Evaluating Intelligibility and Battery Drain of Mobile Sign Language Video Transmitted at Low Frame Rates and Bit Rates

JESSICA J. TRAN, Thomson Reuters
EVE A. RISKIN, RICHARD E. LADNER, and JACOB O. WOBBROCK,
University of Washington

Mobile sign language video conversations can become unintelligible if high video transmission rates cause network congestion and delayed video. In an effort to understand the perceived lower limits of intelligible sign language video intended for mobile communication, we evaluated 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]) at a constant spatial resolution of 320 × 240 pixels. We discovered an "intelligibility ceiling effect," in which increasing the frame rate above 10fps did not improve perceived intelligibility, and increasing the bit rate above 60kbps produced diminishing returns. Given the study parameters, our findings suggest that relaxing the recommended frame rate and bit rate to 10fps at 60kbps will provide intelligible video conversations while reducing total bandwidth consumption to 25% of the ITU-T standard (at least 25fps and 100kbps). As part of this work, we developed the *Human Signal Intelligibility Model*, a new conceptual model useful for informing evaluations of video intelligibility and our methodology for creating linguistically accessible web surveys for deaf people. We also conducted a battery-savings experiment quantifying battery drain when sign language video is transmitted at the lower frame rates and bit rates. Results confirmed that increasing the transmission rates monotonically decreased the battery life.

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

Additional Key Words and Phrases: Intelligibility, comprehension, American Sign Language, bit rate, frame rate, video compression, web survey, communication model, Deaf community, battery power, mobile phone, smartphone

ACM Reference Format:
Jessica J. Tran, Eve A. Riskin, Richard E. Ladner, and Jacob O. Wobbrock. 2015. Evaluating intelligibility and battery drain of mobile sign language video transmitted at low frame rates and bit rates. ACM Trans. Access. Comput. 7, 3, Article 11 (November 2015), 26 pages.
DOI: http://dx.doi.org/10.1145/2797142

This work was funded in part by Google and by the National Science Foundation under grant IIS-0811884. Any opinions, findings, conclusions, or recommendations expressed in this work are those of the authors and do not necessarily reflect those of any funder or supporter.

Authors' addresses: J. J. Tran, Thomson Reuters, 3 Times Square, New York, NY 10036; email: jjtran@uw.edu; E. A. Riskin, Electrical Engineering Department, University of Washington, Seattle, WA 98195; email: riskin@uw.edu; R. E. Ladner, Computer Science and Engineering Department, University of Washington, Seattle, WA 98195; email: ladner@cs.washington.edu; J. O. Wobbrock, The Information School, University of Washington, Seattle, WA 98195; email: wobbrock@uw.edu.

DOI: http://dx.doi.org/10.1145/2797142

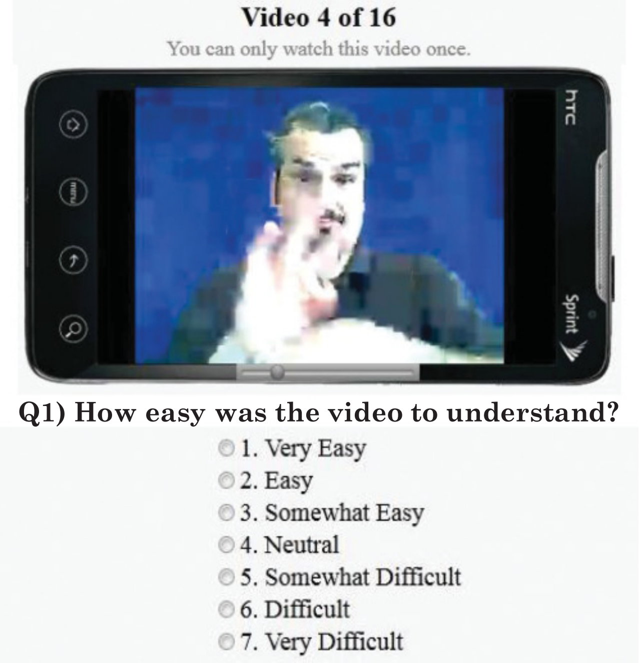
1. INTRODUCTION

With over 1.9 billion smartphone users at the end of 2013, smartphones are rapidly changing the way people communicate and receive information [Cisco 2015]. The growth of smartphone users has led to video being the fastest growing contributor to mobile data traffic [Cisco 2015]. Streaming video providers such as 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) such as Skype, FaceTime, and Google Hangouts. In 2010, Skype received 7 million downloads onto Apple's iPhone alone [Static Brain Research Institute 2012].

Often, high-fidelity video quality is a 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 [Hollington 2013]. The high data rate cost of using FaceTime over limited data plans can quickly become expensive [Chen 2013]. Other mobile video chat apps, such as Skype, transmit video at lower dynamic transmission rates, ranging from 40 to 450kbps depending on network traffic [Cicco et al. 2008]. Overall, commercial mobile video applications place a heavy load on the total available network bandwidth, which may lead to packet loss, delay, blurred video, and a poor user experience.

Deaf 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 [Saks and Hellström 2006]. However, total network bandwidth is limited and network congestion can lead to unintelligible content due to delayed and dropped video. Most US cellular networks no longer provide unlimited data plans and may throttle network speeds to high data rate consumers [Lawson 2011]. 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 prerecorded video and are not intended for real-time mobile video communication.

This article contributes to the continuing effort to make mobile sign language communication more accessible and affordable to deaf people. We optimize how much mobile sign language video transmission rates can be reduced to save bandwidth and battery life while maintaining video intelligibility for ASL video viewed on small mobile devices. This work includes the creation of the *Human Signal Intelligibility Model* (HSIM), a new conceptual model for understanding signal intelligibility and signal comprehension that aid in their operationalization. The HSIM influences our design and execution of a national web survey, as shown in Figure 1, evaluating the lower limits of intelligible sign language video intended to be viewed on small mobile devices. The web survey had 99 respondents watch 16 short ASL videos of a male native ASL signer signing short sentences shown at four low frame rates (1, 5, 10, and 15fps) and at four low fixed bit rates (15, 30, 60, and 120kbps) in a full factorial design. The spatial resolution was held constant at 320 × 240 pixels. Results revealed an intelligibility ceiling effect for video transmission rates, in which increasing the frame rate above 10fps and bit rate above 60kbps did not improve perceived video intelligibility. Notably, this is lower than the recommended ITU-T standards.



Video 4 of 16

You can only watch this video once.

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

Fig. 1. Screen shot of one video from a web survey evaluating intelligibility of sign language video displayed at 15fps at 30kbps at a spatial resolution of 320 × 240 pixels.

We also conducted a power-savings experiment to quantify how much bandwidth and battery life are consumed when transmitting sign language video at the investigated low frame rates and bit rates on an experimental smartphone application. As expected, increasing the frame rate monotonically decreased the battery life.

The main contributions of this work are summarized as follows: (1) the creation of 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; (2) empirical findings verifying an intelligibility ceiling effect for frame rate, in which increasing the frame rate above 10fps does not improve perceived video intelligibility when video is transmitted at a constant bit rate; (3) empirical findings verifying an intelligibility ceiling effect for bit rate, in which increasing the bit rate above 60kbps does not improve perceived video intelligibility; (4) empirical findings validating the bandwidth and power savings associated with reducing video frame rates and bit rates; and (5) demonstration that intelligible mobile sign language can occur at frame rates as low as 10fps and bit rates as low as 60kbps, which is lower than the current recommended ITU-T standards.

This article is an extended version of a paper originally presented at the ACM SIGACCESS Conference on Computers and Accessibility [Tran et al. 2013] and influenced the research presented at the ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '14) [Tran et al. 2014]. Related work is discussed in Section 2. The HSIM in Section 3 and web study in Section 5 were previously discussed in the ASSETS '13 paper. This article describes in more depth how the HSIM influenced the web study design. It also discusses our methodology for creating linguistically accessible web surveys for deaf people (Section 4). This article also describes a new study quantifying battery drain when transmitting video at lower frame rates in Section 7.

Study findings are discussed in Section 8. Conclusions and advice for future investigations of these phenomena are presented in Section 9.

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 is necessary. Video compression is the process of converting video files into a format that takes fewer bits. Compression takes two forms: lossless and lossy. Lossless compression compresses data without losing any information, at the expense of more required storage space. Lossy video compression uses spatial correlation and temporal motion compensation to reduce redundancy in video data, at the expense of introducing visible artifacts that may impact video intelligibility.

Spatial resolution, frame rate, and frame quality are the primary physical parameters that impact video quality. Spatial resolution is described by the number of pixels in each frame. A frame is a single video image and the rate at which frames are shown is measured in frames per second (fps). Frame quality is impacted by the quantization parameter (QP), which directly relates to the bit rate, as described in more detail, to follow.

