# Quality versus Intelligibility: Evaluating the Coding Trade-offs for American Sign Language Video

Frank Ciaramello, Jung Ko, Sheila Hemami
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. Compressed American Sign Language 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 studies the trade-offs between intelligibility and quality when varying the proportion of the rate allocated explicitly to the signer. An intelligibility distortion measure and a quality measure (PSNR) are applied in a rate-distortion optimization framework and a novel encoding technique controls the degree to which intelligibility is emphasized over quality. Understanding the relationship between intelligibility and quality allows the encoder to identify operating points that maximize PSNR while maintaining a minimal level of intelligibility. At fixed bitrates, PSNR can be increased on average by 5 dB with little penalty in intelligibility by providing a nominal amount of rate to the background region. Further increases in PSNR can be achieved at the price of reduced intelligibility.**

## I. 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 [1].

The objective intelligibility measure is used to encode sign language video in a rate-distortion optimization setting and provides bitrate reductions up to 50% compared to a mean-squared-error (MSE) optimized encoder [2]. The intelligibility optimized encoder achieves bitrate reductions by heavily distorting the background video region, which, for some ASL users, can be distracting and annoying. Allowing the user to adjust the level of background distortion addresses this problem, but lowering the distortion in the background region can lead to an unintelligible signer. The goal of this work is to identify optimal operating points that can increase the aesthetic quality of the video while maintaining the intelligibility of the ASL communication.

This work studies the trade-offs between quality, measured as peak signal to noise ratio (PSNR), and intelligibility when applying a general video encoding algorithm and an intelligibility optimized encoding algorithm, each working within the H.264 encoding standard. A novel technique is developed that incrementally varies the amount of rate allocated exclusively to the signer, accommodating various user preferences. It is important to note that ASL video is studied in this work because it provides clearly defined regions-of-interest and an objective distortion measure that reflects human ratings. The methodology presented is applicable to any class of ROI video, particularly videos having multiple regions of varying importance.

This paper is organized as follows. Section II reviews rate-distortion optimization in H.264 and describes the differences between a general purpose video encoder, designed to maximize quality, and an intelligibility optimized video encoder, designed for ASL video. Section III describes how the intelligibility optimized encoder is modified to account for user preferences in background distortion levels. Section IV provides an analysis of this modified coder with respect to the trade-off between PSNR and intelligibility and identifies optimal operating points. Conclusions and future work are discussed in Section V.

## II. Rate-distortion optimization in H.264

Rate-distortion (R-D) optimization for video requires the selection of a set of encoding parameters for each macroblock that minimizes the distortion subject to a target bitrate. In H.264, the video frames are divided into 16×16 pixel macroblocks, each requiring encoding parameters consisting of a quantization step size, defined by the quantization parameter (QP) and macroblock encoding mode, $p$. The optimal macroblock encoding parameters, QP and $p$, are determined by minimizing the Lagrangian R-D cost function, according to

$$\min_{p,QP} J(X,p,QP|\lambda) = D(X,p,QP) + \lambda R(X,p,QP), \quad (1)$$

where $\lambda$ is the Lagrange parameter, $X$ is the current macroblock, $R$ is the bitrate required to encode the macroblock using QP and $p$, and $D$ is the resulting distortion. The Lagrangian optimization facilitates both quality optimization and intelligibility optimization via the application of an appropriate distortion measure.

### A. Quality optimized encoder

General purpose video encoders are designed to maximize the overall quality of the input video, where quality is typically measured as 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).

In this work, the x264 encoder [3], an open source implementation of H.264, is selected as the quality optimized encoder because it provides significant speed improvements over the H.264 JM reference encoder. The R-D optimization algorithm in x264 applies empirical models to select a QP for the each frame, which is subsequently mapped to a $\lambda$, in order to achieve a target bitrate [4]. The remaining encoding decisions are made by minimizing Eq. (1) at each macroblock, using mean squared error (MSE) as the distortion measure. A consequence of selecting a frame-level QP and minimizing MSE is that all macroblocks are considered equally important and PSNR is maximized.

