A Web-Based Intelligibility Evaluation of Sign Language Video Transmitted at Low Frame Rates and Bitrates

Jessica J. Tran\textsuperscript{1}, Rafael Rodriguez\textsuperscript{1}, Eve A. Riskin\textsuperscript{1}, Jacob O. Wobbrock\textsuperscript{2}

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

\textsuperscript{2}The Information School
DUB Group
University of Washington
Seattle, WA 98195 USA
wobbrock@uw.edu

ABSTRACT
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, we evaluated 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 bitrates (15, 30, 60, and 120 kilobits per second [kbps]). We discovered an “intelligibility ceiling effect,” where increasing the frame rate above 10 fps decreased perceived intelligibility, and increasing the bitrate above 60 kbps produced diminishing returns. Additional findings suggest 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. As part of this work, we developed the Human Signal Intelligibility Model, a new conceptual model useful for informing evaluations of video intelligibility.

Categories and Subject Descriptors
K.4.2. [Social Issues]: Assistive technologies for persons with disabilities; H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems – Video.

General Terms
Performance, Experimentation, Human Factors.

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

1. INTRODUCTION
Real-time mobile video communication allows deaf and hard-of-hearing people to communicate in their native 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. U.S. cellular networks do not provide unlimited data plans and may throttle network speeds to high data rate consumers [34]. The high video transmission rates implemented by commercial mobile video applications place a heavy load on the total available network bandwidth. They cause network congestion and delayed video, often interrupting mobile sign language video conversations. Currently, the Telecommunication Standardization Sector (ITU-T) Q.26/16 recommends sign language video to be transmitted at 25 frames per second (fps) at 100 kbps or higher, displayed at 352×288 pixels [21]. Our research demonstrates that intelligible sign language video communication can result at frame rates and bitrates less than recommended by the ITU-T.

In our evaluation of mobile sign language video intelligibility, we discovered a lack of uniformity in the way that signal intelligibility and signal comprehension are operationalized in human-centered evaluations. We introduce a new model, the Human Signal Intelligibility Model (HSIM), to distinguish the components comprising video intelligibility from the components comprising objective video quality and video comprehension. Intelligibility is defined as the capability of a signal to be understood, given that the signal was clearly articulated, captured, transmitted, received, and perceived by the receiver, including the environmental conditions affecting these steps. Comprehension is defined as signal intelligibility plus the 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 limit of intelligible mobile sign language video. Considering the bandwidth needed to support conversations or facilitating real-time mobile sign language learning has been extensively researched [5,11,12,23]. However, unlike the present work, none of this prior work has been intended for facilitating real-time mobile sign language conversations because linguistic accuracy is most important. Therefore, the effect of frame rate reduction on sign language sentences was examined in our study. Respondents rated video quality on a slider ranging from 0 (worst) to 100 (best). These researchers found that respondents favored video shown at 15 fps over 10 fps when shown at a fixed bitrate. Cavender et al. [4] used ASL video clips and also discovered a frame rate preference for viewing ASL video at a fixed bitrate. They evaluated intelligibility of ASL video displayed at two frame rates (10 and 15 fps), three bitrates (15, 20, and 25 kbps), and three region-of-interest (ROI) encoding levels (0, -6, and -12 ROI). (The ROI was an approximation of where the signers face and hands were located.) Our work investigated more frame and bitrate combinations than Cavender et al., and our findings corroborate their findings that respondents rated higher intelligibility for video viewed at 10 fps over 15 fps at a constant low bitrate, which is opposite of what Masry and Hemami found.

The findings from our work and elsewhere [11,12,23] suggests a threshold where increasing the frame rate does not significantly improve video intelligibility. Our research builds upon Cavender et al.’s [4] findings and more rigorously investigates intelligibility of sign language video. Cavender et al.’s laboratory study used prerecorded video filmed with a stationary video camera, which allowed more space in the signing region. By contrast, the videos evaluated in our web study were representative of the angle and signing space constrained by mobile devices. Also, our research goal was to discover how much video quality could be reduced before sign language intelligibility was compromised, a goal not approached by Cavender et al.’s work.

