Joint Packet Scheduling and Content-Aware Playout Control for Video Streaming over Wireless Links

Yan Li[†] ⋆, Athina Markopoulou[†], John Apostolopoulos[‡], Nicholas Bambos[†]
[†]Stanford University, [‡]HP Labs

***Abstract*— Media streaming over wireless links is a challenging problem due to both the unreliable, time-varying nature of the wireless channel and the stringent delivery requirements of media traffic. In this paper, we use joint control of packet scheduling at the transmitter and content-aware playout at the receiver, so as to maximize the quality of media streaming over a wireless link. Our contributions are twofold. First, we formulate and study the problem of joint scheduling and playout control within a dynamic programming framework. Second, we propose a novel content-aware playout control, that takes into account the content of a video sequence, and in particular the motion characteristics of different scenes. We find that the joint scheduling and playout control can significantly improve the quality of the received video, at the expense of only a small amount of playout slowdown. Furthermore, thanks to the content-aware playout, the slowdown takes place mainly in the low-motion scenes, where its perceived effect is limited.**

***Index Terms*— Video-Aware Adaptation and Communication, 5: Multimedia Networking (5.a: priority-based QoS control and scheduling, 5.f: wireless communications).**

I. Introduction

Recent advances in video compression and streaming as well as in wireless networking technologies, are rapidly opening up opportunities for media streaming over wireless links. However, the erratic and time-varying nature of a wireless channel is still a serious challenge for the support of high-quality media applications. To deal with these problems, various network-adaptive techniques have been proposed [1], including radio-distortion optimized packet scheduling [2] and/or power control [3] at the transmitter, and playout speed at the receiver [4], [5]. In past work [6], we investigated the joint control of transmission power at the transmitter and playout speed at the receiver, and achieved significant performance gains over individual controls.

In this work, we consider the transmission of pre-stored media units over a wireless channel with time-varying rate. We investigate the joint control of packet scheduling at the transmitter (Tx) and playout speed at the receiver (Rx), so as to overcome the variations of the channel and maximize the perceived video quality. We briefly note the following intuitive tradeoffs faced by the individual controls in the attempt to maximize video quality. At the Tx side, the dilemma is the following: on one hand we want to transmit all media units; one the other hand, during periods that the bandwidth is scarce, we may choose to transmit the most important units and skip some others, depending on their rate-distortion values. At the Rx side, the dilemma is the following: on one hand, we want to display the sequence at the natural frame rate; on the other hand, during bad periods of the channel, we may choose to slowdown the playout in order to avoid late packet arrivals (leading to buffer underflow and frame losses), but at the expense of the annoying slower playout. A novel aspect of our work, is that we perform content-aware playout variation; that is, we take into account the characteristics of a video scene when we adapt the playout speed. The contributions of this work are twofold:

1) We study the joint control of scheduling and playout; we formulate the problem using dynamic programming and explore the tradeoff in quality degradation between distortion vs. playout variation.
2) We introduce the idea of content-aware playout control and demonstrate that it significantly improves the user experience. The idea is to vary the playout speed of scenes, based on the scene content; e.g. scenes with low or no motion may be less affected by playout variation.

The rest of the paper is structured as follows. Section II introduces the model/formulation. Section III provides simulation results. Section IV concludes the paper.

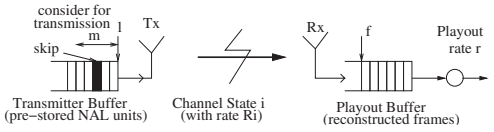

Fig. 1. Joint scheduling and playout control for streaming pre-stored NAL units over a time-varying wireless link.

II. System Model and Problem Formulation

We consider a system shown in Fig. 1, which is comprised of a transmitter (Tx) and a receiver (Rx) communicating over a wireless communication link. Time is slotted. The Tx is equipped with a buffer where the content initially stored. The Rx is equipped with a playout buffer, where received frames are queued up to be played out.

A. Video Source

We use a video sequence pre-encoded using H.264/MPEG-4 AVC [8]. Let N be the total number of frames and $n = 1, ..., N$ be the frame index. Each frame can be further divided into a fixed number, K, of NAL units, i.e. packets for transmission. Let (n, k) be the k^{th} NAL unit in the n^{th} frame, $k = 1, ..., K$. This NAL unit is indexed with $l = (n-1)K + k$, has size b_l (in bits) and leads to a distortion of d_l if not received. To compute d_l, we decode the entire video sequence with this NAL unit missing; this is an approximation as the actual distortion may also depend on the delivery status of prior and subsequent NALs [10]. The distortion model can be extended to capture these loss correlations. Furthermore, we assume that distortions caused by loss of multiple NAL units are additive, which is a reasonable assumption for sparse losses.

