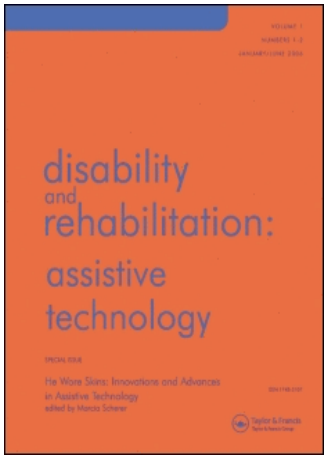

Disability and Rehabilitation: Assistive Technology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t741771157

MobileASL: Intelligibility of sign language video over mobile phones

Online Publication Date: 01 January 2007
To cite this Article: Cavender, Anna, Vanam, Rahul, Barney, Dane K., Ladner, Richard E. and Riskin, Eve A. (2007) 'MobileASL: Intelligibility of sign language video over mobile phones', Disability and Rehabilitation: Assistive Technology, 1 - 13
To link to this article: DOI: 10.1080/17483100701343475
URL: http://dx.doi.org/10.1080/17483100701343475

Disability and Rehabilitation: Assistive Technology, 2007; 1 – 13, i*First* article

informa
healthcare

ORIGINAL ARTICLE

MobileASL: Intelligibility of sign language video over mobile phones

ANNA CAVENDER[1], RAHUL VANAM[2], DANE K. BARNEY[1], RICHARD E. LADNER[1] & EVE A. RISKIN[2]

[1]*Department of Computer Science and Engineering, Box 352350, University of Washington, Seattle, WA 98195, USA, and* [2]*Department of Electrical Engineering, Box 352500, University of Washington, Seattle, WA 98195, USA*

Abstract

For Deaf people, access to the mobile telephone network in the United States is currently limited to text messaging, forcing communication in English as opposed to American Sign Language (ASL), the preferred language. Because ASL is a visual language, mobile video phones have the potential to give Deaf people access to real-time mobile communication in their preferred language. However, even today's best video compression techniques can not yield intelligible ASL at limited cell phone network bandwidths. Motivated by this constraint, we conducted one focus group and two user studies with members of the Deaf Community to determine the intelligibility effects of video compression techniques that exploit the visual nature of sign language. Inspired by eye tracking results that show high resolution foveal vision is maintained around the face, we studied region-of-interest encodings (where the face is encoded at higher quality) as well as reduced frame rates (where fewer, better quality, frames are displayed every second). At all bit rates studied here, participants preferred moderate quality increases in the face region, sacrificing quality in other regions. They also preferred slightly lower frame rates because they yield better quality frames for a fixed bit rate. The limited processing power of cell phones is a serious concern because a real-time video encoder and decoder will be needed. Choosing less complex settings for the encoder can reduce encoding time, but will affect video quality. We studied the intelligibility effects of this tradeoff and found that we can significantly speed up encoding time without severely affecting intelligibility. These results show promise for real-time access to the current low-bandwidth cell phone network through sign-language-specific encoding techniques.

Keywords: *Video compression, eye tracking, American Sign Language (ASL), deaf community, mobile telephone use*

1. Introduction

MobileASL is a video compression project that seeks to enable wireless cell phone communication through sign language.

1.1 *Motivation*

Mobile phones with video cameras and the ability to transmit and play videos are rapidly becoming popular and more widely available. Their presence in the marketplace could give Deaf[1] people access to the portable conveniences of the wireless telephone network.

The ability to wirelessly transmit video, as opposed to just text or symbols, would provide the most efficient and personal means of mobile communication for members of the Deaf Community: deaf and hard of hearing people, family members, and friends who use ASL. Some members of the Deaf Community currently use text messaging, such as Short Message Service (SMS), instant messaging (IM), or Teletypewriters (TTY). However, text is cumbersome and impersonal because (a) English is not the native language of most Deaf people in the United States (ASL is their preferred language), and (b) text messaging is slow and tedious at 5 – 25 words per minute (wpm) [1] compared to 120 – 200 wpm for both spoken and signed languages. Many people in the Deaf Community use video phones which can be used to call someone with a similar device directly or a video relay service. Video relay services enable phone calls between hearing people and Deaf people through the use of a remote human interpreter who translates video sign language to spoken language. This requires equipment (a computer, camera, and internet connection) that is generally set up in the home or work place and does not scale well for mobile use [2]. Video cell phones have the potential

Correspondence: Anna Cavender, Department of Computer Science and Engineering, Box 352350, University of Washington, Seattle, WA 98195, USA. E-mail: cavender@cs.washington.edu

ISSN 1748-3107 print/ISSN 1748-3115 online © 2007 Informa UK Ltd.
DOI: 10.1080/17483100701343475

to make the mobile phone network more accessible to over one million Deaf people in the US [3].

Unfortunately, the Deaf Community in the US cannot yet take advantage of this new technology. Our preliminary studies strongly suggest that even today's best video encoders cannot produce the quality video needed for intelligible ASL in real time, given the bandwidth and computational constraints of even the best video cell phones.

Realistic bit rates on existing GPRS networks typically vary from 30 – 80 kbps for download and perhaps half that for upload [4]. While the upcoming 3G standard [5] and special rate multi-slot connections [4] may offer much higher wireless bit rates, video compression of ASL conversations will still play an important role in realizing mobile video phone calls. First, there is some uncertainty about when 3G technology will become broadly available and, when it does, it will likely be initially restricted to highly populated areas and suffer from dropped calls and very poor quality video as is currently the case in London [6]. Furthermore, degradations in signal-to-noise ratio conditions and channel congestion will often result in lower actual bit rates, packet loss, and dropped calls. More importantly, fair access to the cell phone network means utilizing the already existing network such that Deaf people can make a mobile video call just as a hearing person could make a mobile voice call: without special accommodations, more expensive bandwidth packages, or additional geographic limitations. As such, video compression is a necessity for lowering required data rates and allowing more users to operate in the network, even in wireless networks of the future. The goal of the MobileASL project is to provide intelligible compressed ASL video, including detailed facial expressions and accurate movement and gesture reproduction, at less than 30 kbps so that it can be transmitted on the current GPRS network. A crucial first step is to gather information about the ways in which people view sign language videos.

1.2 *Contributions*

We conducted one focus group and two user studies with local members of the Deaf Community in Seattle to investigate the desire and/or need for mobile video phone communication, the technical and non-technical challenges involved with such technology, and what features of compressed video might enhance understandability.

The purpose of the focus group was to elicit feedback from the target user group about the ways in which mobile video communication could be useful and practical in their daily lives.

