Quality versus Intelligibility: Studying Human Preferences for American Sign Language Video

Frank M. Ciaramello and Sheila S. Hemami

Visual Communications Laboratory
School of Electrical and Computer Engineering, Cornell University
Ithaca, NY, 14853

ABSTRACT

Real-time videoconferencing using cellular devices provides natural communication to the Deaf community. For this application, compressed American Sign Language (ASL) video must be evaluated in terms of the intelligibility of the conversation and not in terms of the overall aesthetic quality of the video. This work presents a paired comparison experiment to determine the subjective preferences of ASL users in terms of the trade-off between intelligibility and quality when varying the proportion of the bitrate allocated explicitly to the regions of the video containing the signer. A rate-distortion optimization technique, which jointly optimizes a quality criteria and an intelligibility criteria according to a user-specified parameter, generates test video pairs for the subjective experiment. Experimental results suggest that at sufficiently high bitrates, all users prefer videos in which the non-signer regions in the video are encoded with some nominal rate. As the total encoding bitrate decreases, users generally prefer video in which a greater proportion of the rate is allocated to the signer. The specific operating points preferred in the quality-intelligibility trade-off vary with the demographics of the users.

1. INTRODUCTION

Real-time, two-way transmission of American Sign Language (ASL) video over cellular networks provides natural communication among members of the Deaf community. When compressing and evaluating ASL video, traditional video quality estimators are inadequate; quality must be measured as the intelligibility of the signer, and not as the overall aesthetic quality of the video. Information in ASL is communicated through facial expressions and hand gestures and the intelligibility of compressed ASL video can be objectively computed by measuring the distortions in the signer's face, hands, and torso. This objective intelligibility measure, denoted the computational intelligibility model (CIM), accurately estimates subjective ratings of intelligibility provided by fluent ASL users.[1]

An intelligibility optimized encoder allocates rate within a frame according to the CIM and provides bitrate reductions up to 50%, at fixed levels of intelligibility, when compared to a mean-squared-error (MSE) optimized encoder.[2] The MSE optimized encoder nominally provides consistent levels of distortion across the entire frame, but is unable to produce intelligible video at low bitrates. The intelligibility optimized encoder achieves bitrate reductions by heavily distorting the background video region, while maximizing the fidelity of the signer. A subset of participants in a subjective experiment qualitatively reported distractions due to heavily distorted backgrounds, even when they considered the videos to be intelligible.[3] Allowing the user to adjust the level of background distortion addresses this problem, but lowering the distortion in the background region necessarily increases the distortion in the signer and can lead to an unintelligible video.

The goal of this work is to evaluate the preferences of ASL users in terms of this quality versus intelligibility trade-off, specifically identifying when a user is willing to sacrifice intelligibility (as measured by the CIM) for an increase in video quality (as measured by PSNR). A paired comparison experiment is conducted to identify user preferences for videos in which the relative amount of rate allocated between the signer and the background varies in a systematic way. A rate-distortion algorithm that jointly optimizes the CIM and PSNR is used to generate the test videos for the experiment and is summarized in Section 2. A detailed description of the subjective experiment is provided in Section 3. The experimental results, summarized and discussed in Section 4, support the need for a user-controlled trade-off between intelligibility and quality.

F.M.C: E-mail: fmc3@cornell.edu; S.S.H: E-mail: hemami@ece.cornell.edu

2. A QUALITY-INTELLIGIBILITY CODING TRADE-OFF

Rate-distortion (R-D) optimization for H.264 video requires the selection of a set of encoding parameters for each macroblock (16×16 block of pixels) that minimizes the distortion, subject to a target bitrate. Depending on the distortion measure being used by the encoder, the resulting rate allocated to any particular macroblock can vary significantly between macroblocks within a frame. The authors have developed a R-D optimization algorithm that has a single, user-specified parameter that can be adjusted to vary the percentage of rate allocated explicitly to the signer.[2] This coder achieves rate-distortion performance defined by the convex combination of a strictly quality optimized and a strictly intelligibility optimized encoder, providing a trade-off between the clarity of the signer and the amount of distortion in the background macroblocks of the video frame. The performance of this encoder, as well as some illustrative examples, are summarized in this section.