H.264/MPEG-4 AVC is a standard for lossy video compression that is commonly used for recording, compressing, and decompressing video [Richardson 2004]. H.264 is best known as the codec standard for blu-ray discs and different streaming Internet sources such as YouTube, Vimeo, the iTunes store, and web software such as Adobe Flash Player and Microsoft Silverlight. The MobileASL application, discussed in Section 2.4, uses x264, which is an open-source version of H.264 [Aimar et al. 2005].

H.264 is a block-based, motion-compensation codec [Aimar et al. 2005]. Motion estimation is used to create motion vectors for intra- and interframe coding. Intraframe coding uses information contained only in the current frame (it performs no temporal processing) [Oppenheim and Schafer 1975]. Interframe coding takes advantage of temporal redundancy between neighboring frames, which allows for higher compression rates. The encoder divides each frame into blocks of pixels, called macroblocks. A block-matching algorithm tries to find a closely matching block in the previous decoded frame. If a matching block cannot be found, then that block is intracoded (I-block); otherwise, a two-dimensional motion vector, which provides an offset from the coordinates in the decoded picture to the coordinates of the reference picture, is formed [Oppenheim and Schafer 1975]. The difference between the new block and the previous one is transformed, via Discrete Cosine Transform (DCT) [Ahmen et al. 1974], and the resulting DCT coefficients are quantized. The motion vectors and DCT coefficients are losslessly compressed and sent to the decoder [Oppenheim and Schafer 1975].

The DCT transforms an image into different frequencies. The DCT has a strong "energy compaction" property in which the signal information is concentrated in a few low-frequency coefficients and the highest-frequency components are quantized to zero [Oppenheim and Schafer 1975]. Trade-offs in video quality can be made by varying the QP and frame rate. The DCT coefficients are divided by the QP. A low-QP value requires more bits to encode than a high-QP value, resulting in higher video quality. Conversely, a high-QP value results in the DCT coefficients being quantized more heavily, which requires fewer bits. In addition to the QP, the frame rate can be varied. For a fixed bit rate, there is a trade-off between frame quality (typically objectively measured by Peak Signal-to-Noise Ratio (PSNR) [Oppenheim and Schafer 1975], which is controlled by the QP parameter and frame rate. More frames per second means that the individual frames will be of lower quality to maintain a fixed bit rate.

2.2. Evolution of Mobile Networks

A new mobile generation wireless system is introduced in the United States approximately every 10 years. The first 1G system, Nordic Mobile Telephone, was introduced in 1981 and was the first fully automatic cellular phone system transmitting data at 1200bps. 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 14.4kbps. Next, 3G (EDGE and CDMA) started becoming available in 2002 and provided an upload data rate of 118.4kbps and download data rate of 296kbps. 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. Finally, 4G Long Term Evolution (LTE) began appearing in 2012 and provides download data peak rates of 300Mbps and upload peak rates of 75Mbps. The quality of service aims for a data-transfer latency of less than 5ms. Today, major cellular phone companies such as Sprint, T-Mobile, AT&T, and Verizon are expanding their 4G LTE networks to provide higher data speeds in more locations across the United States. However, consistent access to 4G LTE service is currently location-dependent.

Although 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 to 4GB per month depending on the data plan [AT&T 2014; T-Mobile 2014; Verizon 2014]. The average US consumer uses 733MB of data per month; however, those users generally check websites and email [Fitzgerald 2013]. 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 (160Kbps) requires 1.2MB per minute or 72MB per hour; music streaming at 320Kbps is equivalent to 2.4MB per minute or 144MB per hour; a Netflix video in standard definition can consume up to 0.7GB per hour and a Netflix video in HD can consume 1GB to 2.8GB per hour [Marshall 2014].

2.3. Commercial Mobile Video Applications

Commercial mobile video applications have evolved with the expanding networks. Skype is a free voice-over-IP (VoIP) service that allows people to communicate through instant message, voice, and video on computers and mobile devices [Skype 2011]. Skype transmits video at high bit rates with mobile-to-mobile calls at 500kbps and video calls between a mobile phone and a computer at 600kbps [Microsoft 2013]. Before 2013, Apple's FaceTime mobile video chat application could only work over Wi-Fi networks. Once Apple devices supported iOS6 (released in September 2012), FaceTime began working on AT&T's tiered data plans at the data consumption rate of 3MB of data per minute [Zeman 2010].

Video relay services (VRS) allow deaf people to communicate over video telephone with a hearing person in real time via a sign language interpreter. Major VRS companies such as Purple Communications, Inc. [Purple 2014], Sorenson VRS [Sorenson 2014], ConvoRelay [Convo 2011], and ZVRS [ZVRS 2014] provide VRS apps for mobile devices. In compliance with the ITU-T standard, these applications attempt to transmit video at rates of at least 25fps and 100kbps, which may lead to video delay or dropped video calls. VRS users tend to use video phones or computers with a broadband connection to utilize interpreting services without the worry of dropped video calls.

All of these 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

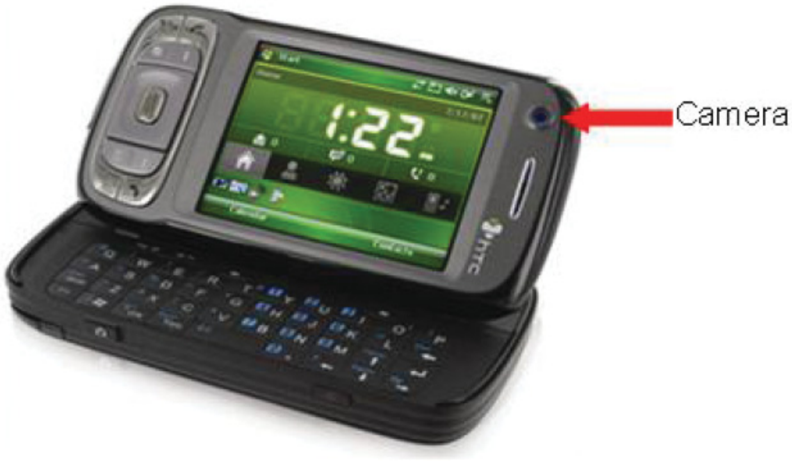


Fig. 2. HTC TyTNII cell phone.

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 people. Instead, network providers begin throttling down network speeds after 2GB of data usage per month [Lawson 2011]. This research contributes to the MobileASL project's [Riskin et al. 2012] goal of providing deaf people equal access to mobile video communication without needing to pay more for services.

2.4. The 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 [Riskin et al. 2012]. One of the goals was to transmit real-time, two-way video using the 3G GSM EDGE network, which has 296kbps download and 118kbps 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 (30kbps at 8–12fps) [Chon et al. 2009].

MobileASL was developed using the Windows Mobile 6.1 platform for the HTC TyTNII cellular phone [Chon 2011]. This phone, shown in Figure 2, was selected because it has a front-facing camera and screen, which can prop itself up at an angle during conversations. The phone weighs 6.7oz; has a 400MHz processor; and 1350mAH battery life. The MobileASL app uses the open-source x264 implementation of the H.264 standard [Aimar et al. 2005] with the ARMv6 SIMD instruction set [ARM 2008] and a NAT-enabled protocol [Chon 2011]. The app uses a peer-to-peer networking application that allows video transmission on both Wi-Fi and AT&T 3G/4G cellular networks.

Since intended users of MobileASL are deaf, 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 [Cherniavsky et al. 2009]. When MobileASL with ROI encoding encodes 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. [2006] conducted a focus group in 2006 investigating intelligibility of sign language video constrained by mobile phone technology. They explored the need and desire for mobile video phones and addressed potential challenges in 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 needed to have an easy and intuitive display; and the software needed the ability to make video calls between different types of video software. They also conducted a laboratory study evaluating video intelligibility at two frame rates (10 and 15fps), three bit rates (15, 20, and 25kbps), and three ROI encoding levels (0, -6, and -12 ROI), during which participants viewed prerecorded videos and were asked to subjectively rate perceived intelligibility. They discovered a frame rate preference of 10fps for viewing ASL video at a fixed bit rate of 25kbps.

Masry and Hemami [2003] evaluated subjective video quality perception of non-ASL streaming video content transmitted at 10, 15, and 30fps and six bit rates (40, 100, 200, 300, 600, and 800kbps). Respondents viewed fifteen 30s 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 15fps over 10fps when shown at a fixed bit rate of 800kbps.

The findings from our work and elsewhere [Holm 1979; Hooper et al. 2007] suggest that there is a threshold above which increasing the frame rate does not significantly improve video intelligibility. Our research builds on Cavender et al.'s [2006] 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 prior MobileASL research.