### *B. Intelligibility optimized encoder*

The intelligibility optimized encoder, implemented as a modified version of x264, incorporates an intelligibility distortion measure into the R-D optimization in Eq. (1). The objective intelligibility measure is a function of the distortion only in linguistically relevant regions, i.e., the signer's face, hands, and torso, and the measure accurately estimates subjective intelligibility ratings of ASL video [1].

For the purposes of R-D optimization, macroblocks in the input video are segmented into either face, hands, torso, or background, using skin color detection and morphological processing. The intelligibility distortion measure can be modeled as the sum of weighted MSE in each of the segmented regions, computed according to

$$\begin{aligned} D_{Intell}(n) &= \alpha_F D_F(n) + \alpha_H D_H(n) + \\ &\quad \alpha_T D_T(n) + \alpha_{BG} D_{BG}(n), \end{aligned} \tag{2}$$

where $D_F$, $D_H$, $D_T$, and $D_{BG}$ are the MSE for the face, hands, torso, and background regions in frame $n$. The region weights are given by $\alpha_F = 1.6$, $\alpha_H = 0.5$, $\alpha_H = 0.1$, and $\alpha_{BG} = 0$. A temporal pooling mechanism, which computes both the mean and the temporal variation of $D_{Intell}(n)$, provides a single distortion value for the entire video, which is denoted $D_{Intell}$ [5].

Because $D_{Intell}$ is a distortion measure, it is inversely proportional to intelligibility. The varying weights control the relative importance of each type of macroblock in the region of interest (ROI); a distortion in the signer's face will result in a lower intelligibility than the same level of distortion in the signer's torso. Distortions in background macroblocks do not contribute to $D_{Intell}$; $\alpha_{BG}$ and $D_{BG}$ are included in Eq. (2) to explicitly account for all macroblocks.

In contrast with the quality optimized encoder, which computes a global QP for the entire frame, the intelligibility optimized encoder uses a trellis-based, R-D optimization procedure that computes the optimal QP for each macroblock [6]. Applying this procedure to a collection of ASL videos over a range of $\lambda$ values provides a functional relationship between $\lambda$ and the optimal QP for each region, given by $2^{\frac{QP_k-12}{3}} = \frac{\lambda}{0.65\alpha_k}$, where $QP_k$ is the quantization parameter and $\alpha_k$ is the weight for region $k$, consisting of face, hands, torso, and background. This functional relationship is similar to one developed for arbitrary video content in H.264 [7]. For increasing values of $\alpha_k$, corresponding to increasing region importance, the quantization step size will decrease. As a result, more important regions in the video frame are encoded with a lower quantization step size and are allocated more rate.

A value of $\lambda$ specifies the QP selection for each macroblock, reducing the optimization in Eq. (1) to selecting only the optimal macroblock mode. Ultimately, this allows a single parameter, $\lambda$, defined for the entire frame, to select the proper encoding parameters for every macroblock. Rate control is performed at the frame-level by adjusting $\lambda$ according to $\lambda(n+1) = \lambda(n) - \left(\frac{R_{target}}{R_{actual}} - 1\right)$, where $R_{target}$ and $R_{actual}$ are the target bits and actual bits for frame $n$ [8].

## III. Varying ROI Priority - blending intelligibility and quality

The intelligibility optimized and quality optimized encoders represent two encoding extremes, either allocating all the rate only to the signer or distributing the rate evenly among every macroblock. When optimizing strictly for intelligibility, the rate allocated to the background is minimized independent of the resulting distortion, creating severe compression artifacts in the background macroblocks. Participants in subjective experiments report varying levels of distraction caused by heavily distorted backgrounds [9]. Quality optimized video provides similar levels of distortion across the entire frame, eliminating extreme distortions in the background. However, when optimizing strictly for quality, distortions in the signer can lead to unintelligible video. These two encoding extremes alone are incapable of accommodating the preferences of ASL users and maintaining intelligible video. This motivates the need for an encoder that provides a variable trade-off between intelligibility optimization and quality optimization.