Special Session on Content-Aware Video Coding, Adaptation, and Communication. Organizers: Z. Li, A. Katsaggelos
Emails: liyan@stanford.edu (⋆correspondence author), amarko@stanfordalumni.org, japos@hpl.hp.com, bambos@stanford.edu

A video sequence can consist of several scenes. Each scene s contains a group of video frames. Each scene has a different amount of motion, which we are interested in characterizing so that we can later take into account in our content-aware playout control. Finding the appropriate metric to characterize the amount of motion in a scene is an open research problem. In this paper, we define the *motion intensity* M_s as the sum of the absolute values of all the motion vectors in the scene s averaged over the number of frames in the scene. We found that this heuristic captures well the motion of standard scenes (see the discussion in Fig. 2). However, our formulation can incorporate more sophisticated metrics if they become available.

B. Wireless Channel

We model the wireless channel as a Markov chain with transition probabilities q_{ij}; in channel state i, the bandwidth available to the video stream is R_i. The reasoning behind this model is the following. Fast fading, slow shadowing, path loss and interferences all affect the signal to interference/noise ratio (SIR); in turn, the SIR dictates the physical transmission rate and packet error rate, thus the channel throughput. Assuming that the physical and MAC layers use coding and retransmissions to combat channel variations, the wireless channel appears to the application layer as error-free but with time-varying throughput. The throughput R_i in state i can be calculated as $R_i^p(1-PER)$, where the R_i^p is the physical channel rate after the coding, and PER is the packet error rate with the corresponding codes; this is reasonable if the channel varies slower than a packet's transmission time, which is the case for low mobility or in-home environments.

C. Transmission Control and Costs

We assume that all N frames of the sequence reside at the Tx. This is a realistic assumption, when the media server/proxy is co-located with the Tx or the path between the server and the Tx is not the bottleneck. Let l be the NAL unit at the head of the Tx. In this baseline model, transmission happens always in-order, the skipped units are dropped and the remaining units at the Tx have no gaps. Due to space limitations, we omit the extension of the model that allows for out-of-order transmission and skipped NALs to be considered for later transmissions.

For the rest of the paper, the term *time slot* refers to the time period over which we adjust the transmission rate (by choosing how many units to transmit). At each time slot, the Tx considers the next m units for transmission and advances the transmission index from l to $l+m$. From the considered m units, some are dropped to conform to the channel throughput R_i; which units to drop are chosen so as to minimize the total distortion $D_{tx}(m, R_i, l)$:

$$\min \sum_{k\in\Theta} d_k \quad \text{subject to} \sum_{k=l,k\notin\Theta}^{l+m-1} b_k \leq R_i, \qquad (1)$$

where $\Theta \subseteq \{l, l+1, ..., l+m-1\}$ is the set of NAL indices to be dropped. This minimization can be solved by a greedy algorithm with each NAL ranked by its distortion-to-size ratio, d_k/b_k.

The control parameter m is chosen from a range of possible values. Large values of m advance the index further (thus helping playout) but may drop more units (thus introducing more distortion). We assume that for all practical cases, no more than 50% of units should be dropped, and we limit the range to $m \in [m_l, 2m_l]$, where m_l is the maximum number of consecutive units that can be transmitted without exceeding the channel rate R_i:

$$m_l = \arg\max \sum_{k=l}^{l+m_l-1} b_k \leq R_i \qquad (2)$$

D. Content-Dependent Playout and Costs

Let f be the number of fully decodable frames at the Rx, i.e. frames whose all NAL units have already been received or dropped at the Tx side; because of the in-order transmission, the units transmitted in the future will belong to subsequent frames. Note that when some NAL units are missing, the distortion has already been captured by the cost D_{tx} at the Tx. We constrain $f \leq F$ to capture the physical buffer size.[1]

Adaptation Range and Timescale. The Rx can control the value of the playout rate $r \in \{r_1, r_2, ..., r_n\}$, where $r_1 < r_2 < ... < r_n$ and r_n is the video sequence normal rate (say 30fps). New packets arrive every time slot, but we adapt r more infrequently, say every T timeslots, in order to avoid noticeable perceived effects of rapid playout variations. Similarly, to avoid large magnitude variations, we constrain r to increase or decrease only by one level, say from the previous r_k to $r \in \{r_{k+1}, r_k, r_{k-1}\}$; however, at the scene boundaries, r can take any value in $\{r_1, r_2, ..., r_n\}$. When adapting the rate of a group of frames that span two scenes, we assign the group to the scene where most frames in the group belong to.