At one extreme, the strictly quality optimized R-D algorithm is designed to maximize the overall quality of the input video by allocating rate evenly to each macroblock in a frame. In this encoder, quality is measured in terms of PSNR. Although PSNR is unable to accurately estimate subjective quality across different videos and different distortion types, it can still be applied as a measure of video quality under certain constraints. In particular, when encoding a single video, it is fair to assume that increasing PSNR corresponds to an increase in subjective quality (or, more conservatively, a non-decrease in subjective quality). Given this assumption, PSNR is used here as an estimate for subjective quality.

At the other encoding extreme for sign language video, it is more appropriate to apply a R-D optimization algorithm designed to maximize an intelligibility criteria, rather than a quality criteria. This intelligibility optimized encoder allocates rate within a frame according to an intelligibility distortion measure, which is a function of the distortion only in linguistically relevant regions, i.e., the signer's face, hands, and torso.[1] The intelligibility distortion measure can be written as the sum of the weighted MSE in each of the relevant regions, computed according to

$$D_{Intell} = \frac{1}{N}\sum_{n=1}^{N} \alpha_F D_F(n) + \alpha_H D_H(n) \alpha_T D_T(n) + \alpha_{BG} D_{BG}(n) + D_{temporal}, \quad (1)$$

where D_F, D_H, D_T, and D_{BG} are the MSE for the face, hands, torso, and background regions in frame n for a video sequence having N total frames. The region weights of $\alpha_F = 1.6$, $\alpha_H = 0.5$, $\alpha_H = 0.1$, and $\alpha_{BG} = 0$ maximize the prediction accuracy of the intelligibility distortion measure with respect to ground-truth intelligibility ratings. Distortions in background macroblocks do not contribute to D_{Intell}; α_{BG} and D_{BG} are included in Eq. (1) to explicitly account for all macroblocks. The impact of temporal variations in the distortions is quantified by $D_{temporal}$.[1]

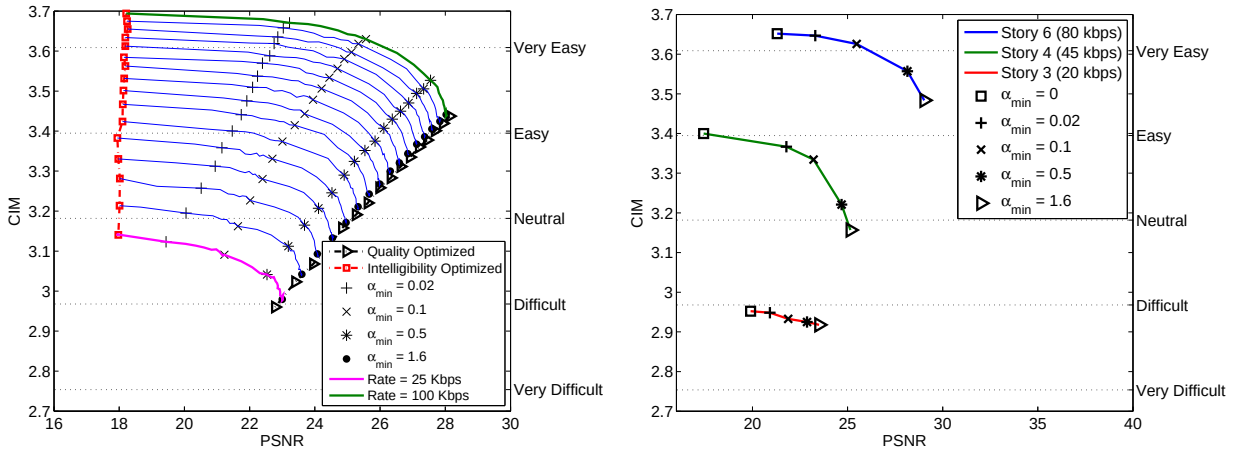
Note that D_{Intell} is a distortion measure and is inversely proportional to intelligibility. The varying weights control the relative importance of each type of macroblock; a distortion in the signer's face will result in a lower intelligibility than the same amount of distortion in the signer's torso. The intelligibility distortion is mapped to the CIM according to

$$CIM = \log_{10} \frac{C}{D_{Intell}}, \quad (2)$$

where $C = 110^2$ is a constant chosen empirically to map to an intelligibility scale.