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. Our work fills this gap by identifying the lower limits of intelligible mobile sign language video.

2.5. Sign Language Comprehension

Sign language learning requires more 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 [Chen and Thropp 2007; Hooper et al. 2007; Johnson and Caird 1996; Sperling et al. 1985] but not so for holding sign language conversations. Johnson and Caird [1996] investigated whether perceptual ASL learning was affected by video transmitted at 1, 5, 15, and 30fps. 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 1fps and 5fps were sufficient for novice ASL learners to recognize learned ASL gestures. Although this work suggests frame rates as low as 1fps and 5fps can support sign language recognition, it does not evaluate conversational sign language, which the present research investigates.

Hooper et al. [2007] defined 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 18fps and displayed at 240 × 180, 320 × 240, and 480 × 360 pixels at 700kbps. They found video-display size did not affect comprehension, but varying frame rates did. Students performed better after viewing video at 12fps than at 6fps, and at 18fps than at 6fps. However, there was no significant difference in performance between 18fps and 12fps.

Sperling et al. [1985] define intelligibility as the ability to correctly recognize signs. Under this operationalization, they investigated ASL video intelligibility transmitted at 10, 15, and 30fps 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 15fps and 30fps did not have a noticeable difference in intelligibility, but lowering the frame rate to 10fps 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 were small and may have influenced respondents' ability to recognize signs shown at 10fps. Also, their work was conducted in 1985, when the video compression algorithms were not as efficient as today; therefore, more visual artifacts may have been introduced in the stimuli used. Our work goes beyond sign recognition and investigates video intelligibility to support two-way conversations.

2.6. Evaluating Video Quality

We aim to discover whether frame rate or bit rate has more impact on ASL video intelligibility. A subjective experiment, conducted by Yadavalli et al. [2003], evaluated frame-rate preferences passively viewed for low-, medium-, and high-motion sequences displayed at 352 × 240 pixels; three frame rates (10, 15, and 30fps); and three bit rates (100, 200, and 300kbps). A limitation of this work was the type of video content used for evaluation. Specifically, a boat moving across a body of water, camera panning from one side of a room to another, and a soccer match were used for low-, medium-, and high-motion video, respectively. Viewers preferred video at 15fps across all bit rates and video sequences, which suggests that 15fps represents a compromise rate between frame and motion quality. At 300kbps, respondents preferred video at 30fps, 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 bit rate) is adequate. But, unlike this prior work, we require respondents to actively watch and understand ASL video content.

Measuring subjective video quality is time-consuming, content-specific, and requires many subjects to produce generalizable findings. By contrast, PSNR is commonly used in video compression to measure *objective* video quality after lossy compression [Wiegang et al. 2003]. However, PSNR has been shown to not always accurately represent humans' subjective judgments about video quality [Feghali et al. 2007; Nemethova et al. 2006; Thu and Ghanbari 2008; Tran et al. 2011; Wang et al. 2002]. Numerous researchers have attempted to map PSNR to subjective responses by creating new objective video quality perception metrics [Winkler and Mohandas 2008; Feghali et al. 2007; Bae et al. 2009]; however, these objective measures have all 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 [2011] 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 yield statistically

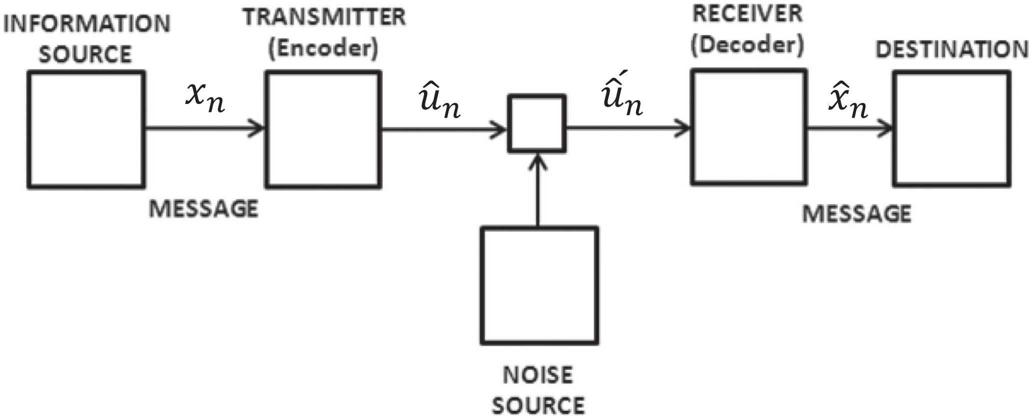


Fig. 3. Block diagram of Shannon's communication system [Shannon 1948, p. 10].

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 measuring video objectively and details the components impacting subjective sign language intelligibility.

3. HUMAN SIGNAL INTELLIGIBILITY MODEL (HSIM)

In evaluating mobile sign language video intelligibility, we discovered a lack of uniformity in the way that "signal intelligibility" and "signal comprehension" were operationalized in the literature of 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 automatically follows [Arons 1997; Harrigan 1995; Heiman and Tweney 1981; Hooper et al. 2007; Omoigui et al. 1999]. As part of this research, we present the *Human Signal Intelligibility Model* (HSIM), a new conceptual model informing video intelligibility evaluations and disentangling video intelligibility from video comprehension.

3.1. Existing Communication Models

Before introducing the components comprising the HSIM, we first discuss three extant conceptual models used to explain the human communication process: Shannon's Theory of Communication [Shannon 1948]; Berlo's Source-Message-Channel-Receiver model [Berlo 1960]; and Barnlund's transactional model of communication [Barnlund 1970]. Shannon's Theory of Communication originates from information theory, while Berlo's and Barnlund's models of communication originate from communication theory. This section will also address the limitations of existing communication models and how intelligibility is defined, which led to our creation of the HSIM.

3.1.1 Shannon's Theory of Communication. In his famous work, Shannon [1948] 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 3).

In this block diagram, the information source generates a signal, x_n, which is a lossy compressed, generating signal $\hat{u}_n$. Noise is introduced to the compressed signal during

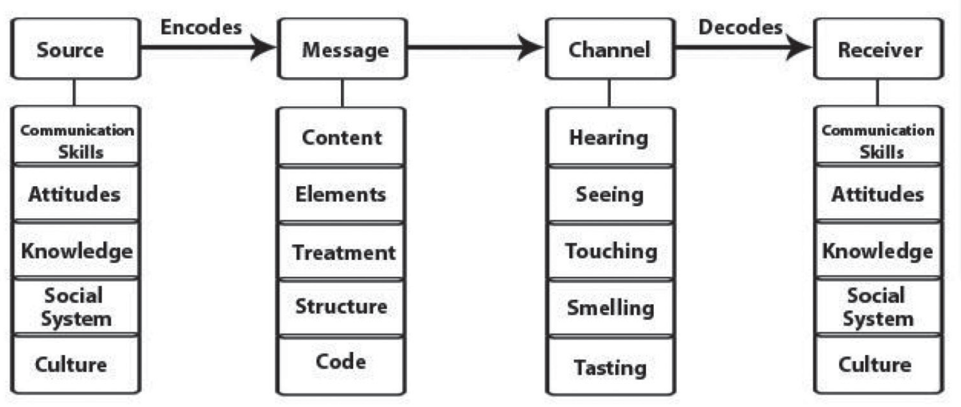


Fig. 4. Berlo's SMCR Model of Communication [Berlo 1960, p. 3].

transmission due to packet loss and network congestion, resulting in signal $\hat{u}_n$. Finally, the transmitted signal is decoded, producing signal $\hat{x}_n$. From this model, one could argue that objective metrics could be used to measure video quality, with high-quality scores implying intelligible content. An objective measure of quality might just measure the difference between x_n and $\hat{x}_n$. However, we argue that there are more components to intelligibility and comprehensibility of a video signal and using objective measures alone is not sufficient for human-centered evaluations. The environment in which video is recorded and displayed, as well as the humans sending and receiving video, also need to be considered. Shannon's channel model focuses only on the communication channel itself without considering the surrounding environment or properties of human senders and receivers.

3.1.2. Berlos's SMCR Model of Communication. Existing communication models [Berlo 1960; Barnlund 1970] that attempt 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 [Berlo 1960]. He created the source, message, channel, receiver (SMCR) model of communication, as shown in Figure 4, to represent an exchange of ideas that may hold influence and authority with one's culture.

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 one's 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 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 describe human communication with 20 different components, the SMCR model does not clearly identify which elements produce intelligible communication.

