# Improving Compressed Video Sign Language Conversations in the Presence of Data Loss

Jaehong Chon[†] Sam Whittle[‡] Eve A. Riskin[†]
Richard E. Ladner[§]
[†] Department of Electrical Engineering, Box 352500,
University of Washington, Seattle, WA 98195-2500
[‡] Google, Seattle, WA 98103
[§] Department of Computer Science and Engineering, Box 352350,
University of Washington, Seattle, WA 98195-2350
Email: {jaehong, riskin}@ee.washington.edu,
samuelw@google.com, ladner@cs.washington.edu

**Abstract**

The goal of the MobileASL (American Sign Language) research project is to enable sign language communication over the U.S. cellular network, which is low bandwidth and lossy. Data loss can greatly impact the quality of compressed video because of temporal and spatial error propagation. We investigate techniques to minimize the effect of data loss for improving compressed video sign language conversations. As both computational power and bandwidth are limited on cellular devices, we must carefully allocate these resources. Specifically we focus on utilizing feedback to recover from data loss.

## 1 Introduction

The goal of the MobileASL (American Sign Language) research project is to enable sign language communication over the US cellular network. American Sign Language (ASL) is the primary language of more than 500,000 Deaf individuals in the U.S. [1] and is greatly preferred to text for its expressiveness, speed, and ease of use.

For ASL conversation, the quality of the video is measured by individual frame quality, frame rate, delay, and jitter. The cellular network is of limited bandwidth and there are no guarantees that packets sent will be successfully received.

Compressed video is highly vulnerable to packet loss because video coding uses spatial and temporal correlation between video frames. The most important problem is error propagation due to inter-frame prediction. Periodically encoding some frames as I-frames can prevent this problem, but because this requires many more bits, it is not suitable for low bit rate video communication over cellular networks.

Unlike video streaming services, which provide buffering to prevent impairments, mobile video calls are real-time and the video quality must recover quickly when packet loss occurs. This is the main cause for impairments over wireless networks [2]. Developing an effective algorithm to overcome the effect of packet loss is critical to improving conversations for sign language communication on mobile devices.

This paper is organized as follows. In Section 2, we investigate related work. We describe our algorithms in Section 3. In Section 4, we present our experiments and results. We conclude in Section 5.

# 2 Related Work

A variety of error-resilient techniques have been proposed to mitigate the effects of packet loss and inter-frame error propagation. Because they cause delay, error control schemes within the source coding layer such as interleaving and forward error control are not suitable for video telephony and video conferencing services. Schemes that request retransmission of corrupted frames or lost packets also cannot be used.

## 2.1 Error Resiliency Features in H.264

The H.264 standard employs various error resiliency features such as Flexible Macroblock Ordering (FMO), Arbitrary Slice Ordering (ASO) and Data Partitioning. Partitioning the picture into independently-decodable regions called slice groups enhances robustness to data loss by managing the spatial relationship between the regions, and preventing errors from being scattered across the whole frame. Since some encoded bitstreams in a frame are more important than others, this enables unequal error protection. For example, low and high frequency transform coefficients can be assigned to different slices and can be processed using different error protection algorithms. In addition, prediction beyond the slice boundaries is forbidden to prevent error propagation from cross-slice predictions.

For MobileASL, however, since the frames are small and encoded at a low bit rate, each predicted frame (P-frame) fits within a single packet, meaning each occurrence of packet loss corresponds to the loss of an entire frame. Thus, the data partitioning and slice structuring done as a result of FMO and ASO in the H.264 standard are not useful for MobileASL.

## 2.2 Intra Refresh

H.264 uses intelligent intra-block refresh by rate-distortion optimization to select between inter/intra mode at the encoder. It has been extensively explored to prevent or minimize the effect of error propagation [3]. In rate-distortion optimization, packet-loss probability models have been proposed to predict the error of the decoder's reconstructed video [4, 5]. Rate-distortion optimization can reduce the required bit budget but doesn't guarantee prevention of temporal error propagation due to frame dropping.

