Quantifying the Effect of Disruptions to Temporal Coherence on the Intelligibility of Compressed American Sign Language Video

Frank M. Ciaramello and Sheila S. Hemami

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

ABSTRACT

Communication of American Sign Language (ASL) over mobile phones would be very beneficial to the Deaf community. ASL video encoded to achieve the rates provided by current cellular networks must be heavily compressed and appropriate assessment techniques are required to analyze the intelligibility of the compressed video. As an extension to a purely spatial measure of intelligibility, this paper quantifies the effect of temporal compression artifacts on sign language intelligibility. These artifacts can be the result of motion-compensation errors that distract the observer or frame rate reductions. They reduce the the perception of smooth motion and disrupt the temporal coherence of the video. Motion-compensation errors that affect temporal coherence are identified by measuring the block-level correlation between co-located macroblocks in adjacent frames. The impact of frame rate reductions was quantified through experimental testing. A subjective study was performed in which fluent ASL participants rated the intelligibility of sequences encoded at a range of 5 different frame rates and with 3 different levels of distortion. The subjective data is used to parameterize an objective intelligibility measure which is highly correlated with subjective ratings at multiple frame rates.

1. INTRODUCTION

Real-time, two-way transmission of American Sign Language (ASL) video over cellular networks has the potential to significantly benefit members of the Deaf community. Unfortunately, current generation US networks are limited in their capacity. Current GPRS networks provide as little as 15 kbps uplink and 30 kbps downlink. At such low bitrates, ASL videos encoded using techniques designed to maximize fidelity yield sign language sequences that are unintelligible.[1,2] Objective assessment techniques are required to evaluate the utility of compressed ASL video in terms of the intelligibility of the sequence.

Understanding how information is communicated in American Sign Language is essential for informing objective techniques that will assess intelligibility. In his pioneering work, Stokoe[3] laid the foundation for the linguistic structure of sign language. He identified three basic units, called primes, that form a sign: the tab (where the sign occurs), the dez (the configuration of the hands), and the sig (the movement that occurs).

This work was expanded by Liddell and Johnson[4] to include the notion of a coherent sequence of the basic units as a fundamental element in sign linguistics, more commonly known as the hold-movement-hold model. In this model, a sign is characterized by an initial articulation of the hands (tab and dez), a movement period during which either the handshape and/or the location change, and a final hold of the new articulation. Not all movements are linguistically relevant. Liddell and Johnson define *movement epenthesis* as the transition from the final hold of one sign to the initial hold of the following sign. The movement epenthesis contains no linguistic information but is simply necessary to physically position the hands correctly.

In addition to manual gestures, a substantial amount of contextual information is added to a conversation by using nonmanual signs in the face.[5,6] A signer's gaze can indicate pronomial references or quotations. Raising or furrowing one's eyebrows indicates a question or a negation, respectively. Nonmanual signs are also essential

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

for forming relative clauses. The contextual detail added through nonmanual signs suggests that accurate interpretation of facial expression is essential for understanding ASL.

The linguistic information in ASL is contained spatially in the hand configurations and facial expressions and temporally in the linguistically relevant movements, such as the changing of a handshape or location. Quantifying the effect of spatial distortions in the signer's face and hands has been addressed in the authors' previous work.[7] This work aims to quantify the effect of temporal artifacts on sign language intelligibility.

Temporal artifacts can reduce the perceived temporal coherence of the video. Temporal coherence in a video sequence is the consistency of the objects between frames that allows for the perception of smooth motion. Evidence suggests that the human visual system (HVS) extrapolates the current visual stimulus to predict location of objects in the next perceived moment.[8] In compressed digital video, distortions can cause inconsistencies between the perceived motion and the expectations of the HVS. For example, motion compensated video encoders can create prediction errors in which a moving object leaves a trail of residual pieces of itself as it moves across the frame. This results in spatial distortions which are temporally correlated with relevant objects in the video.

Viewers track relevant, moving objects and are more sensitive to distortions in and around these objects.[9] These types of distortions reduce the observer's ability to identify coherent motion and can affect the perceived quality of the video. If the distortions are in regions important for ASL communication (e.g. a signer's hands), then the overall intelligibility may also be reduced. Seshadrinathan et al. use an estimate of the optical flow of a video sequence to model such temporal distortions and demonstrate improvements in video quality assessment.[10]

Reductions in frame rate can also disrupt temporal coherence. In ASL video, frame rate reductions can obscure the linguistically relevant movements and will significantly affect intelligibility. Several studies have been performed to understand the effect of reduced frame rate on ASL comprehension. Parish et al. described an ASL sequence as a series of *events*.[11] For example, in an ASL sequence, an event may be moving a signing hand from one position in space to another. Parish generated reduced frame rate sequences of individual signs that were either subsampled at a constant rate or subsampled by selecting frames corresponding to event boundaries. This work demonstrated that by preserving event boundaries, observers were more likely to identify signs correctly. Harkins et al. also found that for individual signs, reductions in frame rate had little impact on the observer's intelligibility; recognition accuracy was greater than 91% at 6 fps.[12] However, these intelligibility experiments were performed for individual signs.