The intelligibility optimized encoder described in Section II-B is modified to include a global distortion weight, $\alpha_{min}$, which specifies the minimum weight to be applied to all regions in the frame. Specifically, if $\alpha_{min} \geq \alpha_k$, the region weight $\alpha_k$ is set equal to $\alpha_{min}$. 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 face, hands, or torso.

Modifying $\alpha_{min}$ controls the degree to which the ROI is prioritized over the rest of the frame. A region is considered prioritized if the corresponding distortion weight is larger than $\alpha_{min}$. A prioritized region has a lower QP and lower distortion than the rest of the frame. For example, the intelligibility optimized encoder corresponds to $\alpha_{min} = 0$; the entire ROI (face, hands, torso) is given priority over the background. When $\alpha_{min} = 0.1 = \alpha_T$, the distortions in the background and the torso are weighted equally, and only the face and hands are prioritized because of their higher distortion weight. As $\alpha_{min}$ increases, only the most important macroblocks are prioritized. At the extreme, when $\alpha_{min} \geq \alpha_F$, all of the

(a) Original video frame

(b) Prioritize all of the ROI. $\alpha_{min} = 0$, PSNR $= 18.44$ dB, $D_{Intell} = 0.616$

(c) Prioritize all of the ROI with nominal background distortion weight. $\alpha_{min} = 0.02$, PSNR $= 21.74$ dB, $D_{Intell} = 0.642$

(d) Prioritize only the face and hands. $\alpha_{min} = 0.1$, PSNR $= 23.43$ dB, $D_{Intell} = 0.668$

(e) Prioritize only the face. $\alpha_{min} = 0.5$, PSNR $= 25.21$ dB, $D_{Intell} = 0.759$

(f) Quality optimized. $\alpha_{min} = 1.6$, PSNR $= 25.73$ dB, $D_{Intell} = 0.846$

Fig. 1. Comparison of distortions for different levels of region-of-interest (ROI) priority. The encoding option $\alpha_{min}$ specifies the minimum distortion weight to be applied to any region. As $\alpha_{min}$, the torso, hands, and face are allocated fewer additional bits relative to the rest of the frame, causing a decrease in intelligibility. Figure 2 specifies the relationship between $D_{Intell}$ and subjective ratings of intelligibility.

regions are weighted equally and the encoder behaves as the quality optimized encoder.

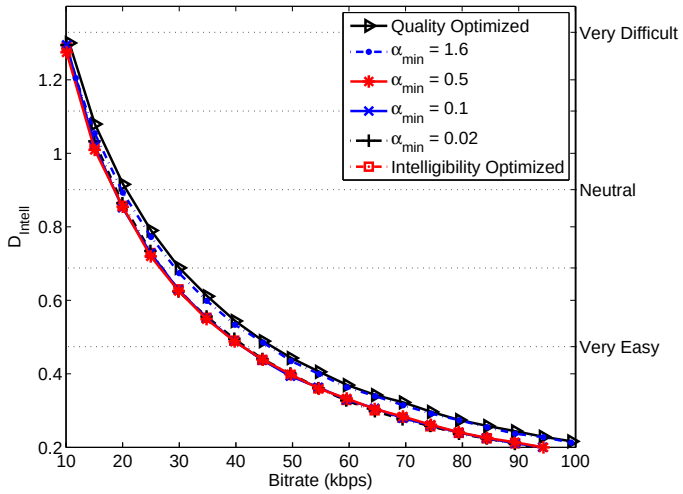
To illustrate, consider a sample ASL video, recorded in an outdoor setting with a highly active background and encoded at 55 kbps with different values of $\alpha_{min}$. Five values for $\alpha_{min}$ are selected to emphasize different operating points: $\alpha_{min} = 0$ prioritizes the entire ROI, $\alpha_{min} = 0.02$ prioritizes the entire ROI and provides a nominal amount of rate to the background, $\alpha_{min} = 0.1 = \alpha_T$ prioritizes only the signer's face and hands, $\alpha_{min} = 0.5 = \alpha_H$ prioritizes the signer's face, and $\alpha_{min} = 1.6 = \alpha_F$ prioritizes no regions and behaves as the quality optimized encoder. Frames from this video are presented in Figure 1. As $\alpha_{min}$ increases, the relative priority of the ROI necessarily decreases and intelligibility decreases, as illustrated in Figures 1(b) through 1(f). Decreasing ROI priority is reflected in an increase in the objective intelligibility distortion measure; $D_{Intell}$ increases from 0.616 to 0.846. For the subjective intelligibility ratings associated with these values, refer to Figure 2. Conversely, as $\alpha_{min}$ increases, PSNR increases from 18.44 dB to 25.73 dB. As this example demonstrates, varying $\alpha_{min}$ can provide a user with control over the level of background distortion while still prioritizing the most important regions of the signer. The following section analyzes PSNR and $D_{Intell}$ over a range of encoding bitrates and $\alpha_{min}$ values, in order to identify appropriate operating points.