Removing units from the Rx. Let t track the timeslot within a cycle of T timeslots: $t = 1, 2, ...T$. etc. In the first timeslot of a cycle ($t = 1$) the control chooses a new value for r, to use for the entire cycle. At every timeslot t, we remove and display $r^t = \lceil \frac{rt}{T} \rceil - \lceil \frac{r(t-1)}{T} \rceil$ frames from the buffer. This reduces the number of full frames in the Rx, f, by $(f - r^t)^+$.

There may be timeslots, when the playout control chooses to remove more frames than the currently available in the buffer, $r^t > f$. Then, the Tx is notified to drop the NALs that miss their deadlines, and the Tx continues with subsequent units. Then this leads to an additional distortion cost $D_{rx}(r, f, l) = \sum_{k=l}^{(f_e+(r^t-f))K} d_k$, where $f_e = \lfloor (l-1)/K \rfloor$ be the frame index of last fully decodeable frame at the Rx buffer. At the Tx side, the index l is updated to $(f_e + (r^t - f))K + 1$.

Units arriving to the Rx buffer. At each time slot, packets arrive at the Rx. We assume a store-forward operation where packets that arrived in the current time slot are not available for display at the same time slot. This is a conservative assumption, as some packets may arrive before the end of the timeslot; alternatively, appropriate channel models could account for the packet arrival distributions. Taking into account both arrivals and playout, in every

[1] For small values of F, the benefit from the control is amplified; in general though, we expect memory to be cheap and thus F to be large enough and have a negligible effect on performance.

time slot, the new NAL index l' at the Tx and new receiver buffer level f' are updated as the following:

$$l' = \begin{cases} (\lfloor (l-1)/K \rfloor + (r^t - f))K + 1 + m & , r^t > f \\ l + m & , r^t \leq f \end{cases}$$
$$f' = (f - r^t)^+ + (\lfloor l'/K \rfloor - \lfloor l/K \rfloor) \tag{3}$$

Playout Costs. Choosing slower playout extends the playout deadlines of the NAL units in transmission (thus reducing distortion due to dropping late units) but may also produce an annoying perceived effect. This effect is scene-dependent. For example, playout speed variations are more perceptible in scenes with significant or constant motion (e.g. a camera pan) rather than in motionless scenes. To capture this effect we introduce the following two costs:

- Let $C_s = g_1(r, M_s)$ be the slowdown cost due to playing slower than the natural rate; M_s is the motion intensity of the scene s the current r frames belong to. If the r frames cross the scene boundary, we take a weighted average of the two costs in the two scenes. The function g_1 should be increasing with M_s and decreasing with r. In this paper, we use the simple function: $C_s = M_s(r_n - r)$.
- Let $C_v = g_2(r, \vec{r})$ be the playout variation cost, due to variations of r from one period to the next. The vector $\vec{r}$ records the past L playout rates and is reset at scene boundaries. The function g_2 should be decreasing in r and increasing in M_s. In this paper, we use the simple function: $C_v = |r - r_{last}|$; i.e. we ignore the effect of M_s (already accounted for in C_s) and we consider only the last chosen r_{last} instead of a longer history $\vec{r}$.

These costs extend the ones proposed in [5], [6] by including the motion intensity of a scene. However, our problem formulation is general enough to also incorporate any new and improved perceptual costs.

E. System State and Optimal Control

The state of the system is $(l, i, f, \vec{r}, t)$; l is the unit at the head of the Tx; i is the state of the channel (leading to rate R_i); f is the state at the Rx and $\vec{r}$ is the playout history; finally, $t \in \{1, ..T\}$ tracks whether we can adjust the playout rate in the current time slot (true for $t = 1$). The controls exercised, (m, r), are subject to the constraints described in the previous subsection. The associated costs are: C_s, C_v for the playout slowdown and variation costs; and D_{tx}, D_{rx} for the distortion cost due to packets dropped at the Tx, to meet transmission rate constraints (D_{tx}) or because they missed their playout deadlines (D_{rx}). The system becomes a controlled Markov chain and the optimal control can be computed from the dynamic programming equations. Let $J(.)$ be the optimal cost to go.