In the same study, Harkins et al. found that for fully formed sentences, recognition accuracy dropped significantly at frame rates below 10 fps. Cherniavsky et al. performed a study that varied the frame rate of fingerspelling segments versus signing segments of an ASL conversation.[13] Fingerspelling in ASL occurs when each letter of a word is explicitly spelled and is typically used for proper nouns and other words with no associated sign. In this study, periods of signing were presented at frame rates of 5, 10, and 15 fps. Periods of fingerspelling were presented with increased frame rates, e.g. when signing was 5 fps, fingerspelling was either 5, 10, or 15 fps. The results suggest that the overall intelligibility of a sequence depends more on the frame rate during signing than during fingerspelling.

The previous work demonstrates that conversational ASL is less robust to reductions in temporal resolution than individual signs. This implies that relevant events contained in subtle facial expressions are no longer perceptible at low frame rates. Hooper et al. analyzed the effects of frame rate on learner comprehension and drew similar conclusions.[14] In a full sign language conversation, as frame rate is decreased, contextual information in facial expressions tends to be lost more quickly than the information contained in larger hand movements.

This work aims to quantify the reduction in intelligibility caused by disruptions to the temporal coherence of sign language in two ways. Spatial distortions that inhibit an observer's ability to track the relevant motion in ASL video are identified and incorporated into an objective intelligibility measure. A subjective experiment is performed to quantify the relationship between intelligibility and frame rate reductions for conversational ASL video and to determine if this relationship is consistent for multiple levels of spatial fidelity. The results of this subjective experiment are incorporated into the objective intelligibility measure, resulting in a single measure which unifies temporal disruptions caused by both compression artifacts and frame rate reductions. This objective measure exhibits high correlation with experimental subjective intelligibility data.

This paper is organized as follows. Section 2 describes an intelligibility measure developed in the authors' previous work, which is based only on spatial distortions in the signer's face and hands. Section 3 presents a method for identifying and quantifying motion-compensated compression artifacts that reduce the appearance of smooth, coherent motion and improving the intelligibility measure. Section 4 details the experiment performed and presents a model for the impact of frame rate on intelligibility.

2. A SPATIAL INTELLIGIBILITY MEASURE

The intelligibility of ASL depends both on the spatial fidelity of the signer's face and hands and on the accurate representation of the linguistically relevant motion. In a previous study, an objective measure of intelligibility was developed based solely on the quality in the signer's face and hands.[7]

The sign language videos used in this previous study were recorded at a resolution of 320×240 pixels and a frame rate of 30 fps. The video coding was done using x264, an open-source, standards-compliant implementation of the H.264/AVC codec. Three different coding parameters were adjusted to create the compressed video: bitrate, frame rate, and region-of-interest (ROI) rate allocation. Three bitrates (15 kbps, 20 kbps, and 25 kbps) and two frame rates (10 fps and 15 fps) were selected for this study.

Taking into account the importance of details in a signer's face, the ROI rate allocation scheme allows for an increase in the signal fidelity of that region. A fixed region was defined for an entire video sequence around the signer's face, and that region is coded using a lower quantization parameter (QP). For the study, three QP offsets were applied: 0, -6, and -12. Because the sequences were encoded at a fixed bitrate, there is a trade-off between the bits allocated to the face region and the bits allocated to the rest of the frame; as fidelity around the face increases, the fidelity in the rest of the frame decreases.

It is important to note that the test sequences used in this study did not exhibit substantial disruptions in temporal coherence. While two frame rates were evaluated, they were treated as separate data sets and the spatial intelligibility measure was evaluated on each set independently. Furthermore, the ROI coding scheme scheme improves the quality around the face but still allocates sufficient rate to other non-face regions.

At a fixed frame rate and in the absence of severe temporal artifacts, subjective intelligibility can be predicted from the objective intelligibility measure as described in Equations 1 and 2, where $D_I^{spatial}$ is the weighted sum of mean-squared-error in the face and hands for an individual frame and $I_{spatial}$ is the objective intelligibility of an entire sequence, averaged over N frames. The face and hand pixels in each frame are identified by applying a classifier cascade for face detection[15] and skin color segmentation for hand detection.[7]

$$D_I^{spatial} = W_F D_F + W_H D_H \tag{1}$$

$$I_{spatial} = \frac{1}{N} \sum_{n=1}^{N} 10 \log_{10} \frac{255^2}{D_I^{spatial}(n)} \tag{2}$$

The optimal weights, in terms of correlation with subjective data, are $W_F = 0.6$ and $W_F = 0.4$. These values are consistent with the linguistic structure of ASL. Facial details add a significant amount of information to the conversation and ASL observers fixate on this region. Distortions in the face have a greater impact on the overall intelligibility. However, because this measure is based purely on the spatial distortions in the signer's face and hands, it cannot account for compression artifacts that reduce temporal coherence, e.g. temporally correlated distortions or reductions in frame rate. In the following two sections, this intelligibility measure will be extended to incorporate the effects of these temporal artifacts.