We use intra refresh to improve sign language conversations by utilizing feedback. Feedback is a useful strategy that is further developed in the next section.

## 2.3 Reference Picture Selection

Feedback-based reference picture selection tries to avoid using a corrupted frame as a reference frame [6]. Because the decoder sends feedback acknowledgements, the encoder can determine which parts of frames were erroneously decoded and then use multiple reference frames to stop temporal error propagation. In such cases, the computational complexity for motion estimation and motion compensation increases in proportion to the number of reference frames used. Since motion estimation and compensation are the most time-consuming modules of video encoding, the overall encoding time increases, especially for mobile phones. In addition, correlation between frames decreases between reference frames that are farther apart in time [7]. This, in turn, decreases coding efficiency.

## 2.4 Error Concealment Schemes

Typically, error concealment by the decoder utilizes spatial, spectral, and/or temporal correlation of the received video. Spatial and spectral error concealment use the information of the neighboring blocks in the spatial or frequency domains. In contrast, temporal error concealment uses the temporal correlation between adjacent frames.

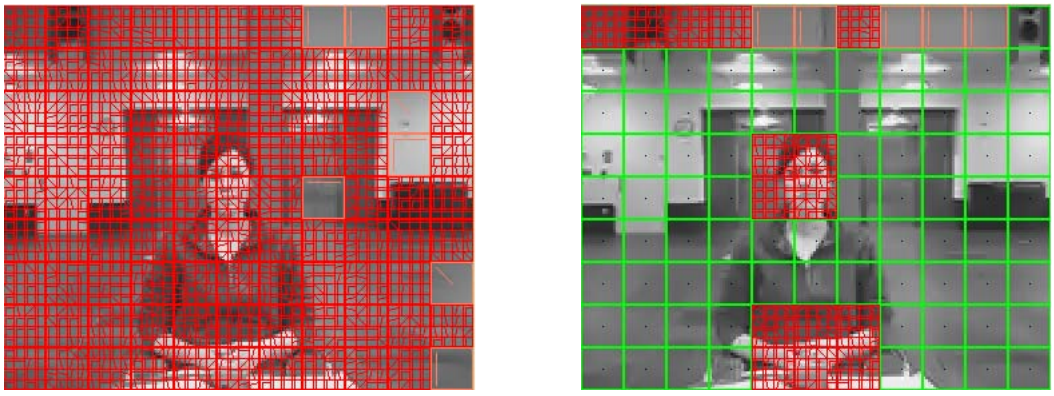
The simplest method to recover the missing frames is to repeat the last received frame with all zero motion vectors, which is called temporal replacement or frame copy [8, 9]. Most of the temporal error concealment techniques assume that only a few macroblocks are lost and these techniques utilize previous frames to estimate the motion vectors associated with the lost macroblocks [10, 11]. However, in MobileASL, one data packet carries a whole frame so entire frame loss must be considered.

To address this, Zhu *et al.* proposed motion vector extrapolation (MVE) in which the motion vectors of macroblocks are extrapolated from the last received frame, and the overlapped areas between the damaged block and the motion extrapolation macroblocks are estimated [12]. Li *et al.* proposed a pixel-based MVE (PMVE) method to conceal the missing frame which extends MVE to the pixel level [8]. Yan and Gharavi proposed a hybrid motion vector extrapolation method based on PMVE, which uses not only the extrapolated motion vectors of the pixels, but also the extrapolated motion vectors of the macroblock to discard the wrongly extrapolated motion vectors [9].

For estimating motion vectors for whole-frame loss, the upper bound of MVE is that the original motion vectors are correctly received, but the residual information is lost. However, this upper bound can neither prevent the temporal error propagation nor remove the corrupt regions until receiving a new independent frame [9]. Therefore, any method based upon MVE that deals with whole-frame loss will not be appropriate for two-way sign language communication.

# 3 Low-Complexity Feedback-Based Encoding

Feedback about dropped frames can be useful to recover data loss in two-way communication. This requires minimal additional complexity because lost frames can be detected by numbering the packets or tracking the frame number contained in a packet.



