Discerning User-Perceived Media Stream Quality Through Application-Layer Measurements

Amy Csizmar Dalal and Keith Purrington
Department of Mathematics and Computer Science
Carleton College
Northfield, Minnesota 55057
Email: {adalal,purringk}@carleton.edu

***Abstract*— The design of access networks for proper support of multimedia applications requires an understanding of how the conditions of the underlying network (packet loss and delays, for instance) affect the performance of a media stream. In particular, network congestion can affect the *user-perceived quality* of a media stream. By choosing metrics that indicate and/or predict the quality ranking that a user would assign to a media stream, we can deduce the performance of a media stream without polling users directly. We describe a measurement mechanism utilizing objective measurements taken from a media player application that strongly correlate with user rankings of stream quality. Experimental results demonstrate the viability of the chosen metrics as predictors or indicators of user quality rankings, and suggest a new mechanism for evaluating the present and future quality of a media stream.**

I. Introduction and motivation

The design of access networks to accommodate streaming media applications requires an understanding of the performance of streaming media applications. In particular, understanding how network congestion affects the performance of multimedia applications is important in guiding future development of protocols and future deployment of network and server resources to best serve the multimedia application clients.

From an application perspective, we are most interested in the *quality level* of a received media stream. In this context, we consider quality in terms of the user-perceived quality of a media stream: i.e., how a user would rank the stream relative to other media streams that he or she has seen in the past. For the purposes of this study, we are most interested in coarse-grained levels of quality: whether a user would classify a media stream's quality as "good" (the user has few complaints about the stream's quality), "acceptable" (the user wishes the quality were better, but will continue to watch the stream anyway), or "poor" (the user considers the stream to be unwatchable, and may terminate the session prematurely as a result).

The most unambiguous way to discern what a user thinks about the quality of a media stream is to ask the user directly. The five-point scale, or Mean Opinion Score (MOS) [1], was developed for this purpose: to allow users to provide feedback on the subjective quality of a media stream through a ranking mechanism, with a score of 1 indicating poor quality and a score of 5 indicating exceptional quality. The use of the MOS has several drawbacks in practice. It does not scale well to a large number of users. More importantly, the MOS gives us limited information about the reasons behind the quality ranking of a stream. For instance, a ranking of "3" may mean that the user did not care for the encoding of the stream, or it may mean that there was a period of network congestion that affected the timely delivery of the stream to the user.

A more tractable solution involves collecting measurements from the underlying network and/or the media player application, and discerning the user's quality rankings based on these measurements. Identifying the proper metrics is a challenge: of the different types of data we could collect—packet loss, throughput, frame rate—which of these tells us the most about a stream's quality? More importantly, what analysis can we apply to this data to provide a picture of stream quality?

By taking measurements as close to the user as possible—at the application layer—and by judiciously choosing our metrics, we can develop a picture of streaming media quality in a more scalable and accurate fashion than by polling the user. In addition, we can combine these application-layer measurements with network-level measurements to form a complete picture of the rationale behind a quality ranking: we can deduce that the quality of a stream was "poor" because the network packet loss rate was 20% on average during the stream. In turn, by understanding the relationship between user quality ranking and network congestion levels, we can make better-informed decisions about the design of the underlying networks and the design of the application-layer streaming protocols.

Recent work by others has also attempted to address the issue of using objective metrics as a substitute for subjective metrics of stream quality. Commercial tools, such as [2]–[5], rely on synthetic test streams, synthetic applications, and/or arbitrary test points in the network to discern the quality of a media stream from objective measurements. These tools, however, can fail to detect application sensitivities to service quality, such as stream start-up delay or player stall, that are of relevance to actual media clients. Media player applications such as Windows Media Player [6] and RealPlayer [7] provide some limited feedback mechanisms, but only to the origin server. A more mathematical approach, in which measurements of a stream on both the sender and receiver sides are correlated, is presented in [8] and [9].

Our proposed measurement mechanism involves the use of an instrumented media player application, similar to the approaches presented in [10]–[12]. Our goal is to demonstrate not only the feasibility of using application-layer measurements to discern user-perceived quality of an on-demand stream, but also to demonstrate that certain metrics may serve as *predictors* of future periods of degraded stream quality. Thus it is possible to utilize this information to design strategies to mitigate the actual reduction in stream quality (for instance, by selecting a different media server with a better quality profile). We discuss this idea more fully in Section II.