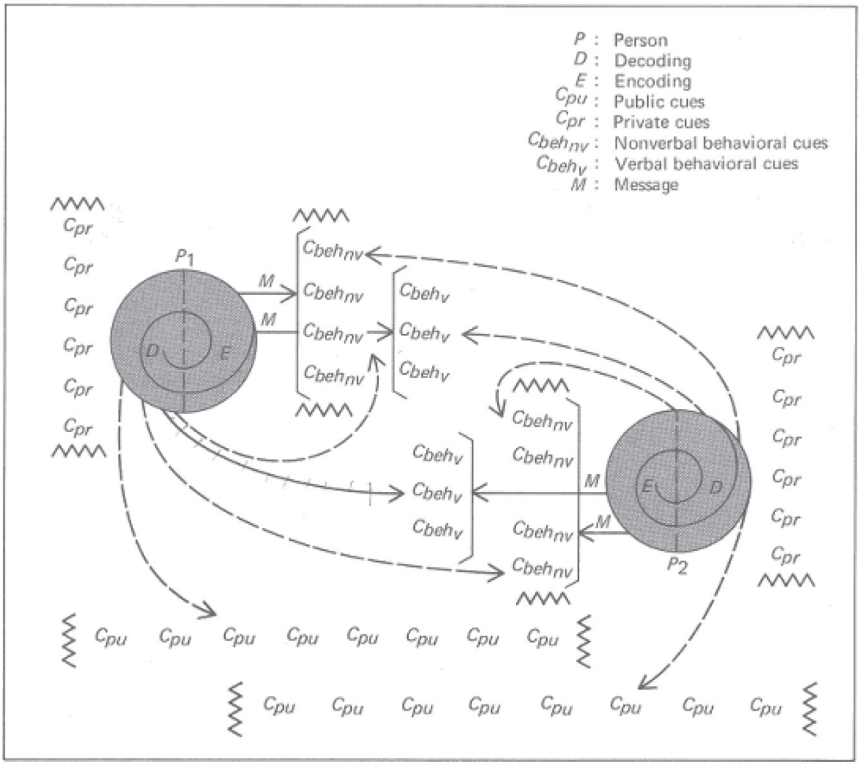


Fig. 5. Barnlund's Transactional Model of Communication [Barnlund 1970, p. 47].

3.1.3. Barnlund's Transactional Model of Communication. Barnlund [1970] proposed a Transactional Model of Communication with seven communication postulates, suggesting that individuals are simultaneously engaging in the sending and receiving of messages, as shown in Figure 5.

The Transactional Model of Communication states that the giving and receiving of messages is reciprocal, not one-way; 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 one another. There are seven communication postulates [Barnlund 1970]: (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 shared meaning for any message to be successfully communicated [Clark 1985]. 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.

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 [Merriam-Webster 2003]; namely, how well the signal was articulated, captured, transmitted, received, and perceived by the receiver, including the environmental conditions affecting these steps. Comprehension, on the other hand, relies on signal intelligibility *and* the human receiver having the prerequisite knowledge, including knowledge of context, to understand the information. For a signal to be comprehended, it must be at least minimally intelligible, but not all intelligible signals must or will be comprehended. Both intelligibility and comprehension are human-centered concepts, unlike objective video-quality measures such as the PSNR, which is a technology-centered concept. These distinctions lead to the creation of the HSIM, described next.

3.3. HSIM Components

We 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 concepts.

The HSIM (1) extends Shannon's theory of communication [Shannon 1948] 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 require comprehension to have been intelligible. If the receiver lacks the requisite knowledge for understanding, the signal will not be comprehended.

The *capability* of a signal (e.g., video) to be comprehended is different than whether a signal is *actually* comprehended in any given instance; 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. Figure 6 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: whether the knowledge of the human receiver is adequate to understand the communicated message. Because of this, the receiver's mind is included in the components comprising comprehension in Figure 6. The knowledge of the human sender, on the other hand, is irrelevant to comprehension by the receiver. For example, the sender could be a robot articulating ASL signs without any real understanding of ASL. The HSIM's definition of signal intelligibility and signal comprehension builds on Koul's definition of speech signal quality. Koul [2003] 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.

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 could 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.

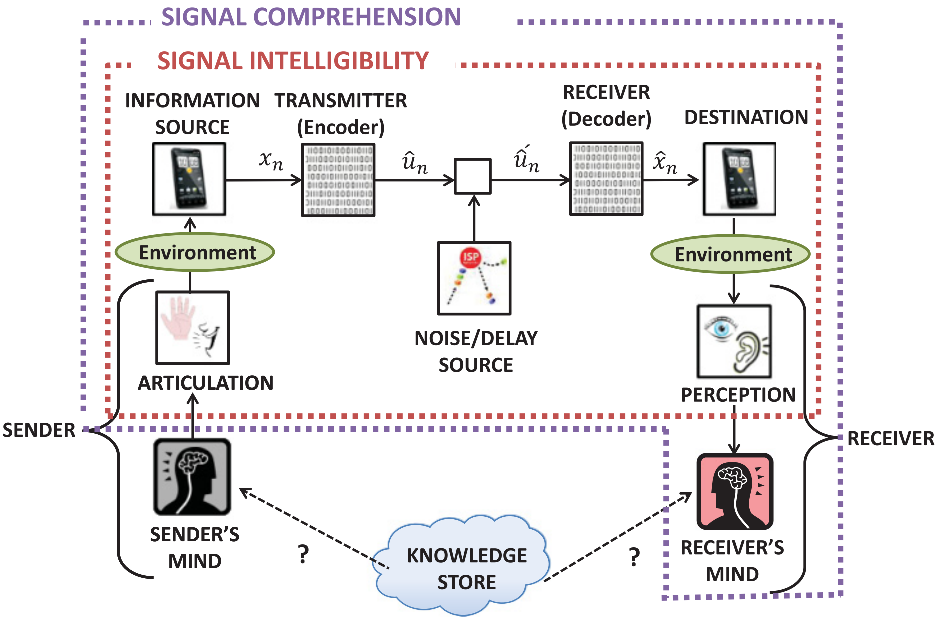


Fig. 6. Block diagram of the Human Signal Intelligibility Model. Note that the components comprising signal intelligibility are a subset of signal comprehension, which is signal intelligibility plus the receiver's mind.

The HSIM 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 the 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 the ability to process information. For instance, the sender could sign perfectly clear ASL, but if the receiver were blind, 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 its 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 and contextual knowledge to understand the signals that they receive. Such minds remove the 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 [Weber 1834] would not be appropriate because differences in video quality will be more evident at lower transmission rates before a signal becomes unintelligible.

The HSIM informs our web study design, which evaluates 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 that all receivers' minds are fully capable of comprehension, participants were screened for ASL fluency. Thereafter, differences in comprehension could be attributed to differences in intelligibility and not knowledge.

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 [Cumming and Rodda 1989; Munoz-Baell and Ruiz 2000]. Under this paradigm, deafness is perceived as the dominant quality of a group of people who share a "condition."

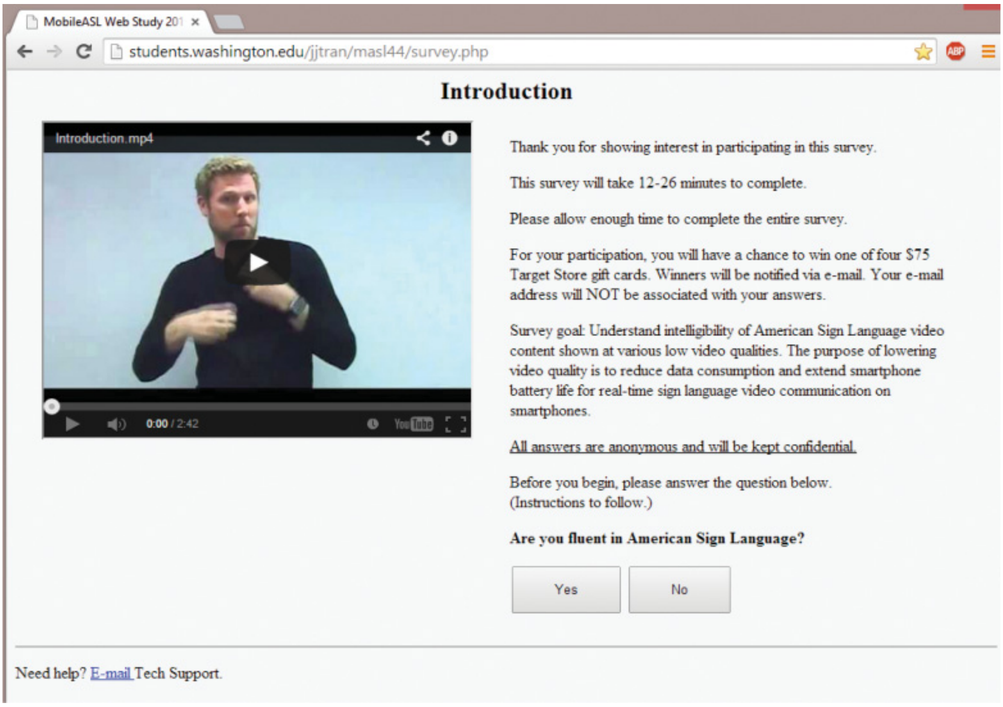
The social model, in contrast, holds that Deaf people are disabled more by their interactions with hearing people than by the physical condition that determines their perception of sounds. This view recognizes the linguistic [Lucas and Valli 2000; Maher 1996] and sociological [Padden and Humphries 2005; Reagan 1995] 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 [Lane 1992], 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, we identify two issues requiring explicit attention: (1) linguistic accessibility and (2) 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 study: (1) the intended audience of Deaf signers; (2) linguistic research that states that the grammar and lexicon of ASL are distinct from that of English [Lucas and Valli 2000; Padden and Humphries 2005]; and (3) the value Deaf Culture places on both linguistic accessibility and self-determination [Lucas and Valli 2000]. For this web study, we include an alternative to textual English by incorporating *ASL instructional videos*, to both increase accessibility and demonstrate our 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 but who are not fluent in ASL (for example, late-deafened individuals.) Figure 7 is an example of the ASL instructional video used alongside the English text.

