language communications while objectively measuring user intelligibility. Participants will be video recorded signing to each other using mobile phones transmitting real-time video at lowered video qualities. In post analysis, intelligibility will be measured by counting the number of repair requests; average number of turns associated with repair requests; and number of conversational breakdowns. Finally, battery life will be quantified once transmission rates are known. An interesting finding from the study will be identifying participants' adaptation techniques (if any) to compensate for lower video quality. I anticipate that participants' signing speed will slow down as the video frame rate is reduced. I also expect participants will use other strategies to compensate for lower video quality such as signing shorter phrases or signing slang terms.

6. CONCLUSION

This research makes three significant contributions: (1) making real-time mobile video communication accessible to deaf and hard-of-hearing people by reducing bandwidth consumption; (2) propose the Human Signal Intelligibility Model (HSIM) which addresses the lack of a universal definition for signal intelligibility, and (3) through rigorous empirical work validate the HSIM to determine video quality intelligibility tradeoffs.

With validation of the HSIM, I anticipate it will be used in future evaluations of intelligibility and comprehension of communication signals like other video streaming media-YouTube, Hulu, and Skype. I also anticipate that the knowledge gained on intelligibility of low video quality will influence user experience of mobile video communication. Mainly, users will be more empowered to control video quality rather than remain subject to cellular networks restrict data usage, which can lead to unintelligible content. Finally, these findings will help society because our work will (1) enable mobile video use by deaf and hard-of-hearing people and (2) serve as a concrete example of how engineers can benefit society and improve lives.

7. ACKNOWLEDGEMENTS

Thanks to Rafael-Sunny Rodriguez for building the web study infrastructure; Gerardo Di Pietro and Jason Smith for ASL video recordings; and our respondents. This work was funded by Google.

8. REFERECNES

[1] Barnlund, D.C. 2008. A transactional model of communication. *Communication Theory*. 47–57.

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

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

[4] Ciaramello, F. and Hemami, S. 2011. Quality versus intelligibility: studying human preferences for American sign language video. *SPIE Human Vision and Electronic Imaging*. 7865, (2011).

[5] Clark, H. and Brennan, S. 1991. Perspectives in Socially Shared Cognition, ch. Grounding in Communication. *American Psychological Association*. (1991), 127–149.

[6] 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.

[7] Kim, J., Tran, J.J., Johnshon, T., Ladner, R., Riskin, E. and Wobbrock, J.O. 2011. Effect of MobileASL on communication among Deaf users. *Extended Abstracts Proc. CHI* (2011), 2185–2190.

[8] McCarthy, J., Sasse, M.A. and Miras, D. 2004. Sharp or Smooth? Comparing the effects of quantization vs. frame rate for streamed video. *Proc. CHI*. (2004).

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

[10] 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).

[11] 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).

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

[13] Tran, J.J., Johnshon, T., Kim, J., Rodriguez, R., Yin, S., Riskin, E., Ladner, R. and Wobbrock, J.O. 2010. A web-based user survey for evaluating power saving strategies for Deaf users of MobileASL. *Proc. ASSETS* (2010), 115–122.

[14] Tran, J.J., Rodriguez, R., Riskin, E. and Wobbrock, J.O. 2013. A web-based intelligibility evaluation of sign language videotransmitted at low frame rates and bitrates. *Proc. ASSETS* (2013). *To appear*.