2. RELATED WORK

The effects of frame rate and bitrate 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. Our work fills this gap by identifying the lower limits of intelligible mobile sign language video.

2.1 Sign Language Learning & Comprehension

Sign language learning is more nuanced than holding sign language conversations because linguistic accuracy is most important. Therefore, the effect of frame rate reduction on sign language learning has been extensively researched [5,11,12,23]. Johnson and Caird [12] 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. In our work, gesture recognition is not enough to support meaningful conversations; therefore, we investigate the impact of low frame rates and low bitrates on sign language sentences.

Hooper et al. [11] defines comprehension as the ability for respondents to accurately retell stories verbatim. They investigated the impact on ASL learning comprehension when ASL video was presented at 6, 12, and 18 fps and displayed at 320×240, 320×240, and 480×360 pixels at 700 kbps. Hooper et al. found 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 not a significant difference in performance between 18 fps vs. 12 fps.

Sperling et al. [23] defines intelligibility as the ability to correctly recognize signs. 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. Our work goes beyond sign recognition and investigates video intelligibility to support two-way conversations.

2.2 Subjective Video Quality

We aim to discover whether frame rate or bitrate has more impact on ASL video intelligibility. A subjective experiment, conducted by Yadavalli et al. [32], 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 bitrates (100, 200, and 300 kbps). Viewers preferred video at 15 fps across all bitrates and video sequences, which suggest 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. Like Yadavalli et al.’s work, we aim to determine whether ASL video becomes more intelligible by increasing the frame rate once frame quality (determined by bitrate) is adequate. But unlike this prior work, we require respondents to actively watch and understand ASL video content.

Masry and Hemami [15] evaluated subjective video quality perception of non-ASL streaming video content transmitted at 10, 15, and 30 fps and six bitrates (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). These researchers found that respondents favored video shown at 15 fps over 10 fps when shown at a fixed bitrate. Cavender et al. [4] used ASL video clips and also discovered a frame rate preference for viewing ASL video at a fixed bitrate. They evaluated intelligibility of ASL video displayed at two frame rates (10 and 15 fps), three bitrates (15, 20, and 25 kbps), and three region-of-interest (ROI) encoding levels (0, -6, and -12 ROI). (The ROI was an approximation of where the signers face and hands were located.) Our work investigated more frame rates and bitrates than Cavender et al., and our findings corroborate their finding that respondents rated higher intelligibility for video viewed at 10 fps over 15 fps at a constant low bitrate, which is opposite of what Masry and Hemami found.

The findings from our work and elsewhere [11,12,23] suggests a threshold where increasing the frame rate does not significantly improve video intelligibility. Our research builds upon Cavender et al.’s [4] findings and more rigorously investigates intelligibility of sign language video. Cavender et al.’s laboratory study used prerecorded video filmed with a stationary video camera, which allowed more space in the signing region. By contrast, the videos evaluated in our web study were representative of the angle and signing space constrained by mobile devices. Also, our research goal was to discover how much video quality could be reduced before sign language intelligibility was compromised, a goal not approached by Cavender et al.’s work.
2.3 Objective Video Quality Measures

Measuring subjective video quality is time consuming, content-specific, and requires many subjects to produce generalizable findings. By contrast, peak signal-to-noise ratio (PSNR) is commonly used in video compression to measure objective video quality after lossy compression [29]. However, PSNR has been shown to not always accurately represent humans’ subjective judgments about video quality [7,17,24,26,28]. Numerous researchers have attempted to map PSNR to subjective responses by creating new objective video quality perception metrics [19,27,30,33]; however, these objective measures have been content-dependent.