These two encoding extremes alone are incapable of accommodating the preferences of ASL users while maintaining intelligible video. The user-specified quality-intelligibility encoding trade-off parameter is denoted α_{min} and specifies the minimum weight to be applied to all macroblocks in the frame. Specifically, if the weight α_k of any region (including the signer's face, hands, torso, or background) is less then α_{min}, then the weight α_k is changed and set equal to α_{min}. This provides a mechanism to increase the quality in the background, while guaranteeing that the background distortion weight is never higher than the distortion weights for the signer's face, hands, or torso.

Systematically varying α_{min} yields the convex combination of the quality optimized and intelligibility optimized encoders, as illustrated in Figure 1. As α_{min} increases, the R-D performance of the encoder sweeps the



(a) CIM vs PSNR for an ASL video.

(b) CIM vs PSNR for only the 3 videos and 5 values of α_{min} selected for the paired comparison experiment.

Figure 1. PSNR vs CIM plots for the quality-intelligibility optimized coder at several rates and values of α_{min}. The left y-axis provides the CIM and the right y-axis provides the subjective rating categories corresponding to the CIM values. In (a), each solid line corresponds to a fixed bitrate and a varying α_{min}. The bitrates vary between 25 kbps and 100 kbps in increments of 5 kbps. PSNR can be increased by several dB without a significant decrease in CIM, when compared to the strictly intelligibility optimized encoder.

space between the two encoding extremes. When encoding a video for a fixed target bitrate, the value of α_{min} determines the operating point in the trade-off between intelligibility and quality, as illustrated in Figure 1(a).

Modifying α_{min} controls the degree to which the regions of interest (ROIs) are prioritized over the rest of the frame. A region is considered prioritized if its corresponding distortion weight is larger than α_{min}. A prioritized region will have lower distortion, on average, than the rest of the frame. To illustrate, consider a sample ASL video encoded at 55 kbps with different values of α_{min}. Five values for α_{min} are selected to emphasize different operating points and are evaluated in the paired comparison experiment: $\alpha_{min} = 0$ prioritizes the entire ROI, $\alpha_{min} = 0.02$ prioritizes the entire ROI and provides a nominal amount of rate to the background, $\alpha_{min} = \alpha_T = 0.1$ prioritizes only the signer's face and hands, $\alpha_{min} = \alpha_H = 0.5$ prioritizes the signer's face, and $\alpha_{min} = \alpha_F = 1.6$ prioritizes no regions and corresponds to the quality optimized encoder. Frames from this video are presented in Figure 2. As α_{min} increases, the relative priority of the ROI necessarily decreases and intelligibility decreases, as illustrated in Figures 2(b) through 2(f). As this example demonstrates, varying α_{min} can provide a user with control over the level of background distortion while still prioritizing the most important regions of the signer.

3. PAIRED COMPARISON EXPERIMENT FOR IDENTIFYING USER PREFERENCES IN THE QUALITY-INTELLIGIBILITY TRADE-OFF

The coder described in Section 2 and the choice of α_{min} controls the trade-off between optimizing a video encoder for intelligibility and optimizing for quality. A paired comparison experiment is conducted to determine subjective preferences in this trade-off. The primary goal is to identify preferred operating points, if they exist, and to determine under what conditions a user likely to desire a particular operating points.

3.1. Stimuli

Reference sign language stories told by a fluent signer at her natural signing pace were filmed at an outdoor location on a busy street having a significant amount of background activity. Videos were recorded at a resolution of 1280×720 pixels and a frame rate of 60 progressive frames per second. For this experiment, the videos are cropped and downsampled in order to match the expected usage conditions, namely a mobile device having a

(a) Original video frame

(b) Prioritize all of the ROI. $\alpha_{min} = 0$, PSNR = 18.44 dB, CIM = 3.47

(c) Prioritize all of the ROI with nominal background distortion weight. $\alpha_{min} = 0.02$, PSNR = 21.74 dB, CIM = 3.44