In $t = 1$, we control both playout and transmission:

$$J(l, i, f; \vec{r}, t = 1) = \min_{m,r} \{C_s + C_v + w(D_{tx} + D_{rx}) + \sum_{j \in \mathcal{I}} q_{ij} J(l', j, f'; \vec{r}', t + 1)\} \tag{4}$$

When $t \neq 1$, only transmission control is active:

TABLE I
TEST SEQUENCE

Frame Numbers in Test Sequence	Original Video Sequence	Frame Numbers Original Sequence
1-60	mother-daughter	101-160
61-120	carphone	171-230
121-180	grandma	1-60
181-240	foreman	271-330
241-300	mother-daughter	391-450
301-360	carphone	281-340
361-420	grandma	61-120
421-480	suzie	31-90
481-540	mother-daughter	901-960
541-600	foreman	144- 1990

$$J(l, i, f; \vec{r}, t \neq 1) = \min_{m} \{w(D_{tx} + D_{rx}) + \sum_{j} q_{ij} J(l', j, f'; \vec{r}, t + 1)\} \tag{5}$$

w is the weighting factor used to explore the performance trade-off between video quality and playout variations. In general, there may be additional weighting factors to stress C_s vs. C_v, and also D_{tx} vs. D_{rx}.

After all NAL units are transmitted, $\{l, i, t\}$ and m are removed from the state and control respectively; then, the Rx gradually increases the playout rate (adjusting upwards every T time slots) and plays out the remaining frames at the natural rate r_n:

$$J(f; \vec{r}) = \min_{r} \{C_s + C_v + J(f - r; \vec{r}')\} \tag{6}$$

The system terminates when all the frames are played out: $f \leq 0$.

III. SIMULATION

A. Simulation Setup

We used the JM8.6 version of the H.264/MPEG-4 AVC *codec* [8], [9]. We simulated *packet loss* by erasing the corresponding NAL units from the RTP stream produced by the encoder. At the receiver side, we decoded the remaining RTP stream with error concealment enabled. In case when an entire frame is lost, we had to implement copy-concealment, (which was not supported in JM 8.6). The video sequences were QCIF at 30fps, encoded using only I and P frames (one I every 10 frames), and packetized using 3 slices/frame and 33MB/slice. The PSNR of the encoded sequence is 36.5dB.

Our *test sequence* is shown in Table I. We concatenated scenes from various standard sequences, which exhibit different degrees of motion. Fig. 2(a) shows the motion intensity M of the resulting test sequence. Recall that M is defined as the weighted sum of the absolute motion vectors in each P-frame; for I-frames, $M = 0$. One can see that our heuristically defined M successfully captures the motion characteristics of these well known scenes. E.g. scene 4 corresponds to the camera pan in Foreman and has the highest M; the scenes from Grandma and Mother-Daughter have the lowest M.

The parameters for the *wireless channel*, are chosen to demonstrate key features of our approach. The rate in the good and bad state was 262 Kbps and 74 Kbps respectively. This results in an average channel rate slightly larger than the average video rate (162 Kbps). The transition probabilities ([0.67 0.33; 0.33 0.67]) were chosen to lead to average state durations around 0.5sec, comparable to

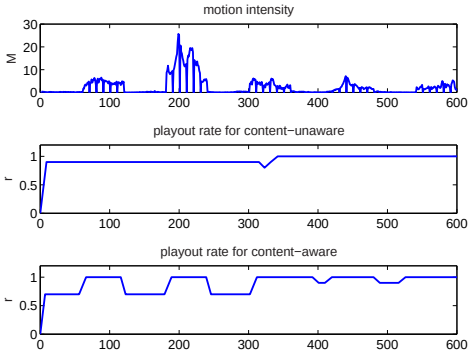

Fig. 2. Motion intensity of the test sequence (a) and Playout rate (as % of the natural rate) without (b) or with (c) motion-awareness.

the coherence time in home or low-mobility environments. The *time slot* (for transmission and reception of a group of packets) is chosen to be 5 frame durations (33ms each), i.e. 0.167sec, to allow for a reasonable number of NALs to be transmitted together. The playout rate is adjusted every T=10 frames.

B. Simulation Results

Fig. 2(b) and 2(c) show the playout rate (normalized w.r.t. the natural playout rate) across frames of the test sequence without and with content-awareness, respectively. The distortion (due to dropped packets) is the same in both cases. The main observation is that, as expected, the content-aware control chooses to slow down more the low motion scenes and leave the high-motion scenes intact; this reduces the perceived effect of slowdown. A secondary observation is that both controls increase the playout rate in the last 180 frames, because buffer underflow is less risky at the end of the sequence.