Following our derivation of the proper subset of metrics to use, we describe, in Section III a series of experiments designed to validate our selection of metrics. The results of these experiments are presented in Section IV. Finally, we conclude in Section V with some thoughts about future work in this area.

II. Objective metrics for measuring subjective stream quality

Identifying metrics for discerning the subjective quality of a media stream is a challenging problem. Collecting measurements at the network level gives us a clear picture of the current congestion conditions, either on a global scale or on a particular network segment. Examples of network-level metrics include the percentage of lost packets, packet delays, and throughput. However, network-level metrics may not clearly indicate the user's perception of or experience with a media stream. In particular, streaming media applications often employ mechanisms, such as aggressive retransmissions or temporarily increasing the transmission rate to several times higher than the normal transmission rate ([13]), to mitigate the effects of network congestion. Another alternative, then, is to take measurements *from the media player application* by periodically polling this application. Measurements that can be pulled from the media player include statistics about the *application-layer* packets (number of application packets received correctly, lost, or retransmitted), information about the current application-layer throughput, and information about the frequency and duration of start-up and mid-stream buffering.

In [14], we describe a process by which to identify a good set of application-level metrics. Based on this process, we have identified four key metrics that characterize the user-perceived quality of a media stream.

The first metric is the number of "lost" application-layer packets. In this context, an application-layer packet is lost if it does not arrive before its scheduled play-out time.

The second metric of interest is the number of packets that the player reports as being "retransmitted". These are the application-layer packets that did not arrive successfully within a specified time window, but which did arrive before the scheduled play-out time. Retransmitted packets serve as a good first-order approximation of the degree of network congestion: the number of retransmission requests increases with increasing packet loss and packet delays at the network layer.

The third metric is the amount of time over which no new application-layer packets arrive: in other words, when the application has not received any new data from the network layer. We refer to this metric as "packet reception pauses".

Finally, we consider the "buffering behavior" of the media player. Specifically, we are interested in how often the player enters a *buffer starvation* period mid-stream, where it has no reserve data to render and must wait for new data to arrive, and the duration of each of these events. We are also interested, although to a lesser extent, in the duration of the *startup buffering* period, or the length of time at the start of the stream between the arrival of the first stream data packet and the first frame rendered by the player.

Of these four metrics, two are what we term *lagging indicators* of stream quality, and two are *leading indicators* of stream quality. The lagging indicators are so termed because they indicate the exact moment at which the quality of a media stream decreases. The two lagging indicators in this study are lost packets and buffer starvation. Lost packets often manifest themselves as uneven transitions between frames or other video artifacts. More rarely, they may manifest themselves as audio glitches, but this is only in cases of severe loss since video data tends to be dropped before audio data. Buffer starvation periods correspond to either a partial or complete stoppage of the stream.

Packet reception pause periods and packet retransmissions, on the other hand, are leading indicators of reduced stream quality because their appearance typically portends the future occurrence of either lost packets or a buffer starvation period. The longer the packet reception pause period (or the more frequently the packet reception pauses occur in succession), the greater the likelihood of the occurrence of a buffer starvation period or a packet loss period. Similarly, an increase in packet retransmissions may indicate that packet loss is imminent.

III. Experimental testbed and procedure

Evaluating the viability of these metrics as replacements for subjective quality measurements requires us to compare these metrics head-to-head with user rankings of the same streams under the same network congestion conditions. To do so, we need a mechanism for collecting application-layer measurements and a mechanism for collecting user quality rankings of media streams. We describe both mechanisms in this section.

A. Data collection infrastructure

Collecting data at the application layer requires us to poll the application for current information about a stream. We have developed a measurement tool ([15]) that leverages the existing installed media player software on a user's machine, without requiring the modification of the media player. Our tool consists of a plug-in that interfaces directly to the installed media player on the client—in this case, Windows Media Player. The plug-in uses ActiveX hooks to query the media

player at uniform intervals about the current state of a stream. This data is logged for later off-line analysis.

B. Network testbed infrastructure

The experimental network consists of a set of 25 client machines on a subnet of a small campus network, and a media server on a separate, isolated subnet. The media server is separated from the rest of the campus network by two routers. The router closest to the media server runs NIST Net ([16]), software which allows one to introduce a known amount of network congestion (packet loss, packet delays, or bandwidth throttling) onto a network. By introducing a known amount of congestion, we can observe how the media players react to perturbations in the network, and how these network disturbances are reflected in the measurements we collect from the media players.