The user study was inspired by strongly correlated eye movement data found independently by Muir and Richardson [7] and Agrafiotis et al. [8]. Both projects used an eyetracker to collect eye movement data while members of the Deaf Community watched sign language videos. Results indicate that over 95% of gaze points fell within 2° visual angle of the signer's face. Skilled receivers of sign focus their gaze on or near the lower face of the signer. This is because contextual information coming from the hands and arms is relatively easy to perceive in peripheral or parafoveal vision, whereas contextual information from the face of the signer (which is also important in sign language) requires the level of detail afforded by high resolution foveal vision.

Based on these results, we conducted a study to investigate effects of two simple compression techniques on the intelligibility of sign language videos. We created a fixed region-of-interest (ROI) encoding by varying the level of distortion in a fixed region surrounding the face of the signer. The ROI encodings result in better quality around the face at the expense of quality in other areas. We also varied the frame rates of the video so that for a given bit rate, either 15 lower quality frames or 10 higher quality frames were displayed per second.

Our second study examined video encoding limits due to the processing power of today's cell phones. Because our goal is to allow real-time communication over cell phones with minimum delay, video must be encoded very quickly. The processing speed of the encoder can be reduced by adjusting parameter settings resulting in a less complex encoding, but this affects the video quality. We studied three different levels of quality and complexity and their effects on intelligibility.

Results from these studies indicate that minor adjustments to standard video encoding techniques, such as ROI, reduced frame rates, and lower complexity encoding parameters may allow intelligible ASL conversations to be transmitted in real-time over the current US cell phone network.

Section 2 next discusses related work. In Section 3, we will share participant responses from the MobileASL focus group. Section 4 explains the video compression user study. Section 5 presents the quality and complexity tradeoff study. Section 6 presents future work and concludes.

2. Related work

Previous research has studied the eye movement patterns of people as they view sign language through the use of an eye tracker [7,8]. Both groups independently confirmed that facial regions of sign language videos are perceived at high visual resolution and that movements of the hands and arms are generally perceived with lower resolution parafoveal vision. A video compression scheme (such as an ROI

encoding) that takes these visual patterns into account is recommended. Video segmentation of the important regions of sign language videos has been implemented using several different methods and shows promise for reducing bit rate through ROI encodings [9,10]. None of these methods have been empirically validated by potential users of the system.

Furthermore, guidelines recommending between 12 and 20 frames per second (fps) have been proposed for low bit rate video communication of sign language [11]. These claims have also not been empirically validated to the authors' knowledge.

Another line of research has pursued the use of technology to interpret between spoken languages and signed languages (for example [12–14]. While these translation technologies may become useful in limited domains, the goal of our project does not involve translation or interpretation.

Rather than focusing on ways to translate between written/spoken and signed languages, we feel that the best way to give Deaf people access to the conveniences of mobile communication is to bring together existing technology (such as large screen mobile video phones) with existing social networks (such as ASL interpreting services). The only missing link in this chain of communication is a way to transfer intelligible sign language video over the mobile telephone network in real time.

3. Focus group

We wanted to learn more about potential users of video cell phone technology and their impressions about how, when, where, and for what purposes video cell phones might be used. We conducted a 1-h focus group with four members of the Deaf Community ranging in age from mid-20s to mid-40s. The conversation was interpreted for the hearing researcher by a certified sign language interpreter. The discussion centered around the following general topics and responses are summarized below:

Physical setup

The camera and the screen should face the same direction. Some phones have cameras facing away from the screen so that one can see the picture while aiming the camera. This obviously would not work for filming oneself, as in a sign language conversation.

The phone should have a way to prop itself up, such as a kickstand. While some conversations could occur while holding the phone with one hand, longer conversations may require putting the phone on a table or shelf.

A slim, pocketable phone was desired. However, connecting a camera that captures better quality video was proposed (similar to using a Bluetooth headset).

A full PDA-style keyboard was desired as text will likely still be an important means of communication.

Features

All participants agreed that the phone should have all of the features currently found in Sidekicks or Blackberry PDA-phones, such as email and instant messaging. The Deaf Community has become accustomed to having these services and will not want to carry around two separate devices.

Even though participants all agreed that video communication is a huge improvement over text, they still felt that text messages would be an important feature. Text may be used to initiate a phone call (like ringing someone), troubleshoot (e.g., 'I can't see you because . . .'), or simply as a fall-back when the connection is bad or when conditions are not favorable for a video call. Participants thought text should be an option during a video call, much like simultaneous text messaging options in online video games.

There should be an easy way to accept or decline a video call. When a call is declined or missed, the caller should be able to leave a video message.

Video calls should be accessible to/from other video conferencing software so that calls can be made between video cell phones and web cams or set top boxes.

Packet loss

Networks are notoriously unreliable and information occasionally gets lost or dropped. The solution to this in a video sign language conversation is simply to ask the signer to repeat what was missed. However, all participants agreed that video services would not be used, or paid for, if packet losses were too frequent.

Scenarios

We discussed several scenarios where the video phone might or might not be useful. Two examples are as follows:

What if the phone rings when driving or on the bus?

There should be an easy way to dismiss the call, or change the camera angle so that the phone could be placed in one's lap while on the bus. Participants proposed that the phone could also be mounted on the dash board of a car. People already sign while driving, even to people in the back seat through the rear-view mirror, so participants thought that this

would not be very different. It could be even more dangerous than talking while on the cell phone and participants thought its use may be affected by future cell phone laws.

What if there was no table available to set the phone down?

One-handed signing for short conversations is not a problem: people sign while drinking, eating, smoking, etc. But, if the location is bad, like a crowded bar, texting may be easier.

One participant succinctly explained, 'I don't foresee any limitations. I would use the phone anywhere: at the grocery store, on the bus, in the car, at a restaurant, on the toilet, anywhere!'

In order for these scenarios to become reality, a better method for encoding (and compressing) video is needed such that intelligible ASL can be transmitted over the low bandwidth cell phone network.

4. User Study #1: User preferences

Inspired by the results from Muir and Richardson [7] and Agrafiotis et al. [8], we conducted a study with members of the Deaf Community to investigate the intelligibility effects of three levels of increased visual clarity in a small region around the face of the signer (ROI) as well as two different frame rates (fps). These factors were studied at three different bit rates comparable to those available in the current US cell phone network, totaling 18 different encoding techniques. Eighteen different sign language videos were created for this study so that each participant could be exposed to every encoding technique without watching the same video twice (i.e. a repeated measures design).

The videos were recordings of short stories told by a local Deaf woman at her own natural signing pace. They varied in length from 0:58 to 2:57 min (mean = 1:58) and all were recorded with the same location, lighting conditions, background, and clothing. The x264 encoder, an open source implementation of the H.264 (MPEG-4 part 10) encoder, was used to compress the videos with the 18 encoding techniques [15]. H.264/MPEG4 AVC is the latest video coding standard and has coding efficiency of two over MPEG-2 [16]. See Appendix A for a complete listing of encoding parameters used for the study videos.