(a) I-frame (b) P-frame

Figure 1: Comparison of number of intra-blocks contained in an I-frame and a P-frame used for refreshing. Intra-blocks are drawn in red and skip-blocks are indicated with green. Most gains will not be as extreme. Note that there are many I-blocks in faces and hands.

Some visual distortion will exist between the time of packet loss and the time that feedback is received. Because MobileASL conversations are real-time and users can ask for clarification if necessary, this short distortion may not be too disruptive.

We investigate feedback to improve compressed video sign language conversations in the presence of data loss [13] using a standard 3G cellular system simulator. We focus on several strategies that utilize feedback with intra refresh mentioned above and compare the results.

## 3.1 Intra Block and Skip Block

Encoding decisions on the macroblock level are guided by frame types. There are two possible frame types for real-time encoding: I-frames and P-frames. An I-frame is a frame in which every macroblock is encoded as an intra-block. I-frames break all temporal dependencies and only have spatial dependencies within the frame at a cost of many more bits. P-frames contain intra-blocks, inter-blocks and skip-blocks. The encoding method for each macroblock is chosen independently to minimize the bits necessary for a set quality. Figure 1 shows how each frame type has different types of macroblocks.

## 3.2 Bursty I-frame Refresh

The simplest way to stop temporal error propagation is to use I-frame encoding whenever a packet is lost. Unfortunately the losses over a wireless link are sometimes bursty, which can produce worse distortion than an equal number of isolated losses [14].

A single I-frame can be used efficiently for both isolated loss and bursty losses. When the loss is detected at the decoder, it requests that the encoder transmit a I-frame. However, additional frames may be lost before the corrected I-frame is received by the decoder and the decoder will generate feedback again. When the encoder receives additional feedback requesting another I-frame within the round trip time, this request can be safely ignored

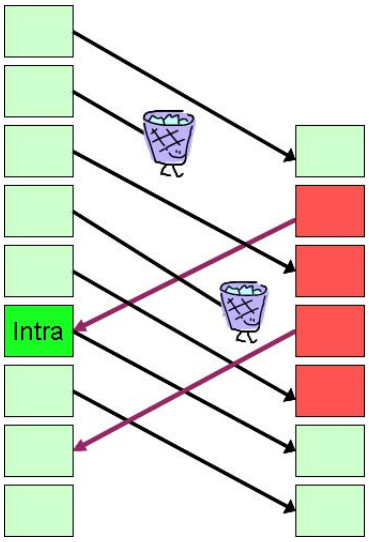



Figure 2: Demonstration of single intra frame recovering from multiple losses. The state of the encoder and decoder are on the left and right respectively. Time advances toward the bottom of the page.

because the initial I-frame will have already terminated the error propagation at the decoder. The first I-frame can still fully recover from future losses that occur within a short time frame.

Figure 2 illustrates how multiple losses are handled. The left column represents the frames produced by the encoder and the right column is the state of frames shown by the decoder, with time running vertically from top to bottom. Arrows indicate the transmission through the network of frames and feedback due to losses of frames. Note how the single I-frame recovers from multiple corruptions. In this example, the round trip time is only 4 frames. As the round trip time increases, the benefits of this method grow equally.

As we explained in a previous section, I-frames require many bits and are thus not well-suited for low-bit-rate communication. We can increase the quantization parameter for I-frames to keep the bandwidth low. When doing so, several additional P-frames are required for full PSNR recovery.

### 3.3 Bursty P-frame Refresh

I-frame refresh based on feedback can be further enhanced by exploiting temporal correlation through inter-frame prediction.

In inter-frame prediction, some macroblocks are coded in the SKIP mode if they are in regions that do not change over time, such as the background. These macroblocks can be reused in a P-frame. Therefore, we can reduce the number of intra-blocks needed to refresh the video by keeping track of which blocks are coded as SKIP at the encoder. In ASL communication, the portion of SKIP blocks is generally high since there is little motion in the background.