3. INCORPORATING TEMPORAL COHERENCE INTO AN OBJECTIVE INTELLIGIBILITY MEASURE

The perception of coherent motion in sign language video can be hindered by certain types of spatial compression artifacts. These artifacts are a consequence of applying the purely spatial measure of intelligibility to a motion-compensated video encoder. The distortion model in Equation 1 was applied to a rate-distortion optimization procedure within the H.264 standard.[1] In this encoder, the macroblocks of each frame in an input ASL sequence are labeled as face, hand, or background. Because the distortion measure in Equation 1 does not include any background distortions, macroblocks labeled as background will be encoded at the lowest possible rate. As a result, background macroblocks will have very large distortions when compared to face and hand macroblocks.

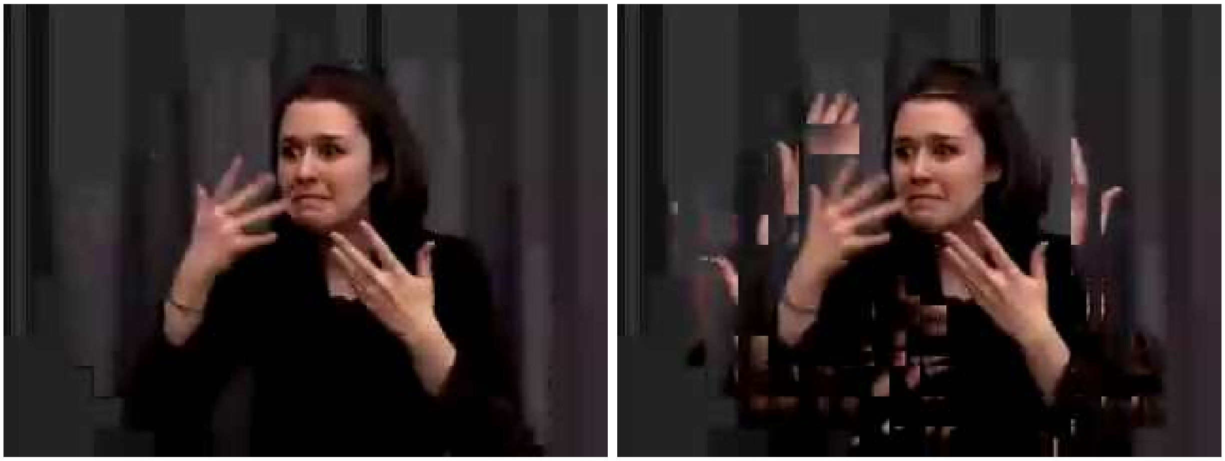
Modern video coders, such as MPEG-2 and H.26x, use motion compensation to exploit temporal redundancies. To achieve very low rates, coded macroblocks (blocks of 16×16 pixels) in a frame can simply be skipped, i.e., not coded at all. The co-located macroblock in the previous frame is copied to the current frame. In the sign language optimized encoder, nearly all background macroblocks are skipped in order to encode them with as little rate as possible. However, if a particular macroblock contains a face or hand in one frame and contains only background in the next frame, encoding the background in this way creates distortions that negatively affect the temporal coherence of a sequence. When these background macroblocks are skipped, residual pieces of the face and hand remain in the macroblock. An example of a frame with many residual face and hand macroblocks is provided in Figure 1(b). These residuals will remain in the frame until the macroblock is once again labeled as face or hand. Fluent ASL observers commented that these types of artifacts made it difficult to follow the hand trajectory and to focus on the signs. Because these distorted macroblocks are temporally correlated with relevant objects, they interrupt the temporal coherence of relevant motion in the ASL sequence.

The intelligibility measure in Equation 1 is unable to account for these problems because it does not measure distortions in background blocks. Properly integrating a temporal factor into the intelligibility measure presents a challenge. The distortions in background blocks that particularly disrupt temporal coherence must be accounted for without over-emphasizing the importance of distortions in less relevant regions of the frame. This is done by identifying only those blocks that contain residual face and hand artifacts.