Both videos and questionnaires were shown on a Sprint PPC 6700, PDAstyle video phone with a 320 × 240 pixel resolution (2.8″ × 2.1″) screen. All studies were conducted in the same room with the same lighting conditions.

4.1 *Baseline video rating*

Original recordings yielded 22 total videos of which 18 were chosen for this study for the following reasons. Undistorted versions of all 22 videos were initially rated for level of difficulty by three separate participants (one Deaf, two hearing) who considered themselves fluent in ASL. The purpose of the rating was to help eliminate intelligibility factors not related to compression techniques. After viewing each video, participants were asked one multiple choice question about the content of the video and then asked to rate the intelligibility of the video using a five-point Likert scale with unmarked bubbles on a range from 'difficult' to 'easy'. We will refer to those bubbles as '1' through '5' here.

The first participant rated all 22 videos as '5', the second rated 20 of the videos as '5' and two as '4', and the third participant also rated 20 of the videos as '5' and two as '4' (although the two were distinct from the ones rated a '4' by the second participant). The four videos that were given a rating a '4' were excluded from the study so that only the remaining 18 videos were used. In fact, *post hoc* analysis of the results from the study found no significant differences between the ratings of any of these 18 videos. This means we can safely assume that the intelligibility results that follow are due to varied compression techniques rather than other confounding factors (e.g., signer speed, difficulty of signs, lighting or clothing issues that might have made some videos more or less intelligible than others).

4.2 *Bit rates*

In an attempt to accurately portray the current US mobile network, we studied three different bit rates: 15, 20, and 25 kb per second (kbps). The optimal download rate of the GPRS network is estimated at 30 kbps, whereas the upload rate is considerably less, perhaps as low as 15 kbps.

4.3 *Frame rates*

We studied two different frame rates: 10 and 15 fps. Preliminary tests with a certified sign language interpreter revealed that 10 and 15 fps were both acceptable for intelligible ASL. The difference between 30 and 15 fps was negligible, whereas at 5 fps signs were difficult to see and fingerspelling was nearly impossible to understand.

Frame rates of 10 and 15 fps were chosen for this study to investigate the tradeoff of fewer frames at slightly better quality or more frames at slightly worse quality for any given bit rate. For example, a video encoded at 10 fps has fewer frames to encode than

the same video at 15 fps, so more bits can be allocated to each frame.

4.4 *Region of interest*

Finally, we studied three different ROI values: −0, −6, −12, where the negative value represents the reduced quantizer step size, out of 52 possible step sizes, in a fixed 6 × 10 macroblock region around the face (a single 320 × 240 pixel frame is composed of 15 × 20 macroblocks). Reducing the quantizer step size in this region results in less compression (better quality) in the face region and more compression (sacrificing quality) in all other regions for a given bitrate. An ROI value of −0 means there is no difference in the aforementioned regions (i.e. a typical encoding). An ROI value of −6 doubles the quality in the face region, distributing the remaining bits over the other regions. And an ROI value of −12 results in a level of quality four times better than a typical encoding around the signer's face sacrificing even more quality in surrounding regions. Snapshots of videos encoded with the three ROI values shown to participants can be seen in Figure 1.

As with frame rate, the ROI values for this study were chosen based on preliminary studies conducted with a certified sign language interpreter.

4.5 *Video order*

Because we can assume that higher bit rates yield more intelligible videos, we chose to structure the order in which videos were shown so that analysis of the data for the three bit rates could safely be separated. Thus, the study was partitioned into three parts, one for each bit rate. The same videos were shown in each partition, but their order within the partition was randomized. The order with which the three parts of study were conducted was determined by a Latin-squares design. The order with which the six different encodings (combinations of 2 frame rates and 3 ROIs) were shown within each part was also determined by a Latin-squares design (meaning each participant watched 18 different encodings in a different order to avoid effects of learning and/or fatigue).

4.6 *Subjective questionnaire*

After each video, participants answered a three-question, multiple choice survey given on the phone's screen and answered using the phone's stylus (see Figure 2a–c). The first question asked about the video content, for example, 'Who was the main character in the story?' This question was simply asked in order to encourage participants to pay close attention to the content of the videos, not necessarily to assess their understanding of the video. It would have been extremely difficult to devise questions of similar linguistic difficulty across all videos. Although we were not interested in the answers to this first question, it is worth mentioning that the correctness of the participants' answers often did not correlate with their answers to the remaining questions. For example, it was not uncommon for the participants to answer the first question correctly and then report that the video was difficult to comprehend and that they would not use a mobile phone that offered that quality of video, and vice versa.

The remaining two questions were repeated for each video. These two questions appear in Figure 2b,c and ask (1)'How easy or how difficult was it to understand the video?' which we will refer to as 'understand', and (2) 'If video of this quality was available on a cell phone, would you use it?' which we will refer to as 'use'. Answers to both questions were constrained by a five-point Likert scale (just as in the Baseline Video Rating) where participants could choose from five bubbles labeled with the ranges (1) difficult . . . easy and (2) no . . . maybe . . . yes.

4.7 *Results of user preferences study*

Eighteen adult members of the Deaf Community (seven women, 11 men) participated in this study.

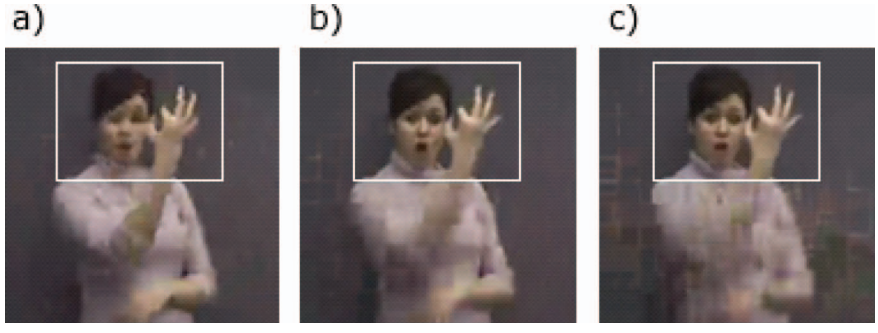

Figure 1. Cropped video frame at (a) −0 ROI (standard encoding), (b) −6 ROI (two times better quality in the face region), and (c) −12 ROI (four times better quality in the face region).