(d) Prioritize only the face and hands. $\alpha_{min} = 0.1$, PSNR = 23.43 dB, CIM = 3.41

(e) Prioritize only the face. $\alpha_{min} = 0.5$, PSNR = 25.21 dB, CIM = 3.32

(f) Quality optimized. $\alpha_{min} = 1.6$, PSNR = 25.73 dB, CIM = 3.23

Figure 2. Comparison of distortions for different levels of region-of-interest (ROI) priority each at 55 kbps. The encoding option α_{min} specifies the minimum distortion weight to be applied to any region. As α_{min} increases, the torso, hands, and face are allocated fewer additional bits relative to the rest of the frame, causing a decrease in intelligibility. Figure 1 specifies the relationship between the CIM and the predicted subjective intelligibility ratings.

display resolution of 320×240 pixels.[4] This reduced resolution is also required for the simultaneous presentation used in the paired comparison methodology.[5] The videos are temporally subsampled to 15 frames per second, which is above the nominal frame rate required for ASL communication.[6]

Three reference stories are selected for the experiment and encoded at one of three bitrates: 20 kbps, 45 kbps, and 80kbps. Each story is encoded at a single bitrate using five different values of α_{min}: 0, 0.02, 0.1, 0.5, and 1.6, corresponding to the five ROI prioritization scenarios illustrated in Figure 2. This combination of bitrates and α_{min} values are selected to yield videos that would be rated as difficult to understand (20 kbps), from neutral to easy (45 kbps), and from easy to very easy (80 kbps), as illustrated in Figure 1(b).

3.2. Method

The subjective experiment uses a paired comparison methodology with simultaneous presentation, as recommended by ITU-T.[5] Each presentation consists of a pair of coded ASL videos displayed synchronously and side-by-side on a single screen. After watching the video pair, the participant is asked to "please select the video you would prefer to see on a cell phone video call." The collection of video pairs consist of videos generated from the same reference story encoded using two different values of α_{min}.

At each bitrate, the 5 test levels of α_{min} yield 10 pair-wise combinations. The 10 pairs are presented to the participant twice, swapping the left/right display order. None of the test pairs contain videos at different bitrates, assuming that videos at higher bitrates will always be preferred over videos at lower bitrates. This results in 20 paired comparisons per bitrate and 60 comparisons per participant. Following 2 practice examples, the 60 pairs are presented in random order. At the completion of the paired comparisons, participants provide demographic data regarding their level of experience with ASL, their use of video-based communication tools such as video relay services and video phones, and their use of text-based communication tools such as Internet chat and text messaging.

3.3. Implementation

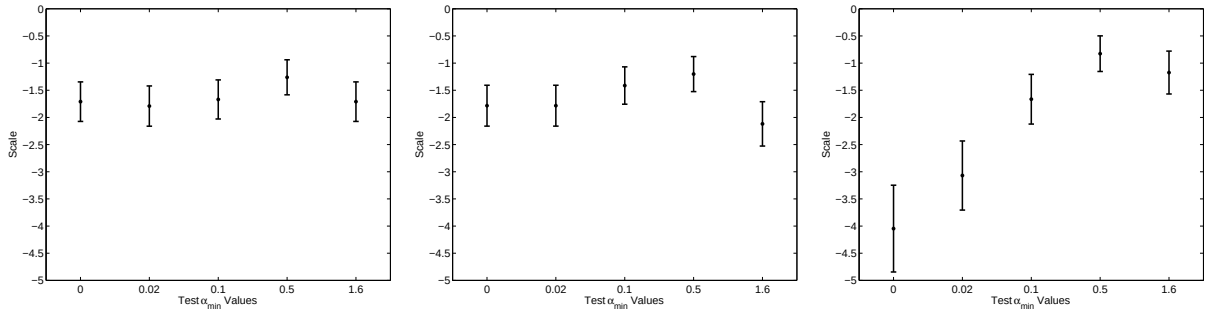
Because of the difficulties in recruiting participants who are fluent in ASL, two versions of the experiment were made available: an on-site experiment in a controlled environment at Cornell University and a web-based experiment, in which ASL users in any location could participate. Despite the limitations of web-based perceptual experiments, such as uncontrolled display environments, varying display technologies, and other real-world variability, web-based experiments drastically increase the observer pool and typically provide results that are consistent with lab-based experiments.[7,8]