In addition to face, hand, and background macroblocks, a fourth macroblock type is defined. A co-located block that contains a face or hand in frame $n-1$ and contains only background in frame n is labeled as NewBG. Unlike the number of face and hand blocks, the number of NewBG blocks in a given frame will vary with frame rate. At lower frame rates, the time difference between frames is larger and the hand moves a farther distance. As a result, there will be more NewBG blocks in each frame. For example, at 60 fps, the NewBG blocks occupy only 1% of the frame on average while at 6 fps they occupy 6% of the frame. If these NewBG blocks are not allocated sufficient rate, they will contain substantial artifacts.

The reduction in intelligibility caused by these artifacts is quantified by measuring the distortion only in NewBG blocks which contain residual face and hand artifacts. NewBG blocks that do not contain face and hand artifacts are treated as all the other background blocks. Distortions in these blocks do not contribute to the overall intelligibility value. In order to differentiate these two cases, the blockwise correlation coefficient is computed for each NewBG block. The correlation is measured between the two co-located blocks in the compressed frames n and $n-1$. If the new background block was coded, the correlation will be low. If the current block was simply copied from the previous frame and contains residual hand or face artifacts, the correlation will be close to 1. The H.264 encoding standard applies a deblocking filter to the coded video at macroblock boundaries. If the deblocking filter is not used, the correlation between skipped, co-located blocks will be exactly equal to 1. It was empirically verified that a threshold of 0.9 was able to account for the changes caused by the deblocking filter. The correlation coefficient is used to differentiate between NewBG blocks that were skipped and NewBG blocks that were coded.

The distortions that impact temporal coherence are quantified by $D_I^{temporal}$, as in Equation 3, where D_{NewBG_F} and D_{NewBG_H} are the mean-squared-error in NewBG blocks that were previously face or hand blocks, which have a blockwise correlation greater than 0.9. W_F and W_H are the same weights that are applied in Equation 1. As previously described, these weights are supported by the structure of ASL and distortions around the signer's face will have a greater perceived impact. This yields a new measure of objective intelligibility which includes the correlated distortions in background blocks, as described in Equations 4 and 5.



(a) New background blocks coded. I = 31.8 dB (b) Background blocks not coded. I = 26.6 dB

Figure 1. Comparison of three types of background qualities. Intelligibility is unaffected when background distortions are small or are uncorrelated with co-located blocks in the previous frame. Intelligibility is reduced when compression artifacts disrupt the temporal coherence of motion.

$$D_I^{temporal} = W_F D_{NewBG_F} + W_H D_{NewBG_H} \tag{3}$$

$$D_I = D_I^{spatial} + D_I^{temporal} \tag{4}$$

$$I = \frac{1}{N} \sum_{n=1}^{N} 10 \log_{10} \frac{255^2}{D_I(n)} \tag{5}$$

Intuitively, new background blocks with high correlation will have a stronger impact on overall intelligibility. Blocks with large distortions but very low correlation will not effect intelligibility, allowing for large distortions in the background as long as they are not caused by residual pieces of the face or hands. This is illustrated in Figure 1. Figure 1(a) has both low distortion and low correlation in new background blocks. Figure 1(b) has very high distortion and high correlation, resulting in a large reduction in intelligibility.

4. QUANTIFYING INTELLIGIBILITY REDUCTIONS CAUSED BY REDUCED FRAME RATE

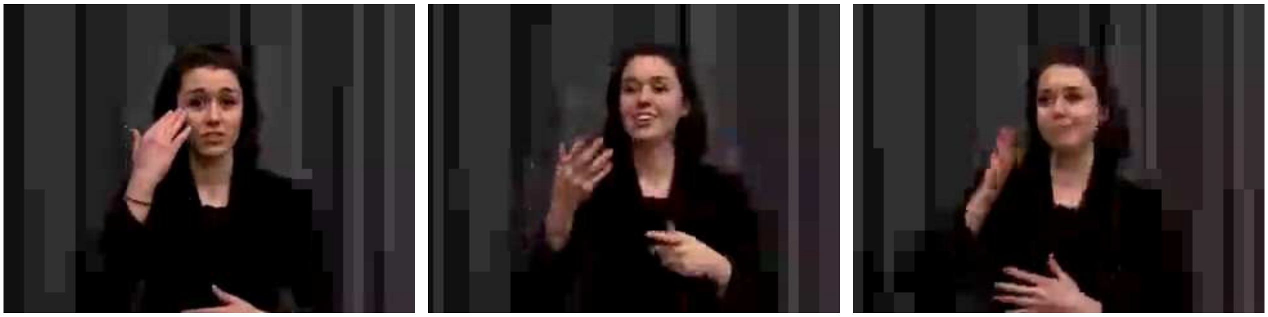
In addition to spatial compression artifacts that affect the perception of motion, reductions in the overall frame rate of a video sequence can have a significant impact on intelligibility. A study was performed in order to quantify how changes in temporal resolution affect an observer's ability to understand sign language stories and to determine if these changes are consistent at different bitrates. The results of this study are used to parameterize a model for the change in intelligibility as a function of frame rate. This model is verified on a separate test data set.