## IV. Characterizing PSNR and $D_{Intell}$ for varying rate and $\alpha_{min}$

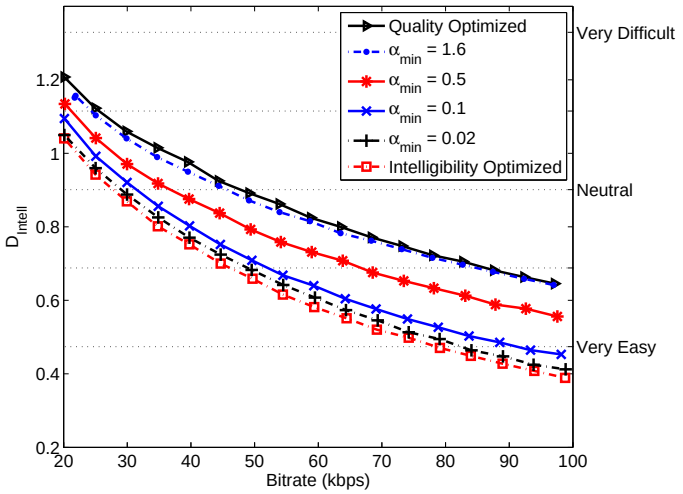
This section analyzes the rate-distortion performance for several fixed values of $\alpha_{min}$ and the relationship between PSNR and $D_{Intell}$ for varying $\alpha_{min}$ at fixed bitrates. The rate-distortion performance of the intelligibility optimized encoder and the quality optimized encoder are compared against multiple values of $\alpha_{min}$ across bitrates ranging from 20 kbps to 100 kbps. Figure 2 compares PSNR and $D_{Intell}$ for two different ASL videos: a video filmed in a studio with a static background and a video filmed on a busy street with high background activity. In each case, the intelligibility optimized encoder achieves significant bitrate reductions at fixed levels of intelligibility over the x264 encoder, demonstrated in Figures 2(a) and 2(b). The bitrate reductions primarily depend on the level of activity in the background region: 10% to 13% for the indoor video and 33% to 47% for the outdoor video.

Because the intelligibility optimized encoder allocates almost zero rate to the background, the PSNR is dominated by the distortions in the background region. As a result, increasing the bitrate for the intelligibility optimized coder yields a negligible increase in PSNR, as demonstrated in Figures 2(c) and 2(d). Because it is designed to minimize MSE, x264 achieves the highest PSNR at fixed bitrates, with 4 dB to 10 dB increases in PSNR over the intelligibility optimized encoder.