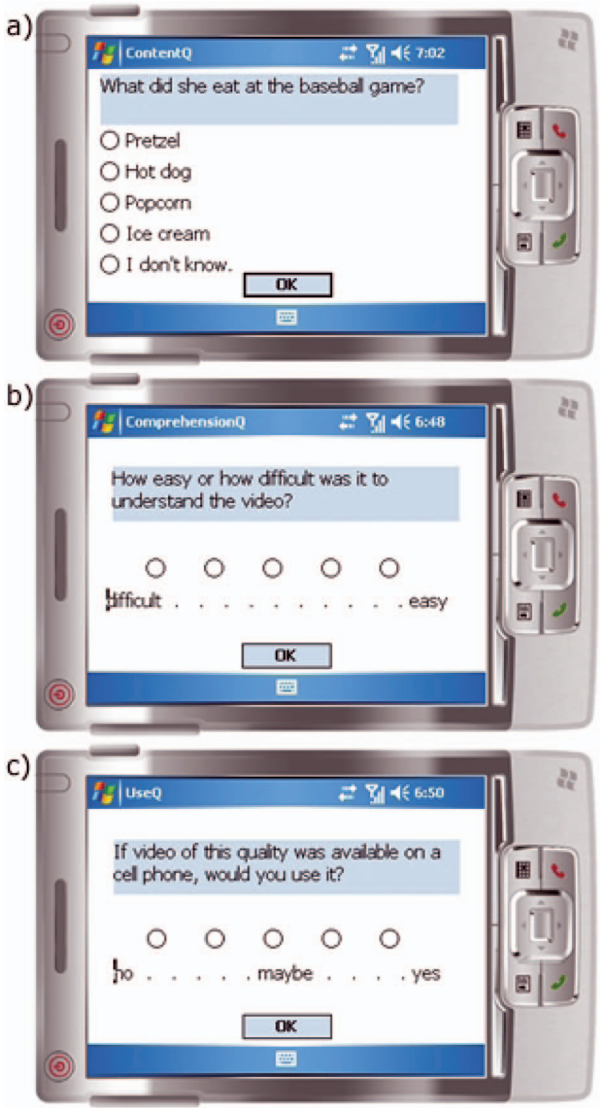

Figure 2. Series of questions in post-video questionnaires asking about (a) the content of the video (different for each video), (b) the understandability of the video, and (c) the participant's willingness to use the quality of video just seen.

Ten participants were Deaf, five hearing, and three CODAs (child of a deaf adult). CODAs are born into Deaf households and often consider ASL their first language and English their second language. All 10 Deaf participants and three CODAs had lifelong experience with ASL; the five hearing participants had an average of 10.6 years (SD = 5.81) experience with ASL.

After a short demographic survey, participants were shown two practice videos. Participants were told that these videos were examples of the worst and best videos to be viewed so that participants had points of reference when rating upcoming videos.

We noticed that some participants tended to favor higher scores, rating many videos as '4' and '5', while other participants favored lower scores, rating videos as '1' and '2'. In order to more fairly evaluate these absolute ratings, we will show both absolute and standardized ratings using z-scores as follows:

$$z = \frac{\mathrm{X} - \mu}{\sigma}$$

where X is the raw score to be standardized, μ is the participant's mean rating, and σ is the participant's standard deviation. This effectively maps each participant's ratings onto a range centered at 0, where videos with a score of zero indicate a participant's average rating. Videos with a negative score are considered below average and videos with a positive score are considered above average by that participant. Scores are in units of standard deviation.

Participants then watched 18 videos and answered the three-question, multiple choice survey for each video. Each video was encoded with a distinct encoding technique as described above. The last 5–10 min of each study session was spent gathering anecdotal information about the participant's impressions of the video cell phone and video quality.

Analysis of survey questions responses indicates that participant preferences for all three variables (bit rate, frame rate, and ROI) were largely independent of each other. For example, the results for ROI encodings held true regardless of changes in bit rate or frame rate.

Also, answers to the two questions 'understand' and 'use' were highly correlated (with a Pearson's correlation coefficient $r(16) = 0.85$ and $P < 0.01$). Because 'understand' and 'use' are so highly correlated, we will only show graphs for 'understand', but note that data for 'use' looks very similar.

4.7.1 Bit rates. As expected, survey responses indicate very strong and statistically significant preferences for higher bit rates: 25 kbps was preferred over 20 kbps, which in turn was preferred over 15 kbps ($F(2, 34) = 51.12$, $P < 0.01$). These results can be seen in Figure 3 and the standardized z-scores can be seen in Figure 4. Higher bit rates were preferred regardless of different frame rates and ROI values of the videos.

4.7.2 Frame rates. For the two different frame rates studied here (10 fps and 15 fps), Figures 3 and 4 show a preference toward 10 fps ($F(1, 17) = 4.59$, $P < 0.05$). A likely explanation is that the difference between 10 and 15 fps is acceptable (in fact some participants did not realize the videos had different frame rates until after the study, during the anecdotal questions). Furthermore, the increased frame quality, due to the fewer number of frames to encode for the same number of bits, may have been a desirable tradeoff.

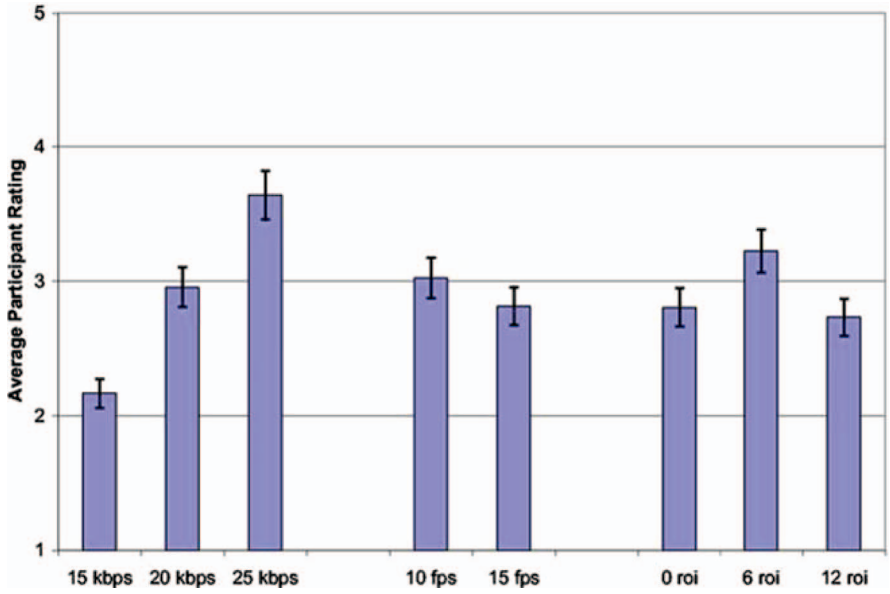

Figure 3. Qualitative results for different bit rates, frame rates, and ROI values, averaged over participants. Error bars represent confidence intervals.