4.1 Experiment Methodology

A study was designed to quantify subjective intelligibility of ASL videos at varying levels of quality and at a range of frame rates. Videos were displayed to participants on an HTC Apache pocket PC with a screen size of 2.8" diagonally and resolution of 240×320 pixels, in order to simulate a mobile sign language conversation. The display on the mobile device was set to the maximum brightness. Participants were given permission to hold the device or set it on a table and were not constrained to a fixed viewing distance. A total of 18 individuals participated in this study. All the participants were fluent in ASL.

Table 1. Average bitrate at each combination of encoding parameters. Because the objective intelligibility measure of each frame was held constant, increasing the frame rate results in an increase in bitrate.

	Frame Rates				
Intelligibility	**6**	**7.5**	**10**	**15**	**20**
32 dB (High)	19.8 kbps	23.1 kbps	27.8 kbps	35.8 kbps	42.3 kbps
28 dB (Medium)	10.6 kbps	12.4 kbps	15.1 kbps	19.5 kbps	23.1 kbps
26 dB (Low)	7.8 kbps	9.1 kbps	11.1 kbps	14.4 kbps	17.1 kbps



(a) High Intelligibility: 32 dB (b) Medium Intelligibility: 28 dB (c) Low Intelligibility: 26 dB

Figure 2. Sample frame taken from videos at each of the three levels of objective intelligibility.

A collection of 16 stories were filmed at the University of Washington at a resolution of 1280×720 pixels and a frame rate of 60 progressive fps. The videos were cropped and downsampled in order to match the resolution of the testing device. On average, the stories were approximately 50 seconds long. All the stories were filmed and displayed in color.

The videos used in the study were encoded according to the H.264 standard using a modified version of the x264 open source encoder. The encoder was modified to use Equation 5 within a rate-distortion optimization setting. The encoder selects quantization step sizes and encoding modes for each macroblock in order to maximize objective intelligibility.[1]

The videos were encoded at three different values of spatial intelligibility, as defined in Equation 5. Specifically, values of 32, 28, and 26 dB were used in the study. An objective score of 32 dB corresponds to a video that would be rated as easy to understand. A score of 26 dB corresponds to a video that would be difficult to understand. In order to accurately quantify the effect of frame rate on intelligibility, the intelligibility measure was held constant for each frame in a sequence, regardless of the encoding frame rate. Therefore, changes in subjective ratings can be attributed solely to changes in frame rate. A multi-pass encoding approach was used to ensure that each frame in the video sequence had the same objective intelligibility value. As a result, the videos are encoded at a variable bit rate rather than a constant bit rate.

For this study, 5 different frame rates were selected: 20, 15, 10, 7.5, and 6 fps. All the frame rates chosen were integer fractions of the original 60 fps recordings: 3, 4, 6, 8, and 10 respectively. The highest possible frame rate was limited by the playback capabilities of the testing device. The cell phone used for the study could not consistently play videos above 20 fps without dropping frames. As demonstrated by both Foulds[16] and Harkins et al.,[12] individual signs can be identified with high accuracy at frame rates as low as 6 fps.

Each video was prepared at all 5 frame rates and encoded at one of the three levels of intelligibility. Videos were assigned to each level such that an observer's total viewing time for each level of intelligibility was approximately the same. The average bitrate of the sequences at each combination of encoding parameters is given in Table 1. Figure 2 provides a sample frame taken from a video at each level of intelligibility and illustrates the visual differences in quality.

During the study, participants first watched two training videos representative of the highest and lowest quality they would see throughout the study. Participants then watched a unique video at each combination of

Table 2. A measure of the goodness of the fit (R^2) for each frame rate and the parameters corresponding to the linear fits. R^2 is relatively unaffected by fixing the slope and fitting only over the intercepts. Fixing the slope provides a consistent model of subjective versus objective intelligibility at every frame rate.

	Frames per Second	**6**	**7.5**	**10**	**15**	**20**
Model parameters and correlation coefficient for independent linear fits at each frame rate	R^2	0.86	0.78	0.67	0.63	0.77
	Slope (m)	0.210	0.308	0.264	0.287	0.292
	Intercept (b)	-6.53	-9.15	-7.42	-8.00	-7.97
Model parameters and correlation coefficient for linear fits with slope fixed across frame rates	R^2	0.79	0.77	0.67	0.63	0.77
	Fixed Slope (m)	0.272	0.272	0.272	0.272	0.272
	Intercept (b)	-8.28	-8.11	-7.64	-7.61	-7.40

frame rate and spatial intelligibility, resulting in a total of 17 observations. Following each video, participants were asked "How easy or how difficult was it to understand the video?" and responded on a 5-point scale.