Fig. 3 shows the tradeoff between % increase in the total playout duration due to slowdown and increase in video quality (PSNR of the decoded sequence), when content-unaware playout is used. The curve is obtained by varying the distortion weight w between $D_{tx} + D_{rx}$ and $C_s + C_v$. By using only the control at the transmitter (i.e. 0% increase in duration) to carefully select the right NAL units for transmission, we observe a 6dB gain over the no-control case. Furthermore, the video quality increases approximately by 2dB for every 10% of playout duration increase, and saturates at the PSNR of the encoded sequence. The most similar work that we are aware of is [4], where the effect of pre-roll delay on quality is studied using a different methodology. In general, the curve in Fig. 3, should depend on the wireless channel.

In Fig. 3, we characterized the effect of playout slowdown in terms of an objective metric (% increase in total duration) but did not take into account the video content. Fig. 4 shows again the tradeoff between video quality and playout variation. The lower curve corresponds to the content-unaware control of Fig. 3, but the effect of playout is now shown in terms of playout cost ($C_s + C_v$) instead of duration increase. Note that better models of perceptual cost of scene-dependent playout slowdown can be easily incorporated in the proposed framework. By using content-aware playout, the tradeoff improves (the curve moves higher and to the left); i.e. for the same

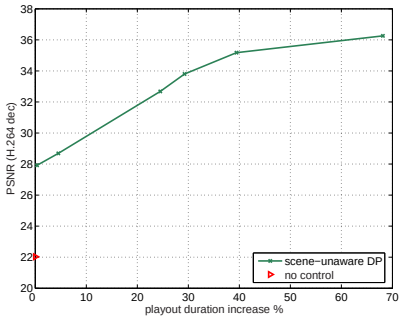

Fig. 3. Video Quality (PSNR) vs. % increase in playout duration.

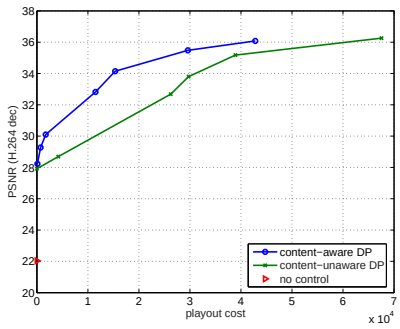

Fig. 4. Tradeoff between Video Quality and Playout Cost, for joint control with/without content-awareness.

distortion, the playout variation has a smaller perceived effect, thanks to the intelligent selection of the preferred scenes for performing the slowdown.

IV. Conclusion

In this paper, we studied the joint control of packet scheduling at the Tx and playout control at the Rx, for video streaming over a time-varying wireless channel; we show that a small increase in playout duration can result in a significant increase in video quality. Furthermore, we proposed to take into account the characteristics, and in particular the motion intensity, of a video sequence in order to adapt the playout control based on the characteristics of each scene in the video sequence; this reduces the perceived effect of playout speed variation for the same increase in video quality.

References

[1] B. Girod, J. Chakareski, M. Kalman, Y.J. Liang, E. Setton and R. Zhang, “Advances in Network-adaptive Video Streaming”, *in Proc. of IWDC 2002*, Sept. 2002, Capri, Italy.

[2] P. Chou and Z. Miao, “Rate-distortion optimized streaming of packetized media,” *submitted to IEEE Trans. on Multimedia*, Feb. 2001.

[3] Y. Li and N. Bambos, “Power-controlled streaming in interference limited wireless networks”, *in Proc. of IEEE BroadNets*.

[4] M. Kalman, E. Steinbach, and B. Girod, “Adaptive Media Playout for Low Delay Video Streaming over Error-Prone Channels,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, no. 6, pp. 841 - 851, June 2004.

[5] M. Kalman, E. Steinbach, and B. Girod, “Rate-distortion optimized video streaming with adaptive playout,” in *IEEE ICIP-2002*, vol. 3, pp. 189-192, Sept. 2002.

[6] Y. Li, A. Markopoulou, N. Bambos, J. Apostolopoulos, “Joint Power-Playout Control Schemes for Media Streaming over Wireless Links,” *in Proc. of IEEE Packet Video*, Dec. 2004, Irvine, CA.

[7] D. Bertsekas, “Dynamic Programming and Optimal Control”, vol. 2, Athena Scientific, 1995.

[8] ITU H.264/ISO MPEG-4 AVC Recommendation

[9] H.264/AVC reference software version: JM 8.6 *http://iphome.hhi.de/suehring/tml/index.htm*

[10] Y.J. Liang, J. Apostolopoulos, and B. Girod, “Analysis of Packet Loss for Compressed Video: Effect of Burst Losses and Correlation Between Error Frames,” *submitted to IEEE Trans. on Multimedia*.