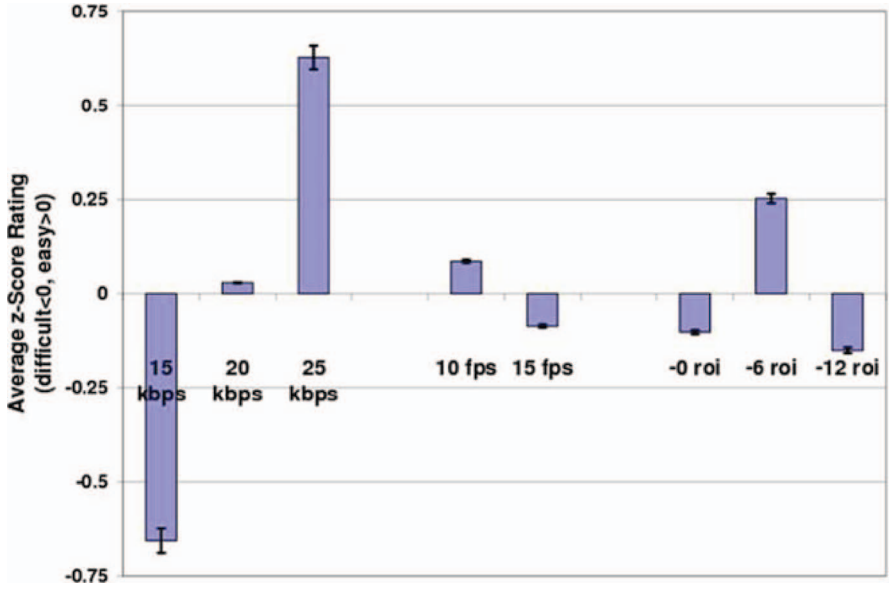

Figure 4. Standardized results for different bit rates, frame rates, and ROI values, using z-scores. Error bars represent confidence intervals.

4.7.3 Region of interest. For the three different ROI encodings, we found a very significant effect ($F(2, 34) = 13.69$, $P < 0.01$). At every bit rate shown, participants preferred an ROI of -6 (the middle value tested). Figures 3 and 4 show that ratings for -0 ROI (no ROI) and -12 ROI were not statistically different. This could indicate that an appropriate range of values was chosen for the study and that there may exist an optimal tradeoff between clarity around the face and distortion in other regions.

Figures 5 and 6 show the average ratings and z-score ratings for the baseline encoding techniques for this study (i.e. 15 fps and no ROI) at the three bit rates tested. The graph also shows the encoding technique that received the best average participant rating: 25 kbps, 10 fps, and -6 ROI. There is a substantial increase in preference for encodings that add both a reduced frame rate and a moderate ROI improvement.

These results indicate that these two simple and effective compression techniques may better utilize low bandwidth connections (such as through the cell phone network) to yield more intelligible ASL. Also, reducing the frame rate to 10 fps and decreasing the quantization by six step sizes in macroblocks around

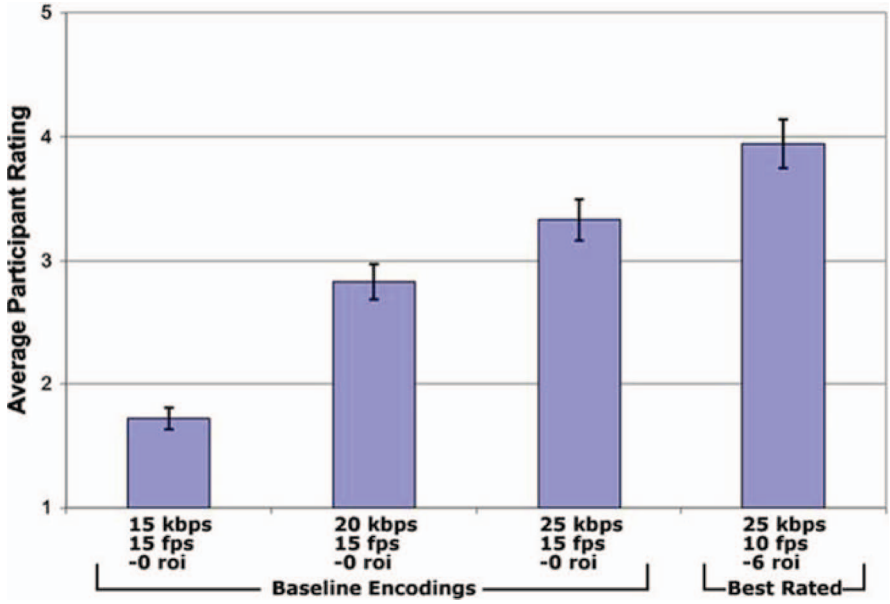

Figure 5. Qualitative results for three baseline bit rate encodings (25, 20, and 15 kbps at 15 fps, −0 ROI) shown against the encoding technique with the maximum average rating (25 kbps, 10 fps, −6 ROI). Error bars represent confidence intervals.

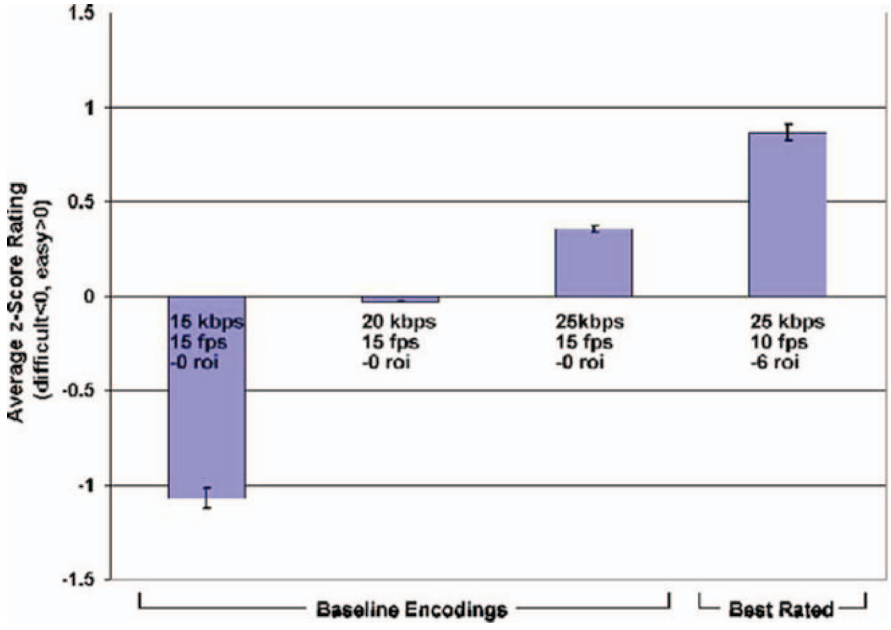

Figure 6. Standardized results for three baseline bit rate encodings (25, 20, and 15 kbps at 15 fps, −0 ROI) shown against the encoding technique with the maximum average rating (25 kbps, 10 fps, −6 ROI) using *z*-scores. Error bars represent confidence intervals.

the face can be executed without increases in encoding time (which is important when encoding on the limited processors of cell phones).