Now, instead of an I-frame, a P-frame that has only intra-blocks and skip-blocks can be used to stop error propagation. The resulting encoded frames will have fewer intra-blocks and thus fewer bits. The unnecessary use of intra-blocks in the background when using I-frame

refresh can cause a flickering effect because of color mismatch between frames, especially when the frames are encoded at a lower bit rate. Therefore, the use of SKIP blocks in the background can further improve the quality when using P-frame refresh.

Furthermore, for bursty losses, this algorithm works similarly to I-frame refresh. After the first P-frame is received, the video quality after the P-frame will be smooth even if further loss occurs within the round trip time. This algorithm can be easily implemented using only an counter per macroblock to keep track of how many times skip-blocks have been continuously used.

# 4 Results

This section contains simulation results. We present comparisons to demonstrate the effectiveness of utilizing feedback.

## 4.1 Simulation Tools

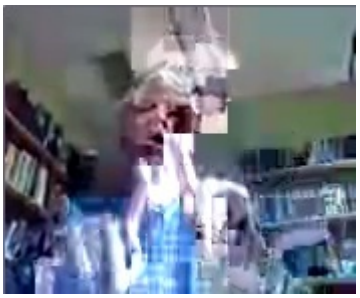
To evaluate the efficiency of the proposed approach, experiments were conducted using a network simulator provided by the 3GPP video ad hoc group [15]. This network simulator has been used in various ways. Tizon *et al.* used it to evaluate the efficiency of their approach on cellular networks [16]. Devadoss *et al.* implemented their own module with the same behavior that takes care of fragmenting packets into RLC (Radio Link Control) frames, reassembling the received ones into IP level packets, and introducing link losses and delays as specified [17]. Singh *et al.* conformed the behavior described in this simulator to investigate rate adaptation mechanisms for conversational 3G video [18].

The software provided by the 3GPP video ad hoc group is implemented to simulate an RTP streaming session over 3GPP networks (GPRS, EDGE, and UMTS) offline. The network parameters throughout the session in the simulator are nearly constant. Error masks, which are used to inject errors at the physical layer, are generated from link-level simulations at various bearer rates and block error rates to simulate packet errors. If the RLC-PDU (Radio Link Control - Protocol Data Unit) is corrupted or lost due to the error at the lower layer, it is discarded and then not given to the upper layer. Moreover, this simulator deals with actual transmission time over the networks using frame sizes of RLC layer and scheduling of the RLC layer. If the maximum transfer delay caused by retransmission or buffering at lower layer is reached, the corresponding RTP packet is also discarded.

We assume a fixed round trip time, measured in frames. For our results, the round-trip time was set equal to seven frames in the simulator. This is reasonable because the video runs at 15 frames per second and the round-trip time of the cell-phone network is estimated to be 500ms. In practice, the round-trip time in frames could be calculated based upon actual network conditions, thus calibrating the recovery to be most effective and efficient.

## 4.2 Simulation Results

The simulation was carried out a video taken on a HTC TyTN-II cell phone. This video is centered on an ASL signer who occupies approximately the center third of the frame (see



(a) Frame Copy

(b) Bursty I-Frame Refresh (+12 QP)

(c) Bursty P-Frame Refresh

Figure 3: Visual comparison of recovery techniques. Figure shows frame number 325 of the test sequence.

| Frame Copy | Bursty I-Frame Refresh | Bursty P-Frame Refresh |
|---|---|---|
| 11.01 dB | 18.50 dB | 18.49 dB |

Table 1: Average PSNR for test sequence.

Figure 3). The phone was set down on the table and the user sometimes moved forward and backward. The resulting background in the video does not change much.

ASL intelligibility is difficult to measure [19]. To substitute for user-studies at this stage of research, we instead measure the peak signal to noise ratio (PSNR), which is a common measure of distortion that results from image compression. The PSNR is calculated for corresponding frames between the simulated video and the source video and a higher PSNR usually indicates a better reconstruction. Although this metric does not directly correspond to the intelligibility of ASL, the average PSNR over all frames of a particular video can indicate the persistence of distortion introduced by data loss. In the simulations, the average PSNR for the decoded video was 7.48dB higher when using our techniques than when frame copy method was used (Table 1).