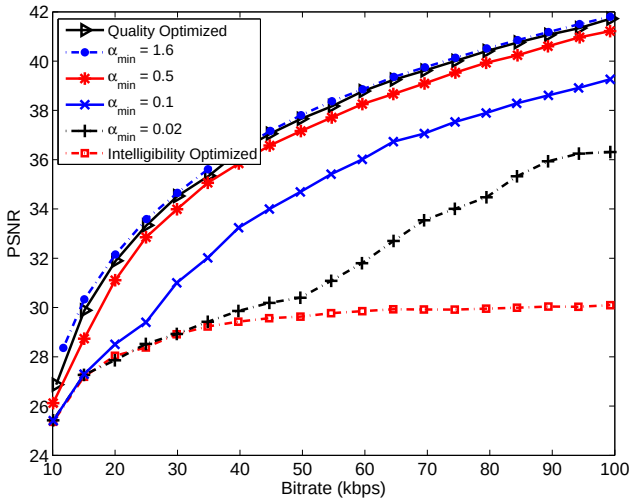
In addition to comparing the intelligibility optimized and quality optimized encoders, Figure 2 also illustrates the effect



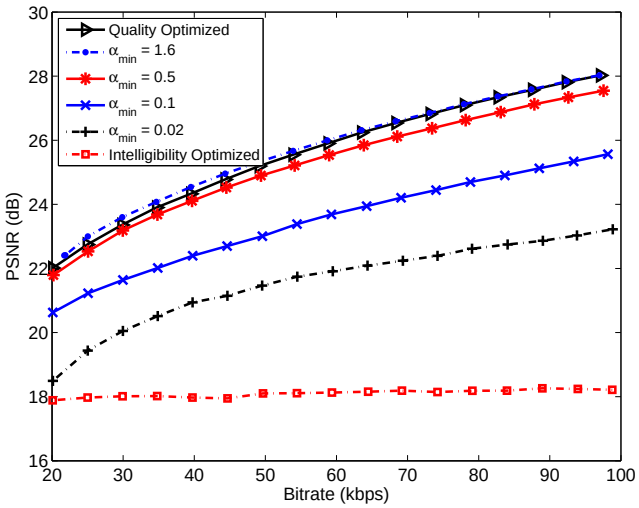

(a) Rate vs $D_{Intell}$ for an indoor ASL video. The intelligibility optimized encoder reduces bitrate by 10%-13% over the quality optimized encoder.




(b) Rate vs $D_{Intell}$ for an outdoor ASL video. The intelligibility optimized encoder reduces bitrate by 33%-47% over the quality optimized encoder.




(c) Rate vs PSNR for an indoor ASL video.




(d) Rate vs PSNR for an outdoor ASL video.

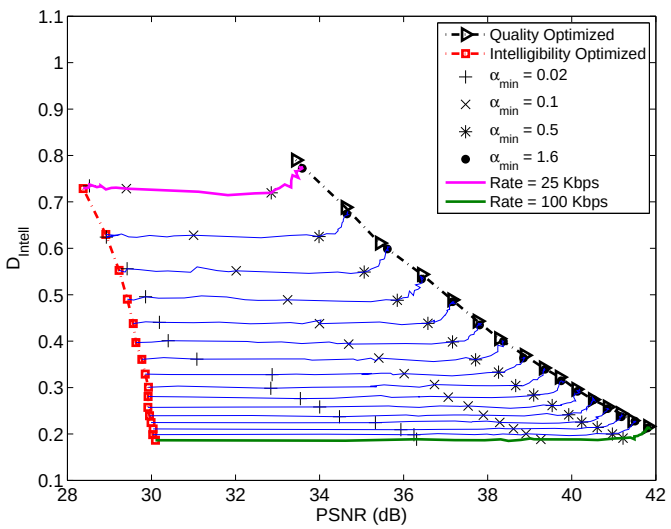
Fig. 2. Rate-distortion plots for the quality optimized coder, the intelligibility optimized encoder, and several values of $\alpha_{min}$. For (a) and (b), the left y-axis provides the objective intelligibility distortion measure, $D_{Intell}$, and the right y-axis provides the subjective rating categories corresponding to the objective distortion values. For (c) and (d), the y-axis provides PSNR. For a fixed level of intelligibility, rate reductions increase for sequences with increasing background activity. When $\alpha_{min} = 0.02$, PSNR increases by several dB and $D_{Intell}$ increases negligibly. When $\alpha_{min} = 1.6$, all the region distortions are weighted equally and the encoder operates identical to the quality optimized encoder.