5. User Study #2: Lowering complexity through encoder parameter selection

In order for video cell phones to be used for sign language communication, both encoding and decoding need to take place at the same time on the processor of the cell phone. Encoding in real-time is a complex task while real-time decoding is less computationally intensive. Minimizing the encoding time will be an important aspect of allowing real-time communication.

The H.264 video encoding standard has a number of different features that can be adjusted. These include the appropriate selection of reference frames, the variable block size used for motion compensation, the mode decision, and the type of motion

estimation used. More complicated encoding can lead to increased coded video quality, but of course, this requires additional time at the encoder. One can consider setting the H.264 parameters as 'tuning a set of knobs' where there is one knob for each H.264 feature. If a knob is turned all the way to high complexity, one can expect improved video quality, whereas if a knob is turned all the way down to low complexity, one can expect lower quality. To be able to encode video in real-time on today's cell phones, a simple encoder is needed, but not at the expense of intelligibility.

The User Preferences Study found that the users most preferred a bit rate of 25 kbps, a frame rate of 10 fps, and an ROI factor of -6. This study was done with the encoding parameters listed in Appendix A. Most of the selected parameters were turned up to high complexity settings: the motion estimation of '*subme* = 6' is the second highest complexity; the pixel motion estimation method (*me*) is the most complex; and using five references frames is significantly more complex than using one reference frame (*ref frames*). To learn more about H.264 features, see [16].

We profiled the encoding speed in terms of fps on a Sprint PPC 6700 PDA-style video cell phone with a 416-MHz Intel PXA-270 processor, running the Windows Mobile 5.0 operating system. With no optimizations to the code, the x264 encoder yielded only 3.1 frames/second on a 320×240 (QVGA) resolution video with low quality encoding settings. However, by taking advantage of the Wireless MMX instructions available on the PXA-270 processor, which allow for the processing of up to 8 pixels in parallel, we are able to achieve 6.2 frames/second on the same QVGA video. Figure 7 shows these different encoding rates before and after code optimization on a video at both QVGA and QCIF resolutions. Three different sets of parameters to the H.264 encoder were tested, corresponding to high, mid, and low quality compression. These settings are explained in more detail in the following section.

The settings used in the User Preferences Study lead to an encoder that can only encode 2.8 fps, whereas based on our user study, we would like to be able to encode 10 fps. As a result, we seek ways to lower the encoder complexity by adjusting H.264 parameter settings.

5.1 *Encoding parameters chosen*

In order to choose encoder parameters that reduce complexity at a given video quality, we encoded 10 ASL video clips with a set of selected parameter setting combinations at 25 kbps. Among the different settings, we chose three that result in highest, intermediate and lowest video quality and complexity. Settings were chosen such that they take the least amount of encoding time for a given video quality and were obtained using an approach proposed in [17]. For specific settings chosen for this study, see Appendix B.

5.2 *Study design*

In order to study these three different encoding techniques, we created six videos (two at each setting combination) so that each participant could be exposed to every encoding technique twice. We used the same sign language videos as in the User Preferences Study and the same x264

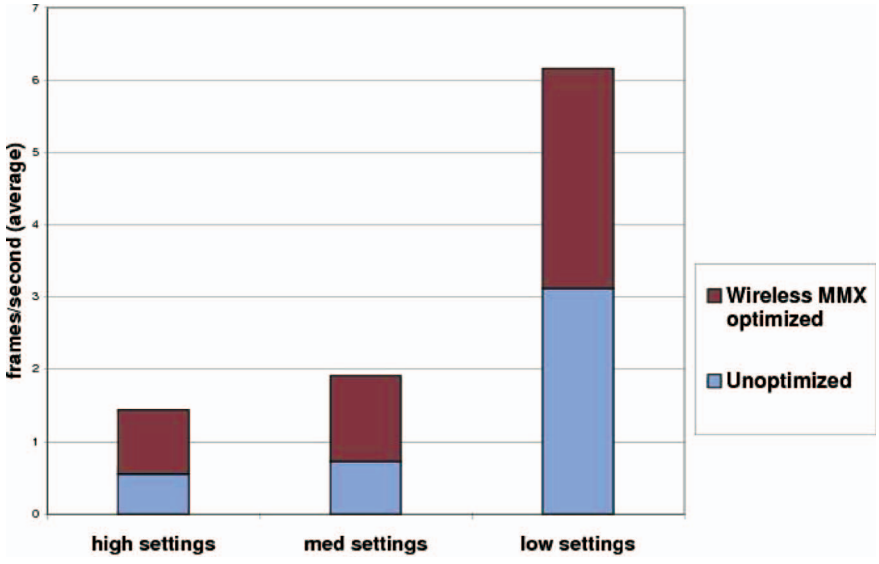

Figure 7. Encoding performance in frames per second for High, Mid, and Low quality/complexity settings with and without code optimization.

implementation of the H.264 encoder [15]. We again used 320 × 240 sized videos.

Participants were first shown one short practice video encoded with the low quality and complexity settings to serve as a point of reference for the following videos. The order in which participants watched the six different videos was randomized to avoid affects of learning and/or fatigue. After each video, participants rated the understandability and usefulness of the video using the same three-question multiple choice survey as in the previous study (see Figure 2).

5.3 *Results of encoder complexity study*

Six adult members of the Deaf Community (two women, four men; different from those who participated in the User Preferences Study) participated in this study. Five participants were Deaf and one was hearing. Two participants had life long experience with ASL, the remaining four had an average of 25.5 years (SD = 6.6) experience with ASL.