Bursty I-frame refresh was investigated for two different cases by adjusting the quantization parameter (QP): No QP change and +12 QP which is called adjusted bursty I-frame refresh. Increasing the QP by 12 lowers the PSNR but the corresponding packet bits is smaller, comparable to bursty P-frame refresh (Figure 4).

Figure 5 shows the packet size of a refresh frame for bursty I-frame refresh (no QP change), adjusted bursty I-frame refresh, and bursty P-frame refresh when the second frame is lost. In bursty I-frame refresh, +12 QP has 4 times fewer bits than no QP change, but it still requires more bits than bursty P-frame refresh.

We compare adjusted bursty I-frame refresh with bursty P-frame refresh. From Figure 6 it appears that the methods have the same average PSNRs. It also shows that frame copy provides the same quality until the first loss but it can't recover without receiving a new refresh frame. Bursty P-frame refresh recovered more quickly from a loss than adjusted bursty I-frame refresh. Each method recovered once feedback was received. As indicated earlier, the number of intra-blocks used by a recovery method greatly affects the size in bits of a refresh frame. The intra frame refresh method produces refresh frames that use the

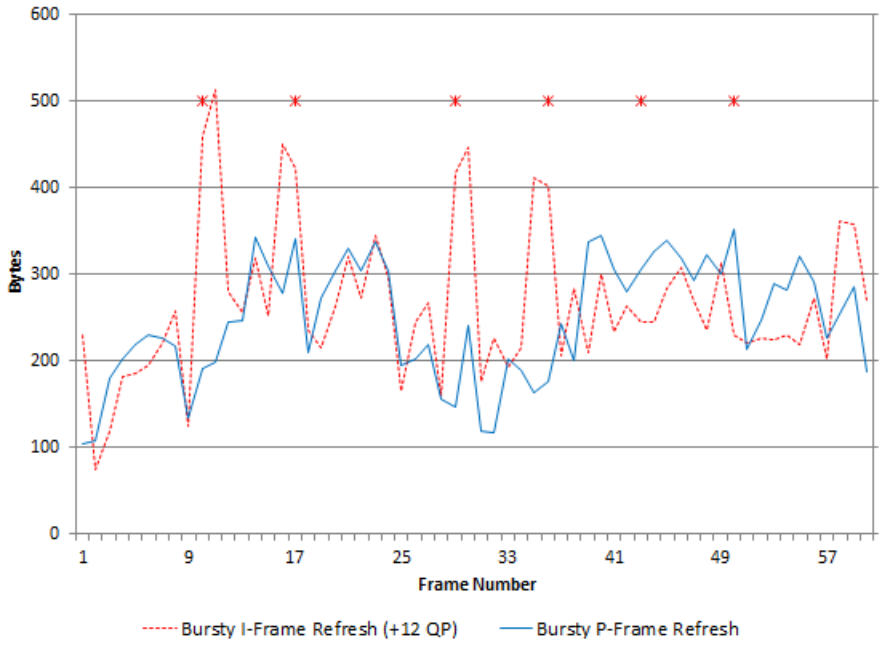



Figure 4: Resulting Packet Size for different recovery techniques on the test sequence. An asterisk indicates the time frames are lost. Note that our target average bit rate is 30 kbps.