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,



Introduction

Thank you for showing interest in participating in this survey.

This survey will take 12-26 minutes to complete.

Please allow enough time to complete the entire survey.

For your participation, you will have a chance to win one of four $75 Target Store gift cards. Winners will be notified via e-mail. Your e-mail address will NOT be associated with your answers.

Survey goal: Understand intelligibility of American Sign Language video content shown at various low video qualities. The purpose of lowering video quality is to reduce data consumption and extend smartphone battery life for real-time sign language video communication on smartphones.

All answers are anonymous and will be kept confidential.

Before you begin, please answer the question below.
(Instructions to follow.)

Are you fluent in American Sign Language?

Yes No

Need help? E-mail Tech Support.

Fig. 7. Example of web survey with ASL interpretation of the English text.

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.

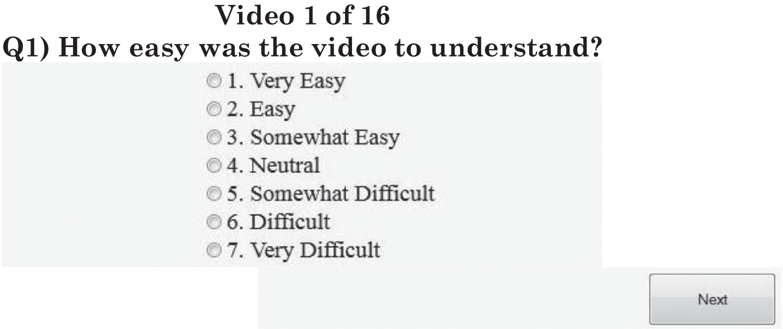
5. WEB SURVEY DESIGN

The HSIM informs the design of our web study, which evaluates sign language video intelligibility transmitted at four low frame rates (1, 5, 10, and 15fps) and four low bit rates (15, 30, 60, and 120kbps), in a full factorial design. These frame rates and bit rates are representative of what would be displayed on mobile devices.

The spatial resolution was held constant at 320 × 240 pixels. The web study was selected over a laboratory study because more parameter settings can be evaluated with participants from across the nation. 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.

5.1. Establishing Language Fluency

Using the HSIM requires establishing language fluency to ensure that all receivers' minds are fully capable of comprehension. Subsequently, we can attribute differences in comprehension to differences in intelligibility and not language fluency. We recruited participants from listservs with known fluent ASL signers. Our 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 deaf or hard-of-hearing?"; "Are you a native ASL signer?"; "From whom did you learn ASL?"; and "How many years have you signed ASL?" We provided instructions to the web study in both ASL and English. ASL



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

Fig. 8. Example of Question 1 shown in web survey.

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.

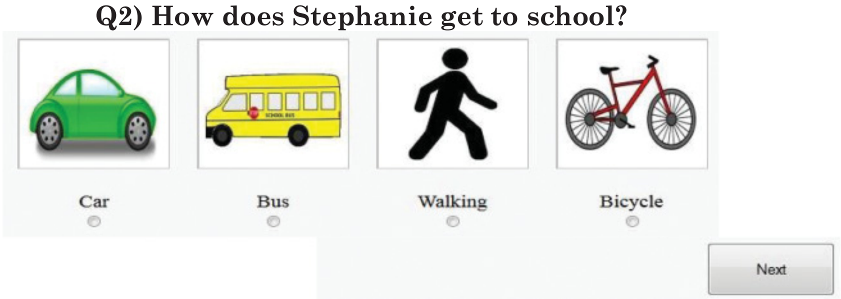
5.2. Video Stimuli

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 prerecorded video on a computer screen parallel to the participant, the videos used in the survey simulated the 45-degree angle and signing space that would be typically displayed on a small mobile device. At the time of video recording, the front-facing camera of smartphones, such as Sprint's EVO phone, recorded only compressed video in the 3GP file format. A tablet was selected to record the videos because it simulated the allowable signing space and display angle. Recording video from a smartphone was not an option due to added video compression. An Acer Iconic tablet running Android Honeycomb 3.2.1 was used to video-record a male, native ASL signer/consultant, signing 16 short ASL sentences that included various amounts of finger spelling and descriptive lexicons. The ASL signer was asked to sign slowly and articulate all signs within the allowable signing space. The ASL signer sat in front of a solid dark-blue background. Video length ranged from 15s to 30s. The tablet recorded uncompressed video in 4:2:2 YUV format at 25fps, 8.73Mbps with 320 × 240 screen resolution.

The original YUV videos were encoded using the open-source H.264 encoder [Richardson 2004] at 1, 5, 10, and 15fps at 15, 30, 60, and 120 kbps, respectively, in a full-factorial design. The encoded videos were converted to MPEG-4 using a publicly available converter [Kurtnoise 2009] that does not contribute additional compression artifacts. The web survey displayed the videos using Apple's QuickTime media player [Apple 2013] since it contributes no additional artifacts.

5.3. Survey Components

The survey consisted of three parts: Part 1, practice videos; Part 2, actual survey; and Part 3, demographic questions. Part 1 displayed two practice videos for participants to familiarize themselves with the survey layout. 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 environment in which the videos would be viewed. Each video was shown once, without the option to repeat or enlarge the video, then removed from the screen and replaced by two questions shown one at a time. Figure 8 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



Q2) How does Stephanie get to school?

Fig. 9. Multiple-choice comprehension question example.

very difficult. Figure 9 is an example of a trivial comprehension question pertaining to the video shown. A four-point, multiple-choice answer appeared with a corresponding image.

The same layout used in Part 1 was used in Part 2 of the survey, in which participants watched 16 different videos at each bit-rate and frame-rate combination. Videos were randomly displayed using a Latin Squares algorithm. The frame-rate and bit-rate settings did not change within each video clip.

Unobtrusive logging was implemented to measure the time it took to answer Questions 1 and 2. The start time began when the question appeared on the screen and the stop time occurred once the "Next" button was clicked. Unobtrusive logging also captured computer screen size, Internet browser, and computer operating system. Finally, the survey concluded with Part 3 asking demographic questions to establish language fluency, as described in Section 5.1, and questions to gather technology use, such as: "Do you own a smartphone or Blackberry?"; "Do you text message on the smart phone or Blackberry?"; "What operating system is on your smartphone?"; "Do you video chat?"; "What video chat program do you use?" "Do you use a video phone?"; "Do you use Video Relay Service (VRS)?"; and "Which VRS service(s) do you use?"

6. 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 to 72 years old (median = 40 years, $SD = 12.7$ years). Of the 77 respondents: 56 were deaf (38 indicated ASL as their native language, 11 have parents who are deaf), 54 indicated ASL as their daily language, and the number of years they have spoken ASL ranged from 5 to 59 years (median = 28 years, $SD = 12.7$). All but 7 respondents owned a smartphone and sent text messages; 65 indicated they used video chat; and 53 used video-relay services.

6.1. Perceived Intelligibility

Results are 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 is 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* procedure that enables the use of ANOVA after alignment and ranking, while preserving interaction effects [Higgins and Tashtoush 1994; Wobbrock et al. 2011].

Table I. Mean Likert Scale Responses for Ease of Understanding Video Quality

	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

Note: *Higher* Likert scores correspond to better comprehension.
Bold numbers indicate higher Likert scores where 7-strongly agree and 1-strongly disagree.