As shown in Figures 8 and 9, although survey responses indicate a mild increase in preference for more complex (better quality) encoder settings, these

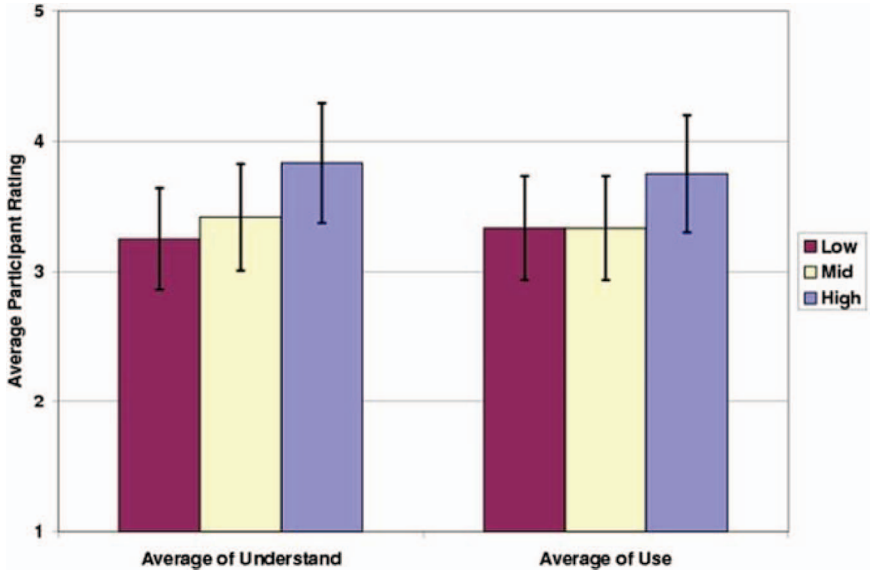

Figure 8. Qualitative results for three different encoder parameter settings (High, Mid, and Low) affecting complexity and quality, averaged over participants. Error bars represent confidence intervals.

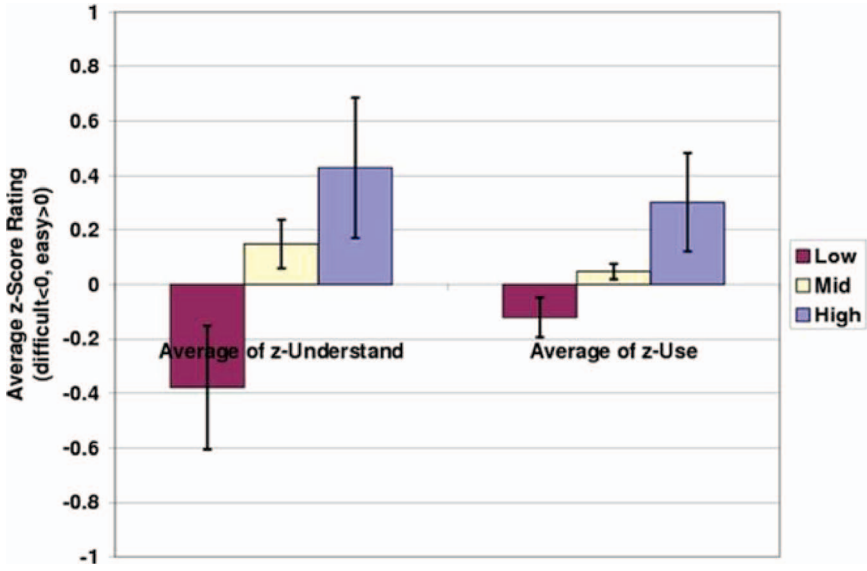

Figure 9. Standardized results for three different encoder parameter settings (High, Mid, and Low) affecting complexity and quality, using *z*-scores. Error bars represent confidence intervals.

results were not statistically significant ($F(2,5) = 0.72$, $P = 0.602$).

These results indicate that we can safely adjust some parameters (such as sub-pixel motion estimation, number of reference frames, and partition size for motion compensation) to improve encoding time without significantly affecting intelligibility. Referring to Figure 7, this user study has shown that we can increase the encoding frame rate from 2 to 6.2 fps (for resolution 320 × 240) without significantly affecting video quality.

Of course, future improvements in cell phone technology, including dedicated H.264 encoding hardware on the phone, may lead to significant increases in encoding speed. But, savings on encoder complexity will continue to be beneficial for sharing processor tasks and extending the battery life of the device.

6. Conclusion

The Deaf Community in the United States stands to benefit from new video cell phone technology as a mobile form of communication in their preferred language, American Sign Language (ASL). Low bandwidth constraints of the current cell phone network create many video compression challenges because even today's best video encoders cannot yield intelligible ASL at such low bit rates. Thus, real-time encoding techniques targeted toward sign language are needed to meet these demands.

This paper discussed the potential for and challenges involved with mobile video sign language communication over the US cell phone network. We investigated the potential needs and desires for mobile sign language communication through a targeted focus group with four members of the Deaf Community.

Motivated by highly correlated visual patterns of receivers of sign found by Muir and Richardson [7] and Agrafiotis et al. [8], we studied the effects of a ROI encoding (where the face regions of the video were encoded at better quality than other regions) and reduced frame rate encodings (where fewer, better quality frames are displayed every second). We studied these two factors at three different bit rates representing a lower-end possible range for transfer over the current cell phone network. Results indicate that reducing the frame rate to 10 fps and increasing the quality of the image near the signer's face may help yield more intelligible ASL for low bit rates. Increased quality per frame for 10 fps was a preferable tradeoff from 15 fps. A tradeoff of six decreased quantization steps near the face of the signer (doubling the quality in that region) was preferred over a typical (no ROI) encoding and over a larger quality increase in the face that caused more distortion in other regions of the video (a quantization difference of 12).

Through a second user study, we also found that we can significantly increase encoding speed (from 2 to 6.2 fps for resolution 320 × 240) without severely affecting video quality by implementing a reduced complexity encoder. This brings us much closer to our goal of encoding at a frame rate of 10 fps in real-time.

These findings are important from a video compression standpoint. Our results indicate that existing mobile phone technology, when coupled with a new means of compression, could be suitable for sign language communication. This combination could provide access to the freedom, independence, and portable convenience of the wireless telephone network from which Deaf people in the United States have previously been excluded.

6.1 *Future work*

The results from this work are currently being used to help define a new video compression intelligibility metric that will inform a compression scheme utilizing the new H.264 standard [18]. We have the following future goals for this project.

The regions used in this study for the ROI encodings were fixed in size and location. While the signer in our videos did not move her upper body outside of this rectangle, and since most signs occur within a 'sign box' surrounding the upper body, it may be more useful and/or more efficient to dynamically choose a region of interest based on information in the video. Many participants thought that the ROI was most problematic when the shape of the hands was lost due to distortion. An interframe skin detection algorithm [9] could likely help with this as it may include regions associated with hands in the ROI encoding. Similarly, because the hands are moving, motion vectors in the video (supposing a stationary background) could help define regions of interest where more bits could be allocated. This would be a valuable idea to test empirically as a moving region of interest may be distracting.

Because of the natural constraints of sign language (the shape, orientation, and location of arms, hands, and face) it may be useful to apply learning algorithms to the motion vectors of several training videos so that the motion vectors in other sign language videos may be more easily predicted, aiding in the speed and efficiency of the compression. For example, an encoder that could predict times of signing, fingerspelling, or 'just watching' could act accordingly: saving bits when no signing is occurring or allocating more bits when high quality video is needed, such as during the highly detailed movements of fingerspelling.