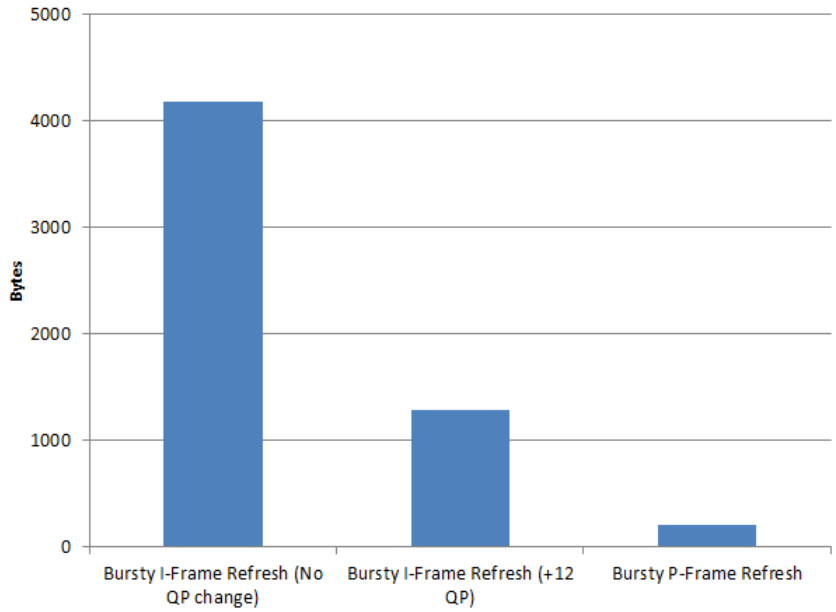



Figure 5: the packet size of a refresh frame for three different cases when the second frame is lost.

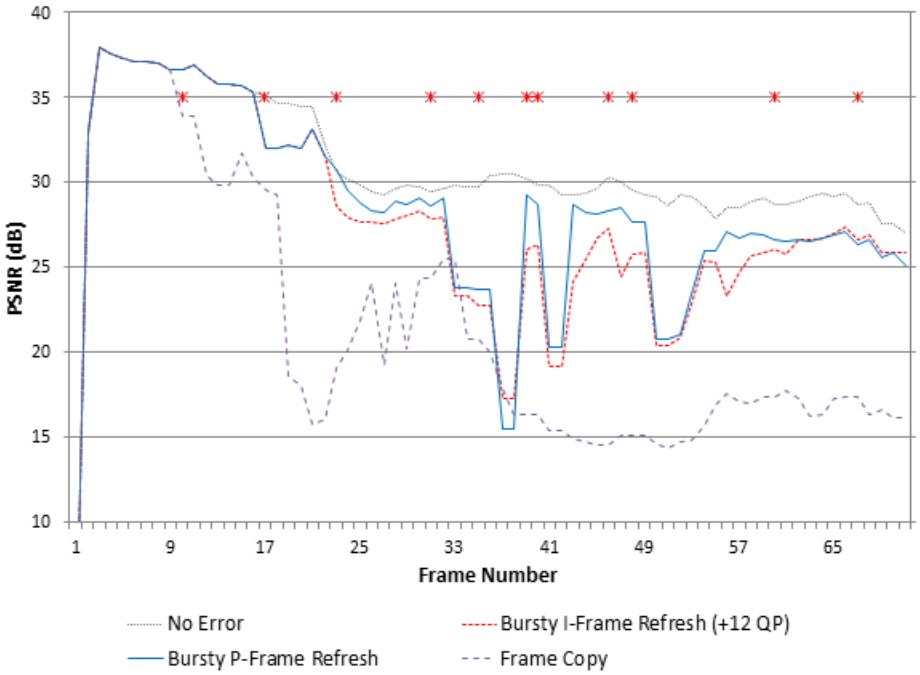



Figure 6: Resulting PSNR for different recovery techniques on the test sequence. An asterisk indicates the time frames are lost. Note that the bursty P-frame refresh recovers more quickly from a loss than adjusted bursty I-frame refresh.

most bits, which is observed in Figure 4.

# 5 Conclusion

We investigated feedback-based recovery schemes for MobileASL. We found that adjusted bursty I-frame method is quite good but it has an undesirable flickering effect because of background changes and larger recovery times. Bursty P-frame refresh is the best algorithm in terms of bit rate and quality because it uses fewer bits and there is no flickering effect. Considering the success of the bursty P-frame refresh in our simulations, we will implement this approach within MobileASL.

# References

[1] Mitchell Ross, "How many deaf people are there in the United States?," 2005. `http://gri.gallaudet.edu/Demographics/deaf-US.php`.