To guarantee synchronous playback of the video pairs, the on-site experiment was implemented in Matlab, using the Psychophysics Toolbox,[9–11] which offers extremely precise control over the video playback timing. For the web-based experiment, an individual video file was created for each pair by decoding the compressed videos, horizontally concatenating the decoded frames, and re-encoding the side-by-side video at a sufficiently high bitrate such that no new compression artifacts were introduced. The video pairs in both the on-site and web-based experiments were identical, though the web-based version offered a shortened experiment, wherein participants only viewed each pair once, without evaluating the left/right swapped pair. Pairs used in the shortened experiment were selected such that every 2 participants evaluated exactly the same set of pairs as a single participant in the full-length experiment. A demonstration of the web-based experiment is available online.*

3.4. Data Processing

The paired comparison methodology acquires data to estimate the probability that stimulus i is preferred over stimulus j. The Bradley-Terry model provides a framework for mapping the pair-wise probability estimates of preference to scale values for each stimulus.[12] The scale values rank the collection of stimuli, determining the relative preference of each value of α_{min}. Because the stimulus pairs in the experiment never contain videos at two different bitrates, scale values are generated independently at each of the three tested bitrates.

*http://foulard.ece.cornell.edu/ASLweb/demo/



(a) 20 kbps. (b) 45 kbps. (c) 80 kbps.

Figure 3. Scale values generated from the complete set of paired comparison data using the Bradley-Terry model. Error bars indicate the 95% confidence intervals.

Table 1. Table of p-values for χ^2_4 hypothesis test on the uniformity of the scale values,[13] for different groups of participants. The null hypothesis indicates that the scale values are not statistically different from a uniform distribution, i.e., each α_{min} is equally preferable. Entries in bold indicate that the null is rejected at 95% confidence ($p < 0.05$). The "ASL FL" and "ASL SL" groups correspond to participants for whom ASL is their first language (FL) or second language (SL). The "Heavy Video Use" and "Light Video Use" groups are divided according to their level of experience with video-based communication technologies.

Bitrate	Complete Set	ASL FL	ASL SL	Heavy Video Use	Light Video Use
20 kbps	0.370	0.097	0.084	**1.7e-4**	**0.003**
45 kbps	**0.017**	0.261	**0.022**	**0.003**	**3.9e-4**
80 kbps	**0**	**8.9e-14**	**4.0e-8**	**0**	**0.003**

4. RESULTS AND DISCUSSION

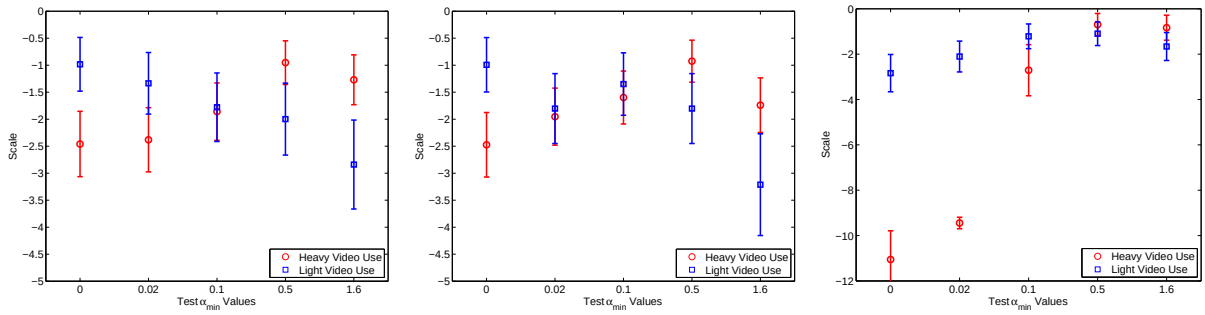
A total of 12 ASL users participated in this experiment: 3 on-site participants and 9 web-based participants. Of the 9 web-based participants, 4 opted for the shortened version, yielding a total of 600 comparisons (200 at each bitrate).