Content intelligibility is most important for sign language video; therefore, objective video evaluations are not the most appropriate way to characterize video quality. Ciaramello and Hemami [6] recognized that sign language video needs to be evaluated in terms of subjective intelligibility. They created a computational model of intelligibility 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. By contrast, our model of subjective intelligibility for sign language video goes beyond objectively measuring video quality and details the components impacting subjective sign language intelligibility.

2.4 Defining Intelligibility

Often, intelligibility and comprehension are loosely defined and used interchangeably in evaluations of video quality. Some researchers focused on measuring signal intelligibility with the intent that if one finds the signal intelligible, then comprehension of content follows [1,8,9,11,18]. In his famous work, Shannon [22] 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. However, the channel model only focuses on the communication channel itself without considering the surrounding environment or properties of a human sender and receiver. Existing communication models [2,3] attempting to distinguish intelligibility from comprehension are poorly defined. Berlo [3] created the source, message, channel, receiver (SMCR) model of communication consisting of twenty different elements; however, it does not clearly identify which elements produce intelligible communication. Barnlund [2] proposed a transactional model of communication suggesting individuals are simultaneously engaging in the sending and receiving of messages. Although Barnlund’s model represents how information is transferred, it does not attempt to distinguish intelligibility from comprehension.

We believe that signal intelligibility and signal comprehension need to be distinguished. 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 intelligibility and the human receiver having the prerequisite knowledge to understand the information. These insights lead us to propose our own intelligibility model, described next.

3. Human Signal Intelligibility Model

We present the Human Signal Intelligibility Model (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 we argue are three usefuly distinct and separable things.

The HSIM (1) extends Shannon’s theory of communication [22] 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 entail comprehension by a receiver lacking the requisite knowledge for understanding.

We claim that 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 2 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, namely 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 2. 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. Our definition of signal intelligibility and signal comprehension builds upon Koul’s definition of speech signal quality. Koul [13]

---

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.
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.

Our 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 block diagram.

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 in which information is conveyed influences the
receiver’s ability to receive the content. For example, a fluent
ASL signer could have a motor impairment that would limit their
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.

4. WEB SURVEY DESIGN

The HSIM informs the design of our web study evaluating how
much frame rate and bitrate can be reduced before intelligibility is
compromised in mobile sign language video. Owing to the need to
ensure all receivers’ minds are fully capable of comprehension,
we screen participants for ASL fluency. Thereafter, we can
attribute differences in comprehension to differences in
intelligibility and not knowledge.

Our web study evaluated sign language video intelligibility
transmitted at four low frame rates (1, 5, 10, and 15 fps) and four
low bitrates (15, 50, 60, and 120 kbps) in a full factorial design.
The web study was selected over a laboratory study because
parameter settings could be evaluated with participants from
across the nation. A mobile web survey was considered, but at the
time of web development, we found too much variability across
mobile devices and mobile web browsers, which we could not
control as an environmental influence.

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. 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.

The web survey began by asking participants to self-report their
fluency in ASL. ASL interpretations of the English text
instructions were shown side-by-side throughout the web survey
to increase accessibility. A professional ASL interpreter, who is a
child of deaf adults, was consulted before filming.

4.1 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. 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. Recording video
from a smartphone was not an option due to added video
compression. We used an Acer Iconic tablet running Android
Honeycomb 3.2.1 to simulate the allowable signing space and
display angle. A male native ASL signer/consultant signed 16
short ASL sentences that included various amounts of finger
spelling and descriptive lexicons. The ASL signer was asked to
sign slowly, and to sign all signs within the allowable signing
space. The ASL signer sat in front of a solid dark blue
background. Video length ranged from 15-30 seconds. The
original YUV videos were encoded using the open source H.264
encoder [20]. 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 [35] since no additional artifacts
were contributed by this player.

4.2 Survey Components

All videos were displayed at 320×240 pixels in the middle of the
computer screen. A picture of the Sprint EVO phone was placed
behind each video to simulate the mobile video appearance. 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 3 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 4 is an example of a trivial comprehension
question pertaining to the video shown. A four point multiple
choice answer appeared with a corresponding image.