[2] A. Khan, L. Sun, E. Ifeachor, J. Fajardo, F. Liberal, H. Koumaras, "Video Quality Prediction Models Based on Video Content Dynamics for H.264 Video over UMTS Networks," *International Journal of Digital Multimedia Broadcasting*, Feb 2010.

[3] C. Zhu, Y. Wang, M. M. Hannuksela, H. Li, "Error Resilient Video Coding Using Redundant Pictures," *IEEE Transactions on Circuit and Systems for Video Technology*, vol. 19, Jan 2009.

[4] T. Stockhammer, D. Kontopodis, T. Wieg, "Rate-Distortion Optimization for JVT/H.26L Video Coding in Packet Loss Environment," in *Proceedings of the International Packet Video Workshop*, Apr 2002.

[5] S. Wan, E. Izquierdo, "Rate-Distortion Optimized Motion Compensated Prediction for Packet Loss Resilient Video Coding," *IEEE Transactions on Image Processing*, vol. 16, May 2007.

[6] Y. Wang, S. Wenger, J. Wen, A. K. Katsaggelos, "Error Resilient Video Coding Techniques," *IEEE Signal Processing Magazine*, vol. 17, July 2000.

[7] K. Lee, T. Kim, B. Seo, J. Suh, "Fast Reference Frame Selection Algorithm for H.264/AVC Based on Reference Frame Map," 2010.

[8] Y. Chen, K. Yu, J. Li, S. Li, "An Error Concealment Algorithm for Entire Frame Loss in Video Transmission," in *Proceeding of IEEE Picture Coding Symposium (PCS)*, 2004.

[9] B. Yan, H. Gharavi, "A Hybrid Frame Concealment Algorithm for H.264/AVC," *IEEE Transactions on Image Processing*, vol. 19, Jan 2010.

[10] M. Chen, C. Chen, M. Chi, "Temporal Error Concealment Algorithm by Recursive Block-Matching Principle," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, Nov 2005.

[11] Y. Xu, Y. Zhou, "Adaptive temporal error concealment scheme for H.264/AVC video decoder," *IEEE Transactions on Consumer Electronics*, vol. 54, Nov 2008.

[12] Q. Peng, T. Yang, C. Zhu, "Block-Based Temporal Error Concealment for Video Packet Using Motion Vector Extrapolation," in *IEEE Conference on Communications, Circuits and Systems and West Sino Expositions*, June 2002.

[13] S. Whittle, "Maintaining Intelligibility of ASL Video in the Presence of Data Loss," *University of Washington, Computer Science and Engineeing, Senior Thesis*, 2008. `http://www.cs.washington.edu/education/ugrad/current/bestseniortheses/Whittle.pdf`.

[14] Y. J.Liang, J. G. Apostolopoulos, B. Girod, "Analysis of packet loss for compressed video: Does burst-length matter?," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr 2003.

[15] 3GPP, Siemens, "Software simulator for MBMS streaming over UTRAN and GERAN," *document for proposal, TSG System Aspects Working Group4#36, Tdoc S4-050560*, Sep 2005.

[16] N. Tizon, B. Pesquet-Popescu, M. Cagnazzo, "Adaptive video streaming with long term feedbacks," in *16th IEEE International Conference on Image Processing (ICIP)*, 2009.

[17] J. Devadoss, V. Singh, J. Ott, C. Liu, Y. Wang, I. Curcio, "Evaluation of Error Resilience Mechanisms for 3G Conversational Video," in *10th IEEE International Symposium on Multimedia (ISM)*, Dec 2008.

[18] V. Singh, J. Ott, I.D.D Curcio, "Rate Adaptation for Conversational 3G Video," in *IEEE INFOCOM Workshops*, 2009.

[19] F. M. Ciaramello, S. S. Hemami, "'Can you see me now?' An Objective Metric for Predicting Intelligibility of Compressed American Sign Language Video," in *Proceeding of Human Vision and Electronic Imaging(HVEI)*, Jan 2007.