of varying $\alpha_{min}$. Setting $\alpha_{min} = 0.02$ applies a nominal weight to the background distortion and results in substantial increases in PSNR with only slight increases in $D_{Intell}$. Further increasing the $\alpha_{min}$ results in increased PSNR at the expense of intelligibility. When $\alpha_{min} = 1.6$, the encoder performs nearly identical to x264, demonstrating that it behaves as a quality optimized encoder at this point.

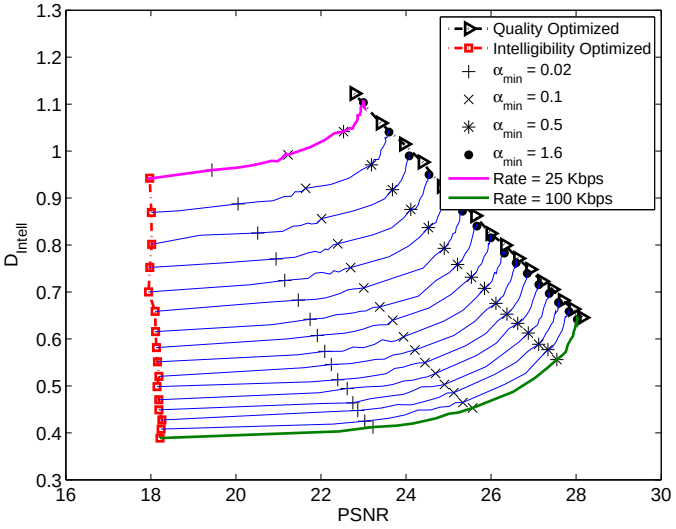
The value of $\alpha_{min}$ controls the priority given to the ROI coder. When $\alpha_{min} = 0$, the encoder is optimizing only for intelligibility. When $\alpha_{min} = 1.6$, the encoder is optimizing only for quality. To explicitly evaluate the trade-off between PSNR and intelligibility afforded by $\alpha_{min}$, the indoor and outdoor videos are encoded at bitrates ranging from 25 to 100 kbps in increments of 5 kbps. $\alpha_{min}$ is varied from 0 to 0.1 in increments of 0.01 and from 0.1 to 1.6 in steps of 0.1.

Sweeping through the range of $\alpha_{min}$ creates a class of encoding scenarios that are the convex combination of the intelligibility optimized and quality optimized encoders, as illustrated in Figure 3. Each curve corresponds to a fixed encoding bitrate and each point in the curve corresponds to a particular value of $\alpha_{min}$. Varying $\alpha_{min}$ from 0 to 1.6 yields combinations of PSNR and intelligibility that span the space between the two encoding extremes: optimizing exclusively for quality or for intelligibility.

The relationship between $D_{Intell}$ and PSNR, as $\alpha_{min}$ varies, depends on the amount of activity in the background region. Increases in $D_{Intell}$ of approximately 0.2 correspond to a difference of 1 point on a 5 point subjective intelligibility scale. An increase in $D_{intell}$ of less than 0.02, i.e., 10% of 0.2, can be considered negligible. When compared to




(a) PSNR vs $D_{Intell}$ for an indoor video having a static background.




(b) PSNR vs $D_{Intell}$ for an outdoor video having an active background.