The media server is a 2.4 GHz Pentium processor machine with 512 MB of RAM, running Windows Server 2003 and Windows Media Server 2003 software. The NIST Net router is a 700 MHz processor machine with 512 MB of RAM, running Linux kernel 2.4.21-27, and NIST Net version 2.0.12. The client machines have 2.4 GHz Pentium processors and 512 MB of RAM and run Windows XP SP1 and Windows Media Player version 9.

Before and during the experiment, we took periodic measurements on the campus network of network packet loss rates, network packet delays, and throughput. Based on these measurements, we found negligible levels of packet loss and delay on the campus network. Thus, it was not necessary to isolate the client machines in addition to isolating the media server.

C. The media streams

Table I lists the audio/video streams used in this study and the network congestion levels used in the experiments. The streams were selected to provide some variety in their length, style, and amount of action. Selecting the loss rates to achieve the desired level of degraded stream quality proved difficult, due in large part to the mechanisms that Windows Media Player utilizes to mitigate the effects of network congestion. We in some cases had to increase the loss rate beyond the parameters of what would be considered "acceptable" network loss rates to overcome Windows Media Player's mechanisms and achieve the desired loss of quality in the streams. Table II describes the categories of loss that were used in the experiments, by qualifying what "mild", "moderate", and "severe" loss "looked" like to the user. Network congestion manifested itself in each stream differently, but the table shows some similarities in terms of the stream behavior and the number of "poor quality" characteristics in each stream within the loss categories.

D. Experimental design

To verify the utility of our metrics in predicting user-perceived quality of a stream, we conducted a series of experiments in which users ranked the quality of media streams under no loss and under the various loss levels described in Section III-C. In the first part of the experiment, we showed a "training clip": a one-minute, thirty-second stream of a news story. The training clip introduced the participants to the baseline level of quality that they could expect from the subsequent streams, under "perfect" conditions (i.e., no network loss or delays).

We then showed the users a series of six streams, selected from the streams listed in Table I. Each user viewed each stream twice: once with no loss and once with either mild, moderate, or severe levels of loss. In total, each user saw three clips with no loss and one clip each at mild, moderate, and severe loss levels. The loss patterns were randomized, so that not all users saw the same level of loss at any given time. The users were not aware of the loss levels shown to them for a particular stream; they also were not told which version of the stream had no loss introduced. The participants then filled out a brief survey in which they ranked the audio, video, and the overall quality of the stream on a seven-point scale, with one being the worst possible quality and seven being the best possible quality. The users were also asked to elaborate on their rankings in each category: why they gave the score that they did. While this part of the survey was optional, we found that all of the participants did in fact provide comments for most or all of their rankings.

The participants viewed the streams using Windows Media Player and our measurement tool. The plug-in collected state information from each stream as the users watched the streams.

We ran these experiments on two separate occasions: one day each in August and October, 2004. From these experiments, we obtained a total of seventeen sets of user rankings. We randomized the network congestion patterns introduced to each user such that at least two users saw each combination of loss patterns: i.e., at least two users saw the exact same loss levels for the exact same streams.

IV. Results and analysis

In this section, we examine the relationships between the metrics described in Section II and the user quality rankings. We determine how closely our selected metrics align with the users' evaluation of the quality of a media stream. Strong correlations between the metrics and the rankings indicate that the metric is a good indicator of user-perceived stream quality, and that the metric in fact can substitute for the user's quality ranking.

Figure 1 illustrates the distribution of the audio, video, and overall quality rankings given for all streams within a loss category (mild, moderate, severe, none). Quality rankings generally decrease as the level of network congestion increases. It is interesting to note that the difference between quality rankings for streams with "moderate" and "severe" loss levels is low. We suspect that in our sample, users thought that the stream with moderate loss was already "poor". Given that the scores skew toward the lower end of the scale, users classify stream quality as "poor" when network-level packet loss reaches 15%.

TABLE I

DESCRIPTION OF TEST STREAMS AND CONGESTION PATTERNS

Stream information			Network congestion level		
Name	**Duration (min:sec)**	**Description**	**Mild**	**Moderate**	**Severe**
Commercial	0:30	Commercial Moderate action	50 ms delay 4% loss	80 ms delay 16% loss	175 ms delay 25% loss
News	4:09	News story Low-Moderate action	50 ms delay 4% loss	60 ms delay 11% loss	80 ms delay 25% loss
Trailer	2:22	Animated move trailer High action	100 ms delay 8% loss	200 ms delay 10% loss	200 ms delay 19% loss

TABLE II

SUBJECTIVE QUALITY DESCRIPTION FOR EACH TEST STREAM

Stream	Mild Loss	Moderate Loss	Severe Loss
Commercial	1-2 sec periods of video "stuttering" at end of stream.	Loss of video motion at start and near end; more pronounced during scene changes. Buffer starvation near end of stream.	"Slide show" effect. Playback ends before all data is rendered.
News	Few 2-6 sec periods of loss of video motion.	4-10 sec periods of video motion loss. Possible buffer starvation mid-stream	More time not playing video than playing video.
Trailer	1-2 sec (isolated) periods of video "stuttering".	Frequent periods of 1-2 sec video stuttering, some loss of video motion.	"Slide show" effect. Long periods of buffer starvation.