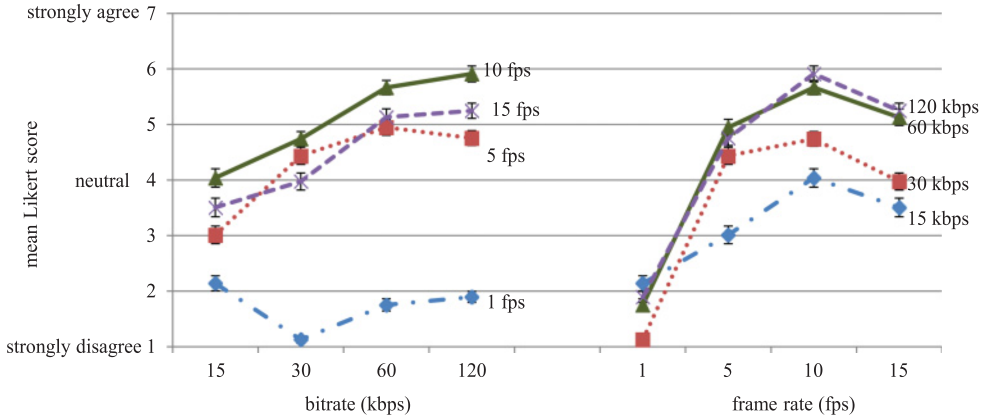


Fig. 10. Plot of 7-point Likert scale rating participants' ease of understanding the video for each frame rate and bit rate averaged over all participants. Error bars represent standard error.

6.1.1. 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 procedure [Holm 1979] were performed for 1fps versus 5fps; 5fps versus 10fps; 5fps versus 15fps; and 10fps versus 15fps. Table I and Figure 10 list the mean Likert score for question 1, in which higher scores correspond to higher agreement with the ease of perceived understanding of video content. As expected, videos displayed at 5fps when compared to 1fps received higher mean Likert scores for video intelligibility (F(1,1139) = 921.07, $p < .0001$). Videos displayed at 10fps when compared to 5fps received higher mean Likert scores for video intelligibility (F(1,1139) = 111.13, $p < .0001$). However, when comparing 10fps versus 15fps, video displayed at 10fps was found to have a higher mean Likert score for intelligible content (F(1,1139) = 77.22, $p < .0001$). As Figure 10 shows, video displayed at 10fps (averaged across four bit rates) received higher mean Likert scores than all other frame rates. An unexpected finding was that video was *not* perceived to be less intelligible at 5fps versus 15fps (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 25fps for intelligible sign language video.

6.1.2. Bit Rate Main Effect. As expected, changing the bit rate was found to have a significant main effect on the ease of understanding ASL video (F(3,1139) = 145.53, $p <$ .0001) as demonstrated in Figure 10. Post-hoc contrast tests with Holm's sequential Bonferroni procedure were performed between 15kbps versus 30kbps; 30kbps versus 60kbps; and 60kbps versus 120kbps. Unsurprisingly, increasing the bit rate from

Table II. Percentage of Correctly Answered Comprehension Questions Across Frame Rate and Bit Rate

	frame rate (fps)			
bit rate (kbps)	1	5	10	15
15	77.90	64.90	97.40	100.0
30	19.48	94.80	100.00	98.70
60	94.80	96.10	100.00	100.00
120	97.40	97.40	97.80	100.00

Note: Accuracy increases with bit rate but not with frame rate.

15kbps to 30kbps to 60kbps was found to improve the perceived ease of understanding ASL video (F(1,1139) = 82.75, $p < .0001$). However, comparing the perceived ease of understanding video displayed at 60kbps versus 120kbps was not found to be statistically significant (F(1,1139) = 4.62, *n.s.*).

6.1.3. Frame Rate × Bit Rate Interaction. There was also significant frame rate × bit rate interaction (F(9,1139) = 23.40, $p < .0001$). Videos transmitted at 10fps, independent of bit rate, received the highest mean Likert scores for ease of understanding video quality, as shown in Table I and Figure 10. This result was also found to be statistically significant when post-hoc contrast tests with Holm's sequential Bonferroni procedure was performed for video transmitted at 10fps versus 15fps, varying the frame rate while the bit rate was held constant (F(1,1139) = 77.22, $p < .0001$). Additionally, displaying the video at 60kbps versus 120kbps was not found to be statistically significant to improve video intelligibility (F(1,1139) = 4.62, *n.s.*), which is reflected by similar mean Likert scores. This suggests that 60kbps is high enough to transmit intelligible video. Video displayed at 1fps received the lowest mean Likert score, which suggests that 1fps is too low to support intelligible sign language video conversations.

6.2. Comprehension Questions

Table II lists the percentage of correctly answered comprehension questions across the frame rates and bit rates. A one-sample Chi-Square test of proportions was performed 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 expected since more bits are allocated to each frame and prior work has demonstrated that increasing the bit rate leads to higher perceived video quality [McCarthy et al. 2004; Nemethova et al. 2006; Wang and Ou 2012]. As Table II demonstrates, 13 of 16 videos received correctly answered comprehension questions with 95% accuracy or higher. This may suggest that the comprehension questions used were too easy; however, the main purpose of the comprehension questions was to ensure that participants were paying attention to the video content.

7. QUANTIFYING BATTERY DRAIN

Reducing the rates at which sign language video is transmitted is only half the solution to extending battery life and reducing bandwidth consumption. 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. Battery life will continue to be a limiting factor for prolonged mobile video communication.

We quantify the battery drain of sign language video transmitted at 320 × 240 spatial resolution at four low transmission rates similar to the ones investigated in the web study, specifically: 5fps/25kbps, 10fps/50kbps, 15fps/75kbps, and 30fps/150kbps. These settings are slightly different from the web study because of the technological limitations of implementation; however, the frame rate and bit rate pairs are still considerably lower than the ITU-T standard. Intuitively, one would expect that transmitting video at lower frame rates and bit rates would result in longer battery life. We conducted this study to quantify and confirm such hypotheses.

7.1. Experiment Setup

Our HSIM influenced the battery study design, specifically which technological components were held constant: the environment, the video content, the network over which transmission occurred, the mobile devices used, and video transmission rates. By doing this, we could attribute battery drain to the video transmission rates and not the technology setup. A Samsung Galaxy S3 smartphone was used to run an open-source video chat software app called IMSDroid [Doubango Telecom 2009], whose encoder was modified to transmit video at 5, 10, 15, and 30fps. IMSDroid is an open-source video conferencing application running on Doubango [Doubango Telecom 2009], a 3GPP IMS/LTE (IP Multimedia Subsystem) framework for embedded systems. IMSDroid is a Java-based frontend to Doubango, which is an 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 a different video encoder. Doubango is the backend framework running 3GPP IMS/LTE, which can run many different types of protocols such as SIP/SDP, HTTP/HTTPS, and DNS. In this experiment, Session Initiation Protocol (SIP) was selected for the VoIP.

To account for network bandwidth and to minimize network congestion, an Asterisk [Asterisk 2014] server was set up as the communication server for the battery study. It controlled the average bit rate per frame. Asterisk is an open-source framework that supports the server side of facilitating VoIP video communication, for which we used SIP [Rosenberg et al. 2002]. A specific configuration file was modified to regulate the bit rate at which video was transmitted, specifically an average of 5kb/frame. Asterisk uses User Datagram Protocol [Postel 1980], which is suitable for fast and efficient transmission of data for video conversations.

Since the bit rate averaged 5kb/frame, the bit rate increased as the frame rate increased: 25, 50, 75, and 150kbps, respectively. The spatial resolution of the video transmitted was 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, thus, the Galaxy S3 was selected. Network traces were conducted on the Asterisk server to monitor the frame rate and bit rate of each video call. A free smartphone diagnostic app, called AndroSensor [Asim 2013], was used to log the discharge of the battery in the experiment.

AndroSensor ran in the background of IMSDroid and logged the battery life percentage in 0.5s increments for 30min. In a preliminary experiment, it was discovered that transmitting video of a person signing consumes more battery life than transmitting a static image. Therefore, all experiments were conducted with the smartphone facing a computer monitor where a person was signing on the screen. Figure 11 is a picture of this experimental setup. 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.