Applying the Bradley-Terry model,[12] scale values for each tested α_{min} are computed at each bitrate. Following the methodology discussed in Ref.13, a χ^2_{t-1} hypothesis test with $t-1$ degrees of freedom ($t = 5$ levels of α_{min}) determines whether the scale values are statistically different from a uniform distribution. If the null hypothesis holds, all values of α_{min} are equally preferable. If the null hypothesis is rejected, at least one α_{min} is preferred over the others. The computed scale values, with 95% confidence intervals, are provided in Figure 3. Table 1 provides the results of the hypothesis tests for uniformity.

At 80 kbps, the scale values demonstrate a preference when $\alpha_{min} \geq 0.1$, as plotted in Figure 3(c). Each of the scale values for $\alpha_{min} \geq 0.1$ have overlapping confidence intervals and can be considered equally preferable. At $\alpha_{min} = 0.1$, because of the relatively high encoding bitrate, the quality-intelligibility optimized coder produces video predicted to be very easy to understand, as seen in Figure 1(b). In this case, the smaller values of $\alpha_{min} = 0$ and $\alpha_{min} = 0.02$ significantly reduce the overall quality (PSNR) while providing only negligible improvements in intelligibility (CIM). This saturation effect implies that when coding an ASL video, when the bitrate is sufficiently high for producing video considered very easy to understand, any additional rate must be allocated to maximize a quality constraint.

At 45 kbps, $\alpha_{min} = 0.1$ and $\alpha_{min} = 0.5$ are preferred over $\alpha_{min} = 1.6$. Referring to Figure 1(b), these two values of α_{min} correspond to the points on the PSNR-CIM curve having the largest slope. These points are preferred because they provide the largest increase in the CIM for the corresponding decrease in PSNR.

At 20 kbps, the scale values are not statistically different from a uniform distribution, indicated by the hypothesis test results in Table 1. As illustrated in Figure 1(b), the PSNR-CIM curve at this bitrate is relatively flat; the relative change in the CIM is small compared to the relative change in PSNR, for varying α_{min}. One



(a) 20 kbps. (b) 45 kbps. (c) 80 kbps.

Figure 4. Scale values generated from paired comparison data of groups of participants who use both video relay services and video phone technology (denoted "heavy video use") and those who do not (denoted "light video use"). Scale values are generated according to the Bradley-Terry model. Error bars indicate the 95% confidence intervals.

Table 2. Table of p-values for χ^2_4 hypothesis test on differences between groups.[13] The null hypothesis indicates that the scale values from each group are statistically equivalent. Entries in bold indicate that the null is rejected at 95% confidence ($p < 0.05$), i.e., the groups are statistically different from each other. The "ASL FL" vs "ASL SL" column compares groups that correspond to participants for whom ASL is their first language (FL) or second language (SL). The "Heavy Video Use" vs "Light Video Use" column compares groups that are divided according to their level of experience with video-based communication technologies.

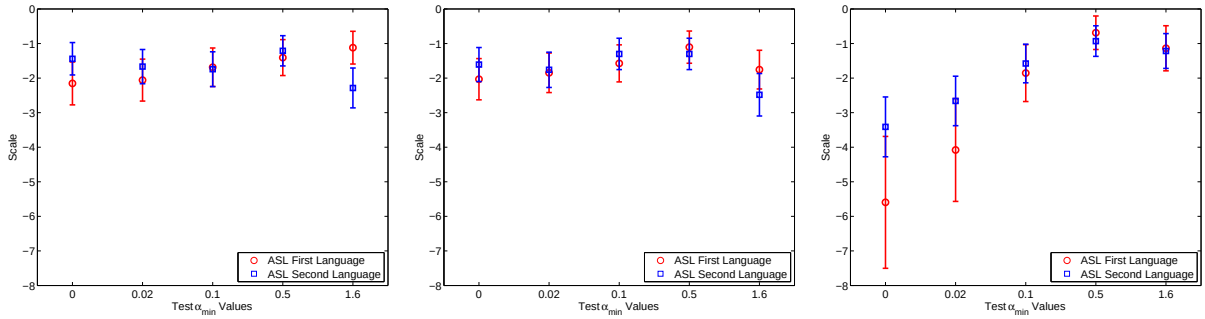
Bitrate	ASL FL vs ASL SL	Heavy Video Use vs Light Video Use
20 kbps	**0.019**	**7.1e-7**
45 kbps	0.321	**5.4e-5**
80 kbps	0.186	**2.9e-7**

might expect a preference for the highest quality video, when the change in CIM is small. However, the lowest quality video ($\alpha_{min} = 0$) is still equally preferable to the highest quality video ($\alpha_{min} = 1.6$).