We unobtrusively logged the time it took to answer the
comprehension question to compare if there was any relationship
between the time to answer the comprehension questions and
Video 1 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

Figure 3: Example of question 1 shown in web survey.

Figure 4: Multiple choice comprehension question example.

5. RESULTS

Our web survey received 300 hits, with 99 respondents completing the survey, all of whom self-reported fluency in ASL. We eliminated results from those who responded with the same answers for all 16 videos, such as selecting all 1s or all 7s. We analyzed data 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 speakers, 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 own a smartphone and send text messages; 65 indicated they use video chat; and 53 use video relay services.

5.1 Perceived Intelligibility

Results will be reported in terms of intelligibility even though comprehension questions were asked. Recall that video intelligibility can be inferred from comprehension questions provided that the receivers' knowledge stores are fully adequate to understand the received signals—in this case, once ASL fluency is established. 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 [31] procedure that enables the use of ANOVA after alignment and ranking, while preserving interaction effects.

5.1.1 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 correction [10] were performed for 1 fps vs. 5 fps; 5 vs. 10 fps; 5 vs. 15 fps; and 10 fps vs. 15 fps. Table 1 and Figure 5 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 5 shows, videos displayed at 10 fps (averaged across all bitrate combinations) 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.

5.1.2 Bitrate Main Effect

Changing the bitrate 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 correction were performed for 15 kbps vs. 30 kbps; 30 kbps vs. 60 kbps; and 60 kbps vs. 120 kbps. Unsurprisingly, increasing the bitrate from 15 kbps to 30 kbps to 60 kbps to 120 kbps were 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 × Bitrate Interaction

There was also a significant frame rate × bitrate interaction (F(9,1139)=23.40, p<.0001). Upon closer inspection, videos transmitted at 10 fps, independent of bitrate, received the highest mean Likert scores for ease of understanding video quality as shown in Table 1 and Figure 5. 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 bitrate to transmit intelligible video. Videos displayed at 1 fps received the lowest mean Likert score, suggesting that 1 fps is too low to support intelligible sign language video.

5.2 Comprehension Questions

We unobtrusively logged the time participants responded to the comprehension questions. 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. We report findings on correctly answered comprehension questions across frame rates (averaged over all four bitrates) and across bitrates (averaged over all four frame rates). Table 2 lists the mean time and standard deviation for respondents who answered the comprehension question correctly.
Table 1: Mean Likert score responses for ease of understanding video quality. Note higher Likert scores correspond to higher perceived intelligibility.

<table>
<thead>
<tr>
<th>Frame rate (fps)</th>
<th>15</th>
<th>30</th>
<th>60</th>
<th>120</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bitrate (kbps)</td>
<td>Mean Likert</td>
<td>std. error</td>
<td>Mean Likert</td>
<td>std. error</td>
</tr>
<tr>
<td>1</td>
<td>2.14</td>
<td>0.14</td>
<td>1.13</td>
<td>0.07</td>
</tr>
<tr>
<td>5</td>
<td>3.01</td>
<td>0.16</td>
<td>4.43</td>
<td>0.15</td>
</tr>
<tr>
<td>10</td>
<td>4.04</td>
<td>0.16</td>
<td>4.74</td>
<td>0.13</td>
</tr>
<tr>
<td>15</td>
<td>3.51</td>
<td>0.17</td>
<td>3.97</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Figure 5: 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.

Table 2: 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 bitrates) and bitrate (averaged over all four frame rates). Bold values indicate highest mean Likert scores and fastest times to submit answer.

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

Figure 6: 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 bitrates) and bitrate (averaged over all four frame rates). Higher Likert scores correspond to higher perceived intelligibility.
We discovered that the fastest mean response times for correctly answering the comprehension questions for both frame rate (averaged over all four bitrates) and bitrate (averaged over all four frame rates) also received 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 bitrates) ($R=-0.66$); and mean response time and mean Likert scores for bitrate (averaged overall all four frame rates) ($R=-0.82$). These results suggest that higher perceived video intelligibility leads to faster content comprehension. Figure 6 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 bitrates) and bitrate (averaged over all four frame rates).