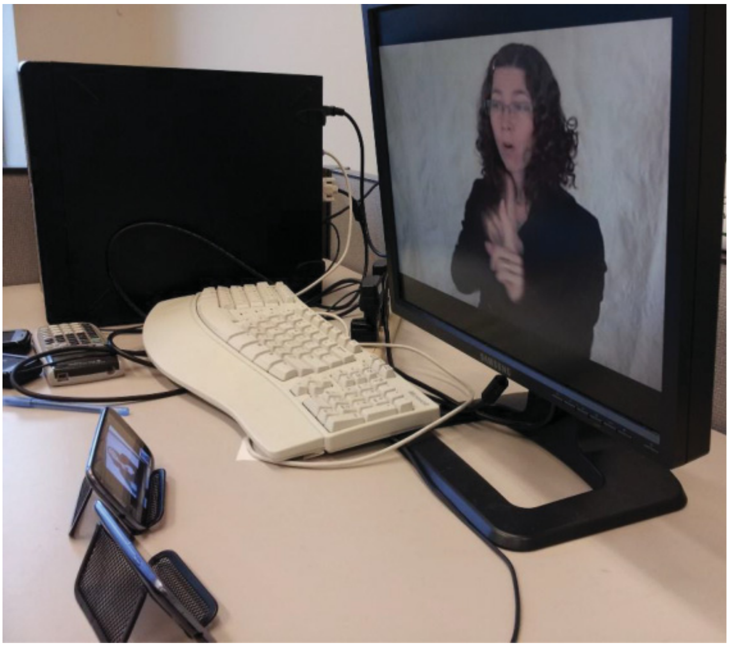


Fig. 11. Experimental setup in which two Samsung Galaxy S3 phones are facing a computer screen with a video of a woman signing in ASL.

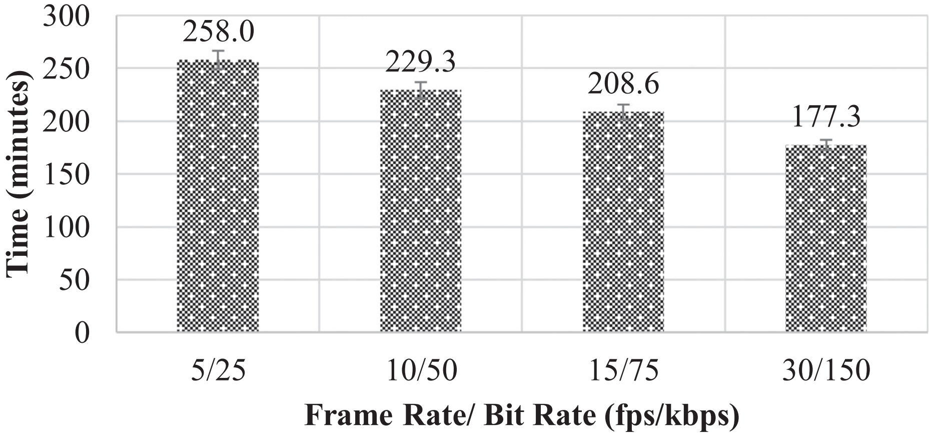


Fig. 12. Estimated average battery life (in minutes) for sign language video transmitted on IMSDroid at each frame rate/bit rate.

7.2. Results

As anticipated, increasing the frame rate at which sign language video was transmitted over the smartphones consumed battery life more quickly because more processing power was 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 would discharge to 0%. Figure 12 shows the extrapolated data for the average battery life of the Samsung Galaxy S3 for each frame rate.

From this experiment, it is estimated that the Samsung Galaxy S3 on standby, with IMSDroid turned off, has a battery life of 1000min and IMSDroid turned on and not transmitting video has a battery life of 750min. The Samsung Galaxy S3 specifications state that a fully charged battery could last up to 8 hours of talk time [T-Mobile 2014]. Our results demonstrate that transmitting video on mobile devices is computationally intensive and depletes a full battery charge in 3 to 4h. As expected, reducing the frame rate/bit rate of video monotonically increases battery life. This result is further corroborated in Tran et al.'s laboratory study, in which fluent ASL signers in pairs held real-time, free-form conversations over an experimental smartphone app transmitting video at the same frame rate and bit rate pairs [Tran et al. 2014].

8. DISCUSSION

8.1. HSIM Influence on Study Design

The HSIM influenced our web study and battery study designs in terms of which components were held constant. We allowed participants to self-report ASL fluency to encourage participation. Language fluency questions were asked in the demographic questions to infer levels of ASL fluency. Recall that, in Section 3, we made the distinction between signal intelligibility and signal comprehension: 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 was recorded and how it was 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.

8.2. Study Findings

8.2.1. Frame Rate and Bit Rate. We anticipated finding frame rate and bit rate pairs in which video quality either begins to affect intelligibility too negatively or diminishing returns begin. Unsurprisingly, respondents overwhelmingly ranked video displayed at 1fps to have the lowest mean Likert scores for ease of understanding the video content. The 1fps option was included 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 [Cavender et al. 2006; Masry and Hemami 2001].

We discovered diminishing returns for videos displayed at 120kbps over video at 60kbps, independent of frame rate. Figure 10 shows how the mean Likert scores for 60kbps and 120kbps, when averaged over all four frame rates, had similar Likert scores and were not found to be significantly different in terms of intelligibility ($F(1,1139) = 0.47$, *n.s.*). Our findings suggest that 60kbps is high enough to provide intelligible video conversations.

Another important finding was that video transmitted at 10fps received a higher mean Likert score than video transmitted at 15fps across all bit rates. The preference of viewing ASL video at 10fps over 15fps was also discovered in earlier ASL video communication research conducted by Cavender et al. [2006]. However, their findings reported a slight, but significant, main effect that frame rate influenced video intelligibility. Our results strongly affirm that ASL video intelligibility peaks at 10fps across all bit rates. At a fixed low bit rate, more bits are allocated per frame at 10fps versus 15fps. This difference is noticeable enough to result in higher intelligibility. Our findings suggest that relaxing the recommended frame rate and bit rate to 10fps at

60kbps will provide intelligible video conversations while reducing total bandwidth consumption to 25% of the ITU-T standard.

8.2.2. Signing Speed. The signing speed used in the video stimuli may have contributed to the nonsignificant intelligibility improvement of video transmitted at 5fps versus 15fps. Our findings suggest that 5fps would be sufficient for intelligible video communication. In future work, we will objectively measure how many signs are perceived by the viewer at 5fps versus 15fps to understand the impact of signing speed and frame rate on video intelligibility.

9. CONCLUSION AND FUTURE WORK

There will be a continued need for investigating trade-offs 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.

9.1.1. Context-Aware Video-Quality Adaptation. 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. 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 work, 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 over current mobile video communication.

9.1.2. Region-of-Interest Improvements. This web study focused on the baseline transmission rates at which to transmit video, without ROI-encoding. A future area of research would be to develop new algorithms that would track the ROIs most important to the signer, specifically the hands and face, and allocate more data to those ROIs.

9.1.3. Mobile Video Communication in Emergency Situations. Emergency response work can greatly benefit from the additional information provided with live video. Findings from this research 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.

9.1.4. Conclusion. We presented the Human Signal Intelligibility Model (HSIM), which 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 bit rates. We found that intelligibility was affected too negatively at 1fps, and that increasing the frame rate and bit rate above 10fps at 60kbps provided negligible gains. Our findings suggest that increasing video transmission rates above 10fps and 60kbps does not increase perceived intelligibility of content. This finding was further investigated by Tran et al. [2014] in a laboratory study. These study results further corroborate the findings that respondents can successfully hold intelligible real-time sign language conversations at transmission rates lower than the recommended ITU-T standard.

Finally, we anticipate that 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.

ACKNOWLEDGMENTS

Thanks to Rafael Rodriguez for the web study infrastructure; Ben Flowers for conducting the battery-savings experiment; Gerardo Di Pietro for the ASL video content; University of Washington's Deaf and Hard of Hearing Services Center; Tobias Cullins for coordinating the ASL interpreters; Samsung for the smartphones used in this study; Sorenson VRS and ZVRS for promoting the web survey; and our respondents.

REFERENCES

N. Ahmen, T. Natarajan, and K. R. Rao. 1974. Discrete cosine transform. *IEEE Transactions on Computers C-23*, 1, 90–93.

L. Aimar, L. Merritt, E. Petit, et al. 2005. x264 - a free h264/avc encoder. Online (last accessed on: 04/01/07). http://www.videolan.org/developers/x264.html.

Apple. 2013. Apple - QuickTime - Download. Retrieved September 30, 2015 from http://www.apple.com/quicktime/download/.

ARM. 2008. The architecture for the digital world. Retrieved September 30, 2015 from http://infocenter.arm.com/help/index.jsp?topic=/com.arm.doc.ddi0419c/index.html.

B. Arons. 1997. SpeechSkimmer: A system for interactively skimming recorded speech. *Proceedings of the CHI*, 3–38.

F. Asim. 2013. AndroSensor. Retrieved September 30, 2015 from http://www.fivasim.com/androsensor.html.

Asterisk. 2014. Asterisk. Retrieved September 30, 2015 from http://www.asterisk.org/.