4.2 Modeling Objective Intelligibility as a Function of Frame Rate

Individual participants' responses are converted to z-scores and the linear fit between subjective responses and the objective intelligibility measure from Equation 5 is calculated for each frame rate independently. The average linear correlation across the frame rates is 0.86. The value of R^2 (which is interpreted as the goodness of the fit) and the model parameters of the linear fits are summarized in Table 2.

The slope of the fits is relatively consistent at each frame rate. Fixing the slope to be the mean of the slopes of each individual fit (i.e., 0.272) and performing the regression only by varying the intercept results in very little reduction in the goodness of the fit. Since fixing the slope has little impact on the correlation, it is fair to assume that a fixed slope model is equally plausible. Such a model is intuitively pleasing, as it implies that the relationship between the objective intelligibility measure and subjective intelligibility ratings is consistent across frame rates. The loss of linguistically important motion as frame rate is reduced can be modeled by an offset of the intelligibility measure, as in Equation 6, where I is taken from Equation 5,

$$I(framerate) \quad = \quad I - f(framerate) \tag{6}$$

At very high frame rates, this offset is likely to be very low and intelligibility will be a function of only the distortion types computed by I. At extremely low frame rates, this offset is likely to be very high and intelligibility will be low regardless of the frame level quality.

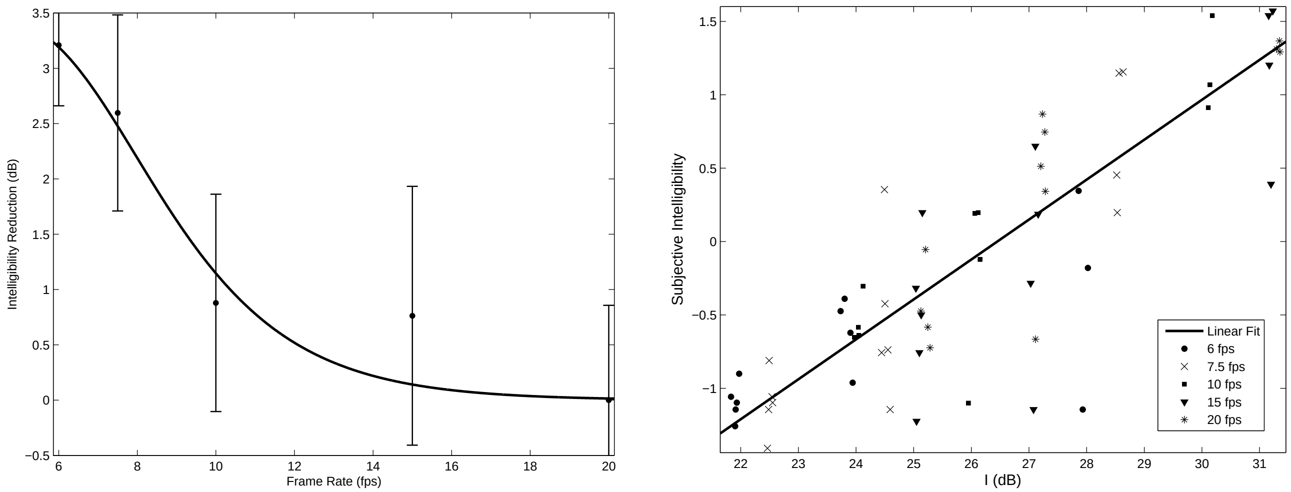
The amount of information loss is a function of frame rate and is parameterized by the subjective data. The linear fit between subjective and objective intelligibility can be written as

$$\begin{aligned} \text{Intell}_{subj} &= m(I - f(framerate)) + b \qquad (7) \\ &= mI + b - mf(framerate) \qquad (8) \end{aligned}$$

where $m = 0.272$ is the slope and $b = -7.40$ is the intercept at 20 fps taken from Table 2. The slope and intercept are fixed in order to model the difference in intelligibility between frame rates. At frame rates higher than 20 fps, all the relevant motion in ASL is being accurately captured. By definition, $f(framerate)$ is equal to 0 at 20 fps. The reduction in intelligibility at each of the test frame rate lower than 20 is computed using the intercepts in Table 2 as $\frac{b(20)-b(framerate)}{m}$.

An inverted sigmoid model is fit to this experimental data. A sigmoid is selected to model $f(framerate)$ because the gains (or reductions) in intelligibility converge as frame rate increases (or decreases). For frame rates below 6 fps, the subjective ratings are unlikely to get any worse; videos at these frame rates are completely unintelligible. Similarly, there will be little gain in intelligibility by increasing the frame rate beyond 20 fps. The value of the intelligibility offset is modeled as

$$f(framerate) \quad = \quad a_1 * \left(1 - e^{-e^{a_2 + a_3 f}}\right), \tag{9}$$



(a) Sigmoidal fit of frame rate reduction to intelligibility. $R^2 = 0.93$ for this model. Error bars indicate the 95% confidence intervals.