Fig. 3. PSNR versus $D_{Intell}$ for videos with different levels of background activity. Each solid line corresponds to a fixed bitrate and varying $\alpha_{min}$. The bitrates vary between 25 kbps and 100 kbps in increments of 5 kbps. Depending on the amount of activity in the background, PSNR can be increased by several dB without a significant increase in $D_{Intell}$, when compared to the intelligibility optimized encoder.

the intelligibility optimized encoder, selecting $\alpha_{min} = 0.5$ increases PSNR in the indoor video between 4.5 dB and 11 dB, depending on the encoding bitrate, with negligible increase in $D_{Intell}$, as illustrated in Figure 3(a). For the high background activity video in Figure 3(b), only a nominal value of $\alpha_{min} = 0.02$ can be selected before the increase in $D_{Intell}$ becomes non-negligible. At this point, PSNR is increased between 1.3 dB and 4.7 dB, depending on the encoding bitrate.

The slope of the PSNR versus $D_{Intell}$ curves is steepest when $0.5 < \alpha_{min} < 1.6$. In this region, when compared to the quality optimized encoder, $D_{Intell}$ is reduced between 0.03 and 0.08 for a corresponding decrease in PSNR of only between 0.5 dB and 0.6 dB. The signer's face is relatively small compared to the rest of the frame and distortions in the signer's face have the largest impact on $D_{Intell}$. Prioritizing the signer's face decreases distortions in the corresponding macroblocks and increases intelligibility without creating substantial distortions in the other regions.

## V. Conclusion and Future Work

This work presented an H.264 video encoder for American Sign Language video that optionally controls the trade-off between optimizing intelligibility, as computed by distortions measured in the signer, and optimizing quality, as measured by PSNR. Even in videos with highly active backgrounds, PSNR can be increased by at least 4 dB without sacrificing intelligibility. For videos with low background activity, it is possible to maximize both PSNR and intelligibility by only prioritizing the signer's face. In future work, a subjective study will be designed to evaluate user preferences for the quality-intelligibility trade-off. Of particular interest is the willingness to sacrifice intelligibility in order to decrease distortion in the background (and consequently increase PSNR).

The intelligibility distortion applied in this work is a region-sensitive distortion measure, computed via the linear combination of weighted MSE, that can be generalized to any ROI video. In future work, advanced spatial distortion measures that incorporate perceptual models will be considered. As long as the distortion measure facilitates the computation of region distortions (e.g., produces a spatial error map), perceptual models, such as masking or contrast sensitivity, can only serve to improve the accuracy of the ROI distortion measure. In addition, non-linear combinations of the region distortions may also improve the subjective intelligibility estimation accuracy. Either of these modifications will affect the functional relationship between quantization step size, QP, and the Lagrange parameter, $\lambda$. In this case the real-time selection of a QP for each region must be re-evaluated.

## References


[1] F. Ciaramello and S. Hemami, "'Can you see me now?' An objective metric for predicting intelligibility of compressed American Sign Language video," *Proc. SPIE Human Vision and Electronic Imaging*, vol. 6492, Jan. 2007.

[2] ——, "Complexity constrained rate-distortion optimization of sign language video using an objective intelligibility metric," *Proc. SPIE Visual Communication and Image Processing*, vol. 6822, Jan. 2008.

[3] x264. Http://developers.videolan.org/x264.html.

[4] L. Merritt and R. Vanam, "Improved rate control and motion estimation for H.264 encoder," *Proc. IEEE International Conference on Image Processing*, vol. 5, 2007.

[5] F. Ciaramello and S. S. Hemami, "An objective intelligibility measure for assessment and compression of American Sign Language video," in preparation.

[6] A. Ortega and K. Ramchandran, "Forward-adaptive quantization with optimal overhead cost for image and video coding with applications to mpeg video coders," in *Proc. of IS&T/SPIE Digital Video Compression '95*, Feb. 1995.

[7] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," in *IEEE Trans. Circuits and Systems for Video Technology*, vol. 13, no. 7, Jul. 2003.

[8] T. Wiegand, M. Lightsone, D. Mukherjee, T. G. Camplbell, and S. Mitra, "Rate-distortion optimized mode selection for very low bit rate video coding and the emerging H.263 standard," in *IEEE Trans. Circuits and Systems for Video Technology*, vol. 6, no. 2, Apr. 1996.

[9] F. Ciaramello. (2009, Dec.) Unpublished subjective test results.