6. DISCUSSION

6.1 HSIM Influence on Study Design

The HSIM influenced our web study design and identified the components that were held constant. We allowed participants to self-report ASL fluency to encourage participation. The demographic questions had language fluency questions to infer levels of ASL fluency. Recall in Section 3, we made the distinction between signal intelligibility and signal comprehension where the latter is defined as signal intelligibility plus human knowledge and the receiver’s mind. Since data analysis was performed on data collected from fluent ASL respondents, we were not concerned with language proficiency influencing our results. We controlled the environment in which the video stimulus were recorded and how they were displayed on the web survey. The videos used in the survey were preprocessed to reduce the potential lag time when loading our web survey. We also asked participants to use a high speed internet connection and allow enough time to view all video sequences.

6.2 Study Findings

6.2.1 Frame Rate and Bitrate

We anticipated finding frame rate and bitrate 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 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 [4,16]. 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 [25] that people perceived the least amount of negative effects when the lowest frame rate and bitrate settings were applied.

We discovered diminishing returns for videos displayed at 60 kbps and 120 kbps independent of frame rate. Figure 5 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 ($F(1,1,139)=0.47, n.s.$). 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 may not be 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. [4] 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 perceived intelligibility. Our findings suggest that relaxing the recommended frame rate and bitrate 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 inverse correlation between mean Likert scores rating perceived video intelligibility and mean response times for correctly answering comprehension questions for both frame rate (averaged over all four bitrates) and bitrate (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. We recognize some videos may be easier to comprehend than others due to varied amounts of finger spelling and descriptive lexicons used. Nevertheless, we observed respondents answered comprehension questions more quickly when viewing ASL video with higher perceived intelligibility, suggesting that measuring response time may serve as a proxy for measuring video intelligibility, a relationship we aim to explore more rigorously in the future.

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. Our findings suggest that 5 fps would be sufficient for intelligible video communication. In future work, we will objectively measure how many signs are perceived by the viewer at 5 fps vs. 15 fps to understand the impact of signing speed and frame rate on video intelligibility.

7. CONCLUSION AND FUTURE WORK

We presented the Human Signal Intelligibility Model (HSIM) that identifies and distinguishes the components comprising signal intelligibility and signal comprehension. The HSIM informed our web study evaluating the lower limits of sign language video transmitted at four low frame rates and four low bitrates. We found that intelligibility was affected too negatively at 1 fps at 15 kbps, and that increasing resources beyond those required for 10 fps at 60 kbps provides negligible gains. Our 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.

In future work, we will conduct a laboratory study to evaluate and further demonstrate that intelligible real-time mobile video calls can be made at lower frame rates and bitrates than those recommended by the ITU-T standard. We anticipate the knowledge gained on low video quality intelligibility will make mobile sign language video more accessible and affordable. Finally, we anticipate the HSIM can be used in other signal evaluations of intelligibility and comprehension such as audio and
other video streaming media. The knowledge gained about intelligibility of low video quality has the potential to positively influence the user experience of mobile video communication.

8. ACKNOWLEDGMENTS
Thanks to Richard E. Ladner, Gerardo Di Pietro, Jason Smith, John Norberg, Christine Liao, Kyle Rector, Nick Parker, Kristy Winter, and Tobias Collins for the web study and ASL video feedback; Alex Jansen and Charles Delahunt for help with statistical analyses; Raja Kushalnagar, Lydia Runnels, Jerry Schnep, Rosalee Wolfe, Christian Vogler, Kathleen Youell, Sorenson VRS, and ZVRS for promoting the web survey; and our respondents. This work was funded in part by Google.

9. REFERENCES