Participants in the Focus Group all agreed that packet loss will be a big concern and could render mobile video technology useless for ASL conversations. An area of future research will be investigating the implications, limitations, and effective ways to handle packet loss for video sign language communication.

The long-term goal of this project is to enable members of the Deaf Community to communicate using mobile video phones. Developing compression techniques that encode and decode in real-time on a mobile phone processor is an important and ongoing aspect of this project. For example, we are working on an encoder with a constant encoding time, which will adjust parameters in the coder based on how long it is taking to encode. Specifically, an encoder could reduce time spent searching for good motion vectors when decreased encoding time is needed whereas improvements to the encoding can be made when more encoding time is available. We will continue to optimize the H.264 encoder for both speed and video quality.

Acknowledgements

We would like to thank Sheila Hemami, Francis Ciaramello, and Alan Borning for their guidance and feedback throughout this project.

Thanks to Jessica DeWitt for helping to create the videos used in this study, Tobias Cullins for arranging interpreters, and all of the people who participated in the study. This research has been supported by the National Science Foundation through two Grants CCF-0514353 and CCF-0104800, an NSF Graduate Fellowship, and a Boeing Professorship.

Note

1. Capitalized Deaf refers to people who are active in the signing Deaf Community and Deaf Culture, whereas lowercase deaf is typically a more medical term.

References

1. James CL, Reischel KM. Text input for mobile devices: comparing model prediction to actual performance. In: CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems, pp 365–371.
2. Keating E, Mirus G. American sign language in virtual space: Interactions between deaf users of computer-mediated video communication and the impact of technology on language practices. Lang Soc 2003;32:693–714.
3. Mitchell R. How many deaf people are there in the United States?, 2005, http://gri.gallaudet.edu/Demographics/deaf-US.php.
4. GSMA, General packet radio service, 2006, http://www.gsmworld.com/technology/gprs/class.shtml.
5. 3GToday, 2006, http://www.3gtoday.com/.
6. 3GNewsroomcom, 3UK disgraced by BBC watchdog programme, October 22, 2003, http://www.3gnewsroom.com/.
7. Muir L, Richardson I. Perception of sign language and its application to visual communications for deaf people. J Deaf Studies Deaf Education 2005;10(4):390–401.
8. Agrafiotis D, Canagarajah CN, Bull DR, Dye M, Twyford H, Kyle J, Chung-How JT. Optimized sign language video coding based on eyetracking analysis. VCIP 2003;5150:1244–1252.
9. Habili N, Lim CC, Moini A. Segmentation of the face and hands in sign language video sequences using color and motion cues. IEEE Trans Circuits Syst Video Techn 2004; 14(8):1086–1097.
10. Schumeyer R, Heredia E, Barner K. Region of interest priority coding for sign language videoconferencing. IEEE First Workshop on Multimedia Signal Processing 1997;531–536.
11. Sector ITS, Draft application profile: Sign language and lip reading real time conversation usage of low bit rate video communication, 1998.
12. Vogler C, Metaxas D. A framework for recognizing the simultaneous aspects of American sign language. Comput Vis Image Underst 2001;81(3):358–384.
13. Starner T, Pentland A, Weaver J. Real-time American sign language recognition using desk and wearable computer based video. IEEE Trans Pattern Anal Mach Intell 1998;20(12): 1371–1375.
14. Bangham J, Cox SJ, Lincoln M, Marshall I, Tutt M, Wells M. Signing for the deaf using virtual humans. In: IEE Colloquium on Speech and Language Processing for Disabled and Elderly.
15. Aimar L, Merritt L, Petit E, Chen M, Clay J, Rullgrd M, Heine C, Izvorski A, x264 - a free H.264/AVC encoder, 2005, http://www.videolan.org/x264.html.
16. Richardson I. Vcodex: H.264 tutorial white papers, 2004, http://www.vcodex.com/h264.html.
17. Vanam R, Riskin EA, Hemami SS, Ladner RE. Distortion-complexity optimization of the H.264/MPEG-4 AVC encoder using the gbfos algorithm. In: Proceedings of IEEE Data Compression Conference, Snowbird, UT, USA, 2007. pp 303–312.
18. Ciaramello F, Cavender A, Hemami S, Riskin E, Ladner R. Predicting intelligibility of compressed american sign language video with objective quality metrics. In: 2006 International Workshop on Video Processing and Quality Metrics for Consumer Electronics.

Appendix A. x264 Parameters used in video phone study

The following lists parameters used for the x264 encodings of videos used the User Preferences Study and the Encoder Complexity Study.

Parameter	Study#1	Study#2
Resolution	320 × 240	320 × 240
– –subme	6	[Varied]
– –bframes	1	1
– –no – b – adapt	Yes	Yes
– –scenecut	–1	–1
–I	9999	9999
– –mixed – refs	Yes	Yes
– –meumh	Yes	Yes
– –directspatial	Yes	Yes
– –ref	5	[Varied]
–A	p8 × 8,p4 × 4, b8 × 8,i8 × 8,i4 × 4	[Varied]
– –trellis	Default	1
– –8 × 8dct	No	Yes

Description/definition of parameters:

Parameter	Description
Resolution	Width and height of video in square pixels
−−subme	Sub-pixel motion estimation, partition decision quality
−−bframes	Number of B-frames between I and P frames
−−no − b −adapt	Disable adaptive B-frame decision
−−scenecut	How aggressively to insert extra I frames
−I	Maximum GOP size
−−mixed − refs	Decide references on a per partition basis
−−meumh	Pixel motion estimation method, umh is uneven multi-hexigon search
−−directspatial	Direct MV prediction mode
−−ref	Number of reference frames
−A	Partitions to consider during analysis
−−trellis	Trellis RD quantization, requires CABAC
−−8 × 8dct	Adaptive spatial transform size

Appendix B. x264 Parameters used in video phone study

The following lists the way in which parameters were varied for the x264 encodings of videos used the Encoder Complexity Study.

We varied the parameters: sub-pixel motion estimation (*subme*), number of reference frames (*ref frames*) and partition size (*part*) to obtain the three settings High, Medium, and Low listed in Table I.

Table I. H.264 encoder parameters for High, Mid, Low video quality and complexity.

Parameters	High	Mid	Low
subme	7	6	1
ref frames	16	12	2
part	p8 × 8,p4 × 4, b8 × 8,i8 × 8,i4 × 4	i8 × 8,p8 × 8	i8 × 8,p8 × 8

Other parameters were kept constant and can be found in Appendix A. It should be noted that higher values for *subme* and *ref frames*, and a larger number of possible ways to partition a frame for motion compensation will result in better quality and higher complexity.