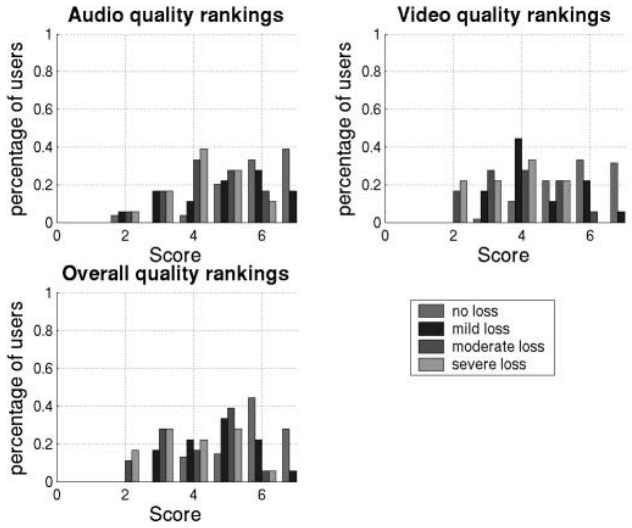

Fig. 1. Histograms of user quality rankings for all viewed streams

To determine the extent to which each of the metrics corresponds to users' quality rankings, we calculate the correlations between the rankings and each of the metrics of interest. Table III lists the degree of correlation between the user rankings and each of the metrics of interest. By separating the data for each clip, we can isolate the loss rate as the primary factor affecting both our metrics and the user-reported scores. The negative correlations indicate that the users' quality rankings decrease as these metrics increase.

Except for the movie trailer clip, the highest correlations between user ranking and quality metrics occur for the video rankings, and the lowest correlations occur for the audio rankings. This indicates two things. First, that a user's perception of stream quality depends more highly on the video quality than on the audio quality. Second, that the video quality of a stream is more highly affected by network congestion than the audio quality is. This intuitively makes sense: a larger portion of a stream is devoted to video than to audio packets, so network-level congestion will affect a larger number of video packets than audio packets. Also, video packets are the first to be "pruned" by the server under periods of congestion.

We can also see from Table III that the degree to which network congestion affects the quality of the clip depends both on the level of congestion and on the nature of the clip. In longer streams, the effect of packet loss and packet reception pauses may be amortized over the length of the stream: a two-second period of loss in a three-minute stream will not be quite as noticeable as a two-second period of loss in a 30- second stream. Also, the same percentage of lost packets spread over a three-minute stream will not be quite as noticeable as in a 30-second stream: users tend to "forget" loss periods if they happened early on in the stream, or if most of the stream was of acceptable quality.

It is interesting to note that the correlations for retransmitted packets are actually higher than those for packets reported by the player as lost, in all cases. In only two cases is the correlation coefficient less than -0.5. This indicates that "recovered packets" are a stronger indicator of user-perceived stream quality (and possibly of network-layer packet loss, or "actual packet loss") than the percentage of packets lost at the application layer.

TABLE III

CORRELATION COEFFICIENTS BETWEEN USER QUALITY SCORES AND QUALITY-CONTROL METRICS

	Commercial			**News**			**Trailer**		
	Audio	Video	Overall	Audio	Video	Overall	Audio	Video	Overall
Packet reception pause duration	-0.4864	-0.7019	-0.6493	-0.2223	-0.4135	-0.3259	-0.3836	-0.3573	-0.4216
Lost packets	-0.4373	-0.7044	-0.6817	-0.3011	-0.5540	-0.4928	-0.5655	-0.4660	-0.5241
Retransmissions	-0.5886	-0.7452	-0.6998	-0.3855	-0.6002	-0.5614	-0.6035	-0.4958	-0.5295
Startup buffering	-0.4437	-0.5705	-0.5037	-0.2611	-0.4616	-0.4461	-0.4863	-0.4506	-0.4508