AT&T. 2014. AT&T. Retrieved September 30, 2015 from http://www.att.com/shop/wireless/data-plans.html#fbid=027qt05YFJ6.

S. Bae, T. N. Pappas, and B. Juang. 2009. Spatial resolution and quantization noise tradeoffs for scalable image compression. *ICASSP*, IEEE, II–945–II–948.

D. Barnlund. 1970. *A Transactional Model of Communication*. Harper & Row. New York, NY.

D. K. Berlo. 1960. *The Process of Communication*. Holt, Rinehart, & Winston, New York, NY.

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

B. Chen. 2013. AT&T allows FaceTime for limited data users. What about unlimited? *The New York Times*. Retrieved September 30, 2015 from http://bits.blogs.nytimes.com/2013/01/16/facetime-limited-data-att/?_php=true&_type=blogs&_r=0.

J. Y. C. Chen and J. E. Thropp. 2007. Review of low frame rate effects on human performance. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans* 37, 6, 1063–1076.

N. Cherniavsky, J. Chon, J. O. Wobbrock, R. Ladner, and E. Riskin. 2009. Activity analysis enabling real-time video communication on mobile phones for deaf users. *UIST*.

J. Chon. 2011. Real-time sign language video communication over cell phones. Ph.D. thesis. University of Washington. 1–105.

J. Chon, N. Cherniavsky, E. Riskin, and R. Ladner. 2009. Enabling access through real-time sign language communication over cell phones. *Asilomar Conference on Signals, Systems, and Computers*, 588–592.

F. Ciaramello and S. Hemami. 2011. A computational intelligibility model for assessment and compression of American Sign Language video. *IEEE Trans. IP*. 20, 11.

L. Cicco, S. Mascolo, and V. Palmisano. 2008. Skype video responsiveness to bandwidth variations. *NOSSDAV*.

H. Clark. 1985. Language use and language users. In: *Handbook of Social Psychology*. Harper & Row, New York, NY, 179–231.

Convo. 2011. Convo. Retrieved September 30, 2015 from https://www.convorelay.com/.

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

Doubango Telecom. 2009. IMSDroid-High Quality Video SIP/IMS client for Google Android. Retrieved September 30, 2015 from http://code.google.com/p/imsdroid/.

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

D. Fitzgerald. 2013. How much smartphone data do you really need? *The Wall Street Journal*. Retrieved September 30, 2015 from http://blogs.wsj.com/digits/2013/08/01/how-much-smartphone-data-do-you-really-need/.

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

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

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

J. Hollington. 2013. Costs associated with using FaceTime. *iLounge*. Retrieved September 30, 2015 from http://www.ilounge.com/index.php/articles/comments/costs-associated-with-using-facetime/.

S. Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 2, 65–70.

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

B. F. Johnson and J. K. Caird. 1996. The effect of frame rate and video information redundancy on the perceptual learning of American sign language gestures. In *Proceedings of the CHI'96 Conference Companion on Human Factors in Computing Systems*, ACM, New York, NY. 121–122.

R. Koul. 2003. Synthetic speech perception in individuals with and without disabilities. 19, 1, 49–58.

Kurtnoise. 2009. Yet another MP4 box user interface for Windows users. Retrieved September 30, 2015 from http://yamb.unite-video.com/index.html.

H. Lane. 1992. *The Mask of Benevolence: Disabling the Deaf Community*. Alfred A. Knopf, Inc., New York, NY.

S. Lawson. 2011. Mobile growth driving out unlimited data. Retrieved September 30, 2015 from http://www.pcworld.com/businesscenter/article/242376/mobile_growth_driving_out_unlimited_data.html.

C. Lucas and C. Valli. 2000. *Linguistics of American Sign Language: An Introduction*. Gallaudet University Press, Washington, DC.

J. Maher. 1996. *Seeing Language in Sign: The Work of William C. Stokoe.* Gallaudet University Press, Washington, DC.

G. Marshall. 2014. How much 4G data do you really need? Retrieved September 30, 2015 from http://www.techradar.com/us/news/phone-and-communications/mobile-phones/how-much-4g-data-do-you-really-need–1176594.

M. Masry and S. S. Hemami. 2001. An analysis of subjective quality in low bit rate video. *International Conference on Image Processing*, IEEE, 465–468.

M. Masry and S. Hemami. 2003. CVQE: A metric for continuous video quality evaluation at low bit rates. *SPIE Human Vision and Electronic Imaging*.

J. McCarthy, M. A. Sasse, and D. Miras. 2004. Sharp or smooth? Comparing the effects of quantization vs. frame rate for streamed video. *Proceedings of the CHI*.

Merriam-Webster. 2003. *The Merriam-Webster Dictionary*. http://www.merriam-webster.com (8 May 2003).

Microsoft. 2013. How much data will Skype use on my mobile phone? http://community.skype.com/t5/Other-features/How-much-data-does-skype-use/td-p/897886.

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

Cisco. 2015. Cisco visual networking index:global mobile data trafic forecast update, 2014–2019. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf.

A. Nemethova, M. Ries, M. Zavodsky, and M. Rupp. 2006. PSNR-based estimation of subjective time-variant video quality for mobiles. *Proceedings of MESAQIN 2006*, Prag, Tschechien, June, 2006.

N. Omoigui, L. He, A. Gupta, J. Grudin, and E. Sanocki. 1999. Time-compression. *Proceedings of CHI*, ACM Press, New York, NY, 136–143.

A. Oppenheim and R. Schafer. 1975. *Discrete-Time Signal Processing*. Pearson.

C. Padden and T. Humphries. 2005. *Inside Deaf Culture*. Harvard University Press, Boston, MA.

J. Postel. 1980. *User Datagram Protocol–RFC 768*. https://tools.ietf.org/html/rfc768.

Purple. 2014. Purple VRS on Your Devices. Retrieved September 30, 2015 from http://www.purple.us/.

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

I. Richardson. 2004. vocdex: H.264 tutorial white papers. http://www.vcodex.com/h264.html.

E. Riskin, R. Ladner, and J. Wobbrock. 2012. MobileASL. University of Washington. Retrieved September 30, 2015 from http://mobileasl.cs.washington.edu/.

J. Rosenberg, H. Schulzrinee, G. Camarillo, et al. 2002. *SIP: Session Initiation Protocol*. RCS 3261. https://tools.ietf.org/html/rfc3261.

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

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

Skype. 2011. Skype. Retrieved September 30, 2015 from http://www.skype.com/intl/en-us/home.

Sorenson. 2014. Sorenson Communications. Retrieved September 30, 2015 from http://www.sorenson.com/.

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

Static Brain Research Institute. 2012. Skype statistics. Retrieved September 30, 2015 from http://www.statisticbrain.com/skype-statistics.

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

T-Mobile. 2014. T-Mobile. Retrieved September 30, 2015 from http://www.t-mobile.com/cell-phone-plans/individual.html#lshop_plans_1.

J. J. Tran, B. Flowers, E. Riskin, R. Ladner, and J. O. Wobbrock. 2014. Analyzing the intelligibility of real-time mobile sign language video transmitted below recommended standards. *Proceedings of ASSETS*, 177–184.

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

J. J. Tran, E. Riskin, R. Ladner, and J. O. Wobbrock. 2013. Increasing mobile sign language video accessibility by relaxing video transmission standards. *Third Mobile Accessibility Workshop at Proceedings of CHI*.

Verizon. 2014. Verizon Wireless. Retrieved September 30, 2015 from http://www.verizonwireless.com/b2c/index.html.

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

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

E. Weber. 1834. De pulsu, resorptione, auditu et tactu. *Anatationes anatomicae et physiologicae*.

T. Wiegang, H. Schwarz, A. Joch, F. Kossentini, and G. Sullivan. 2003. Rate-constrained coder control and comparison of video coding standards. *IEEE Transactions on Circuits and Systems for Video Technology* 13, 7, 688–703.

S. Winkler and P. Mohandas. 2008. The evolution of video quality measurement: From PSNR to hybrid metrics. *IEEE Transactions on Broadcasting* 54, 3, 660–668.

J. O. Wobbrock, L. Findlater, D. Gergie, and J. J. Higgins. 2011. The Aligned Rank Transform for nonparametric factorial analyses using only ANOVA procedures. *Proceedings of CHI*, 143–146.

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

E. Zeman. 2010. "iPhone 4 jailbreak unlocks 3G FaceTime calls. *Information Week*. Retrieved September 30, 2015 from http://www.informationweek.com/mobile/mobile-devices/iphone-4-jailbreak-unlocks-3g-facetime-calls/d/d-id/1091309?

ZVRS. 2014. ZVRS Communication Service for the Deaf, Inc. http://www.zvrs.com/products/softwareapps.

Received June 2014; revised June 2015; accepted June 2015