A uniform distribution of scale values can be attributed to one of two statistical models. In the first model, each individual observer has no preference and is arbitrarily selecting one of the two videos in a pair. This case implies that every value of α_{min} yields the same perceptual response and no value is preferred over another. In this case, the selection of an operating point in the quality-intelligibility trade-off is arbitrary, since all points are truly equal. In the second model, a single observer (or group of observers) demonstrates a preference for a particular α_{min}, while a sampling of the entire population of observers exhibits no preference. In this case, each value of α_{min} is preferred by a specific individual (or group) and that preference varies across individuals (or groups), supporting the need for a user-specified operating point in the quality-intelligibility trade-off.

An analysis of the scale values for different groups of participants provides evidence for the second model. In particular, groups divided according to their use of video-based communication technologies have opposite (and non-uniform) preference rankings. Because the collection of data is sufficiently small, the relevant groups have been identified manually, though one could use a recursive procedure for identifying groups having homogeneous preferences.[14] The 7 participants who reported using video relay services and video phone technology are denoted the "heavy video use" group. The remaining 5 participants are denoted the "light video use" group, because some individuals in this group use Internet chat services, such as Skype, which offer video communication as a secondary feature. At every bitrate, the scale values for each of the two groups are statistically different from uniform, as shown in Table 1. Furthermore, using the methods in Ref. 13, a χ^2_4 hypothesis test identifies a significant difference between these two groups at every bitrate, i.e., these two groups have statistically different preferences. The results of this hypothesis test, with p-values, are provided in Table 2.

At each tested bitrate, the "light video use" group has a significantly higher preference for $\alpha_{min} = 0$ than the "heavy video use" group. Furthermore, at 25 kbps and 50 kbps, the "light video use" group prefers $\alpha_{min} = 0$



(a) 20 kbps. (b) 45 kbps. (c) 80 kbps.

Figure 5. Scale values generated from paired comparison data of participants whose first language is ASL or whose first language is not ASL. Scale values are generated from paired comparison data according to the Bradley-Terry model. Error bars indicate the 95% confidence intervals.

over $\alpha_{min} = 1.6$, as shown in Figure 4. Conversely, the "heavy video use" group demonstrates a preference for $\alpha_{min} = 1.6$, where videos are coded for quality. This preference is most evident at 80 kbps, where the values of $\alpha_{min} \geq 0.1$ are preferred unanimously over $\alpha_{min} = 0$ and $\alpha_{min} = 0.02$, causing the large difference in scale values in Figure 4(c).

Variations in the preferences of the "heavy video use" and "light video use" groups may be attributable to differences in their prior experience of digital video. Video-based communication technologies typically use a quality criteria when coding video (e.g., they maximize PSNR). In this case, the coding distortions are generally distributed evenly across space. The strictly intelligibility optimized coder ($\alpha_{min} = 0$) produces video in which the signer and the background have significantly different distortion levels. This disparity in the spatial distribution of distortion substantially differs from a quality optimized coder, and, consequently, differs from the prior experiences of the "heavy video use" group, resulting in a preference for the coded ASL video that is more consistent with their expectations.

An alternative grouping of ASL users divides the collection based on the level of experience with ASL. The first group consists of those whose first or primary language is ASL, which commonly includes deaf persons or hearing children of deaf adults. The second group consists of those who have learned ASL as a second language. The differences between these groups are only significant at 20 kbps and not to the same degree of confidence as the differences for the "video use" groups. The p-values are summarized in Table 2. Furthermore, the scale values for each of these groups are consistent with those computed from the complete data set, as illustrated in Figure 5. Other partitions of the participants yield similar conclusions; a user's experience with video-based communication serves as the most meaningful predictor of the preferred operating point in the quality-intelligibility trade-off.