While our original intent was to examine mid-stream *buffer starvation* periods, we found that there were too few of these events in our datasets to analyze. We instead analyze the correlation between the duration of the startup, or initial, buffering period of a stream and the user's quality ranking of that stream. The data shows that the duration of the startup buffering period is for the most part strongly correlated with user quality rankings. The correlation between video quality rankings and startup buffering duration is higher than the correlation between audio quality rankings and startup buffering duration, as it is with the other metrics. Thus, even though originally we were not apt to consider startup buffering as an indicator of degraded stream quality, we plan on including this metric in our future studies due to its strong showing here.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have demonstrated that objectively-measured metrics can be used as approximations of the quality score that a user would give a media stream under network congestion conditions. In particular, the degree of application-layer packet loss, the number of packet retransmission requests at the application level, the amount of time in which no new application-layer packets are received, and the duration of the startup buffering period, are all strong indicators of the user-perceived quality of a media stream, if we consider coarsely-quantified quality levels.

Because we can approximate user rankings with carefully-chosen metrics, we can discern the quality of a media stream merely by taking measurements from the media player application. It is not necessary to poll the users directly. Taking measurements rather than polling is a much more scalable solution, and is potentially much less ambiguous than user rankings can be.

Several metrics have previously been identified as "leading" indicators of stream quality: observing these metrics and then observing the stream demonstrates that the metrics provide clues as to future quality levels of the stream, by indicating warning signs of increasing network congestion levels. By exploiting these leading indicators, we can potentially *predict* future quality levels of a stream, without having to measure what is happening at the network layer. Such predictions may lead to new models for allocation of network and server resources for streaming applications: distributing media servers and/or media content closer to the users, for instance, or developing mechanisms to allow for smooth transitions from one media server to another in mid-stream.

ACKNOWLEDGMENTS

The authors would like to thank Hewlett-Packard Laboratories for sponsoring earlier stages of this research. We would also like to thank Michael Tie for his technical assistance and Benjamin Sowell for assisting us with the data collection and analysis. Finally, we would like to express our appreciation for the Carleton College students and staff who participated in the user testing experiments for this paper.

REFERENCES

[1] ITU-T Recommendation P.910, "Subjective video quality assessment methdods for multimedia applications," Recommendations of the ITU, Telecommunications Sector.

[2] "NetIQ's Chariot software," `http://www.netiq.com`.

[3] "Broadstream," `http://www.broadstream.com`.

[4] "Streamcheck," `http://www.streamcheck.com`.

[5] "Keynote Streaming Perspective," `http://www.keynote.com/solutions/html/streaming_perspective1.html`.

[6] "Windows Media Player," `http://www.microsoft.com/windows/windowsmedia/players.asp`.

[7] "RealNetworks' RealPlayer," `http://www.real.com`.

[8] S. Wolf and M. H. Pinson, "Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system," in *Proceedings of SPIE International Symposium on Voice, Video, and Data Communications*, Boston, MA, September 1999.

[9] W. Ashmawi, R. Guerin, S. Wolf, and M. H. Pinson, "On the impact of policing and rate guarantees in Diff-Serv networks: A video streaming application perspective," in *Proceedings of SIGCOMM 2001*, San Diego, CA, August 2001.

[10] Y. Wang, M. Claypool, and Z. Zuo, "An empirical study of RealVideo performance across the Internet," in *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, San Francisco, CA, November 2001.

[11] D. Loguinov and H. Radha, "Measurement study of low-bitrate Internet video streaming," in *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, San Francisco, CA, November 2001.

[12] P. Calyam, M. Sridharan, W. Mandrawa, and P. Schopis, "Performance measurement and analysis of H.323 traffic," in *Proceedings of the 2004 Passive and Active Measurement Wrokshiop*, Antibes Juan-les-Pins, France, April 2004.

[13] J. Nichols, M. Claypool, R. Kinicki, and M. Li, "Measurement of the congestion responsiveness of windows streaming media," in *Proceedings of NOSSDAV*, Kinsdale, Ireland, June 2004.

[14] A. C. Dalal and E. Perry, "A new architecture for measuring and assessing streaming media quality," in *Proceedings of the Workshop on Passive and Active Measurements (PAM 2003)*, La Jolla, CA, April 2003.

[15] ——, "An architecture for client-side streaming media quality assessment," Hewlett-Packard Labs, Tech. Rep. HPL-2002-90, April 2002.

[16] "NIST Net network emulator," `http://www-x.antd.nist.gov/nistnet/`.