(b) Objective intelligibility versus subjective ratings for complete data set.

Figure 3. The offset calculated from the sigmoidal fit in (a) is applied to the measured objective intelligibility. The linear fit on the entire data set (all frame rates) achieves a correlation coefficient of 0.853, as illustrated in (b).

where a_1 controls the upper asymptote, a_2 and a_3 control the convergence locations and growth rate, and f is the frame rate in frames-per-second.

The resulting fit achieves $R^2 = 0.93$ and is plotted in Figure 4.2(a), with values of $a_1 = 3.2$, $a_2 = 4.6$, $a_3 = -0.55$. While the point at 15 fps is the largest outlier, it is still within the 95% confidence interval. New objective intelligibility values are computed using the formula $I - f(framerate)$ and are compared with the subjective intelligibility scores over all the frame rates simultaneously. The frame rate modified objective measure achieves a linear correlation coefficient of 0.85 with the subjective responses, as illustrated in Figure 4.2(b).

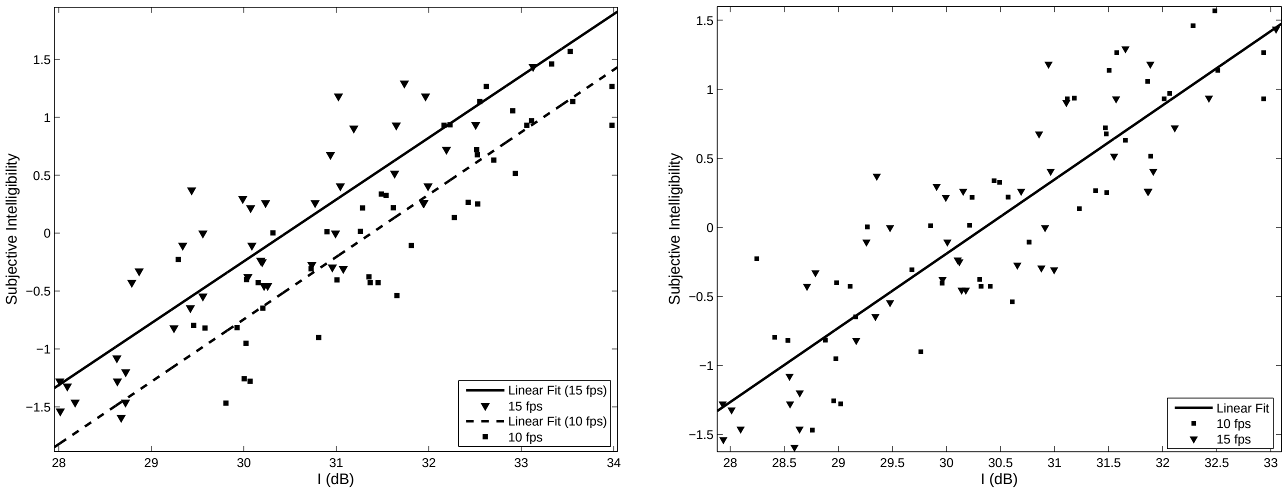
4.3 Verification Using an Independent Test Set

The data from the experiment described in Section 2 is used as a test set to verify the effectiveness of the proposed model at predicting intelligibility consistently at different frame rates. In the test data set, the objective measure of intelligibility in Equation 5 is highly correlated with subjective ratings within a single frame rate. Furthermore, the slope of the linear relationship between objective and subjective intelligibility was virtually identical for the 10 and 15 fps data sets. This result is consistent with the findings presented in Table 2. Figure 4(a) illustrates the linear relationship for the 10 and 15 fps data sets separately.

The frame rate offset function is applied to the test set data in order to unify the multiple frame rates. The resulting unified data set achieves a Pearson correlation coefficient of 0.88. Figure 4(b) illustrates the linear relationship of the complete data set. The high correlation on both the test set and the original data set used to parameterize the model indicate that the proposed sigmoidal model effectively characterizes the impact of frame rate on intelligibility.

5. CONCLUSION

An objective measure of intelligibility for American Sign Language video was presented, which analyzes both distortions in the signer's face and hands and distortions which reduce the temporal coherence of the video. Disruptions to the temporal coherence are quantified in two ways. First, background macroblocks that are highly correlated with co-located face and hand macroblocks in previous frames have a significant impact on the perception of smooth motion. These macroblocks are identified using the block-level correlation. Intelligibility is



(a) Separate frame rate test sets for 10 and 15 fps. The Pearson linear correlation coefficient is 0.88 for the 10 fps set and 0.87 for the 15 fps set

(b) Objective intelligibility versus subjective ratings for complete test set. Pearson linear correlation coefficient is 0.88 for this test set.