5. CONCLUSIONS AND FUTURE WORK

A paired comparison experiment was conducted to evaluate user preferences for coded ASL video. Test videos were generated using a quality-intelligibility coder, which provides a user-controlled parameter, α_{min}, that varies the degree to which a quality criteria is emphasized over an intelligibility criteria. Videos at 3 bitrates were coded using 5 test levels for α_{min}. At 80 kbps, users preferred videos coded according to the quality criteria, because the intelligibility of these videos was sufficiently high. At the lower tested bitrates of 45 kbps and 20 kbps, the preferences varied with user demographics. Participants having significant experience using video-based communication technologies preferred video coded according to the quality criteria while those with little experience preferred video coded according to the intelligibility criteria. The existence of these two classes of individuals confirms the need for a user-centric encoding option, because the most desirable quality-intelligibility operating points vary across individuals and across bitrates.

The parameter α_{min} varies the spatial distribution of the distortions, either distributing distortion evenly across a video frame (in the quality case) or heavily distorting the background region in order to improve the quality in only the signer (in the intelligibility optimized case). For future study, these region-of-interest (ROI)

coded videos will be evaluated by participants who are unfamiliar with ASL. Without any ASL experience, a participant making paired comparisons in this setting will simply prefer the video having higher quality, independent of the ASL content. The impact on the perceived video quality of ROI video of this type can potentially be applied more generally to any videoconferencing scenario.

REFERENCES

1. F. Ciaramello and S. Hemami, "An objective intelligibility measure for assessment and compression of American Sign Language video," *IEEE Trans. Image Proc.* , submitted for publication.
2. F. Ciaramello, J. Ko, and S. Hemami, "Quality versus intelligibility: Evaluating the coding trade-offs for American Sign Language video," *Proc. Information Sciences and Systems (CISS)* , Mar. 2010.
3. F. Ciaramello and S. Hemami, "The influence of space and time varying distortions on objective intelligibility estimators for region-of-interest video," in *Proc. IEEE Int. Conf. Image Proc.*, 2010.
4. J. Chon, N. Cherniavsky, E. Riskin, and R. Ladner, "Enabling access through real-time sign language communication over cell phones," *43rd Annual Asilomar Conference on Signals, Systems, and Computers* **19**(1), 2009.
5. ITU, *P.910: Subjective video quality assessment methods for multimedia applications*, September 1999.
6. J. Harkins, A. Wolff, E. Korres, R. Foulds, and S. Galuska, "Intelligibility experiments with a feature extraction system designed to simulate a low-bandwidth video telephone for deaf people," in *Proc. RESNA Annual Conference*, **14**, pp. 38–40, 1991.
7. N. Moroney, "Unconstrained web-based color naming experiment," in *Proc. SPIE Color Imaging: Device-Dependent Color, Color Hardcopy and Graphic Arts*, 2003.
8. M. H. Birnbaum, ed., *Psychological Experiments on the Internet*, Academic Press, San Diego, CA, 2000.
9. D. H. Brainard, "The psychophysics toolbox," *Spatial Vision* **10**, pp. 433–436, 1997.
10. D. G. Pelli, "The videotoolbox software for visual psychophysics: Transforming numbers into movies," *Spatial Vision* **10**, pp. 437–442, 1997.
11. M. Kleiner, D. Brainard, and D. Pelli, "What's new in psychtoolbox-3?," *Perception* **36**(ECVP Abstract Supplement), 2007.
12. R. A. Bradley and M. E. Terry, "The rank analysis of incomplete block designs i: The method of paired comparisons," *Biometrika* **39**, pp. 324–345, 1952.
13. J. C. Handley, "Comparative analysis of bradley-terry and thurstone-mosteller paired comparison models for image quality assessment," in *PICS 2001: Image Processing, Image Quality, Image Capture, Systems Conference*, pp. 108–112, 2001.
14. C. Strobl, F. Wickelmaier, and E. Karls, "Accounting for individual differences in bradley-terry models by means of recursive partitioning." Technical Report Number 54, University of Munich, 2009.