Figure 4. The frame rate offset model in Figure 3(a) is applied to a test data set consisting of videos at 10 and 15 fps. The independent linear fits for each frame rate is illustrated in (a). The model successfully combines intelligibility at different frame rates, as demonstrated by the high correlation coefficient in (b).

measured independent of frame rate as the log of the weighted summation of the mean squared error in correlated background macroblocks and in the signer's face and hands.

Second, a subjective experiment was designed in order to quantify the loss of information due to reductions in frame rate. The objective intelligibility measure is offset by an amount that is dependent on the frame rate of the sequence. The frame rate offset function is modeled by a sigmoidal function, and, by incorporating this frame rate offset, the objective intelligibility measure achieves a Pearson linear correlation of 0.85 on the training experimental data set and 0.87 on an independent test data set. The effect of variable frame rates and intermittent frame drops are currently being explored. According to the linguistic structure of ASL, not all frames will contain the same level of information. For example, if the dropped frames contain a relevant facial expression, intelligibility will be more affected than when the dropped frames contain only manual signs. Work is being done to automatically detect the linguistically relevant frames in order to improve the objective intelligibility measure.

REFERENCES

1. F. Ciaramello and S. Hemami, "Complexity constrained rate-distortion optimization of sign language video using an objective intelligibility metric," *Proc. SPIE Visual Communication and Image Processing* **6822**, 2008.
2. A. Cavender, R. Ladner, and E. Riskin, "MobileASL: Intelligibility of sign language video as constrained by mobile phone technology," in *ASSETS 2006: The Sixth International ACM SIGACCESS Conference on Computers and Accessibility*, 2006.
3. W. C. Stokoe, *Sign Language Structure*, Linstok Press, Inc, Silver Spring, MD, 1978.
4. S. K. Liddell and R. E. Johnson, "American sign language: The phonological base," in *Sign Language Studies*, **64**, pp. 195–278, 1989.
5. C. Baker and C. A. Padden, "Focusing on the nonmanual components of American Sign Language," in *Understanding language through sign language research*, P. Siple, ed., pp. 27–57, Academic Press, (New York, NY), 1978.

6. S. K. Liddell, “Nonmanual signals and relative clauses in American Sign Language,” in *Understanding language through sign language research*, P. Siple, ed., pp. 59–90, Academic Press, (New York, NY), 1978.
7. F. Ciaramello and S. Hemami, “‘Can you see me now?’ An objective metric for predicting intelligibility of compressed American Sign Language video,” in *Proc. SPIE Vol. 6492, Human Vision and Electronic Imaging ’07*, B. E. Rogowitz, T. N. Pappas, and S. J. Daly, eds., **6492**, 2007.
8. M. A. Changizi, A. Hsieh, R. Nijhawan, R. Kanai, and S. Shimojo, “Perceiving the present and a systematization of illusions,” *Cognitive Science: A Multidisciplinary Journal* **32**(3), p. 459, 2008.
9. Z. Wang, L. Lu, and A. Bovik, “Foveation scalable video coding with automatic fixation selection,” *Image Processing, IEEE Transactions on* **12**, pp. 243–254, Feb 2003.
10. K. Seshadrinathan and A. C. Bovik, “An information theoretic video quality metric based on motion models,” in *Workshop on Video Processing and Quality Metrics for Consumer Electronics 2007 Proceedings*, January 2007.
11. D. Parish, G. Sperling, and M. Landy, “Intelligent temporal subsampling of american sign language using event boundaries,” *Journal of Experimental Psychology: Human Perception and Performance* **16**(2), pp. 282–294, 1990.
12. 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,” *Proceedings of RESNA Annual Conference* **14**, pp. 38–40, 1991.
13. N. Cherniavsky, A. Cavender, E. Riskin, and R. Ladner, “Variable Frame Rate for Low Power Mobile Sign Language Communication,” in *Proceedings of ACM SIGACCESS Conference on Computers and Accessibility*, pp. 163–170, 2007.
14. S. Hooper, C. Miller, S. Rose, and G. Veletsianos, “The effects of digital video quality on learner comprehension in an american sign language assessment environment,” *Sign Language Studies* **8**(1), pp. 42–58, 2007.
15. P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” *Proc. Computer Vision and Pattern Recogntion* , 2001.
16. R. A. Foulds, “Biomechanical and perceptual constraints on the bandwidth requirements of sign language,” in *IEEE Trans. On Neural Systems and Rehabilitation Engineering*, **12**, pp. 65–72, March 2004.