DELIVERY QUALITY SCORE MODEL FOR INTERNET VIDEO

Hojatollah Yeganeh[1], *Roman Kordasiewicz*[1], *Michael Gallant*[1], *Deepti Ghadiyaram*[2] *and Alan C. Bovik*[3]

[1]Avvasi, Waterloo, Canada
[2]Department of Computer Science, The University of Texas at Austin, USA
[3]Department of Electrical and Computer Engineering, The University of Texas at Austin, USA
Email: hyeganeh@ieee.org, rkordasiewicz@avvasi.com, mgallant@avvasi.com
deepti@cs.utexas.edu, bovik@ece.utexas.edu

ABSTRACT

The vast majority of today's internet video services are consumed over-the-top (OTT) via reliable streaming (HTTP via TCP), where the primary noticeable delivery-related impairments are startup delay and stalling. In this paper we introduce an objective model called the delivery quality score (DQS) model, to predict user's QoE in the presence of such impairments. We describe a large subjective study that we carried out to tune and validate this model. Our experiments demonstrate that the DQS model correlates highly with the subjective data and that it outperforms other emerging models.

Index Terms— Quality of experience, Delivery quality score, Mobile video

1. INTRODUCTION

Recent forecasts predict that by 2015 mobile video will represent 66% of global mobile data traffic [1] [2]. According to Google's official reports, more than one billion unique users visit YouTube each month, and over 6 billion hours of video are watched every month on YouTube [3]. Uninterrupted playback is a major component of the viewer experience, and playback interruptions are known to reduce viewing time or engagement [4]. Having a reliable QoE metric for internet video that reflects the impact of these delivery-related impairments is crucial.

Not surprisingly, Quality of Experience (QoE) metrics which focus on user perception of and satisfaction with service performance have received increased attention over the past several years [4–8]. Evaluating the quality of experience specifically for video services is usually done by conducting subjective studies wherein human subjects rate the perceived quality of video clips, with and without impairments [9, 10].

Subjective studies are the gold standard for assessing the quality of video services, representing the most accurate method for obtaining quality scores and ratings [11–13]. Results of a study can then be used to develop models for, or to evaluate the performance of, automated video quality measurement systems. Unfortunately, subjective testing is very expensive in terms of preparation, running time and human resources.

There have been numerous efforts and studies to develop objective quality measures for video, though most of these efforts are focused on more traditional sampling and coding impairments or on packet loss and associated spaitio-temporal artifcats due to unreliable streaming [14, 15]. The vast majority of today's internet video services, and certainly those which contribute to the forecast and marketing headlines above, are consumed over-the-top (OTT) via reliable streaming (HTTP via TCP), where the noticeable delivery-related impairments are limited to startup delay and stalling. Recently, objective methods that focus on these types of temporal artifacts have started to appear in the literature [5, 16, 17].

Avvasi has been working on these problems since 2008, and Avvasi's DQS model considers that a viewer's recent level of satisfaction or dissatisfaction plays an important role in their opinion about the overall QoE. This is supported empirically through multiple subjective studies focused on these artifacts [18]. This model is a continuous function that can provide a score at any point in time during a clip.

This paper provides an overview of Avvasi's DQS model including a description of performance on data collected from subjective studies carried out in collaboration with the Laboratory for Image and Video Engineering (LIVE) at the University of Texas at Austin. Experiments show that the model predicts QoE scores quite well and that it correlates highly with subjective data. Moreover, further analysis demonstrates that the model consistently outperforms another emerging QoE model.

2. DELIVERY QUALITY SCORE (DQS) MODEL

Initial buffering (startup-delay) and re-buffering (stalling) are the most important factors that contribute to reduced quality of experience (QoE) of OTT video services. The problem of estimating delivery QoE can therefore be reduced to predicting the perceptual impact of these delivery-related artifacts. Therefore, the goal is to define an objective model that takes initial buffering and re-buffering events into account, and produces scores which accurately predict a viewer's QoE at any point during the clip.

The model is constructed as a parameterized behavioral model. Behavioral models are often based on state models or state machines. This requires identifying important states and events which are likely to cause state transitions.

The Avvasi model (DQS) states have been determined from observations made from multiple subjective studies. This is further described in section 3.1. Analysis of subjective scores leads us to the following conclusions. First, on average, viewers do not react to initial buffering (or startup delay) with the same level of dissatisfaction as they do to interruptions once playback has started. Second, during playback, viewers do not react to a single interruption (first re-buffering) with the same level of dissatisfaction as they do to repeated interruptions (multiple re-buffering). Each of these three scenarios can be sub-divided into two sub-states, playback and no playback. A complete state diagram is shown in Figure 1 that also includes the initial and end states. Note that some level of startup

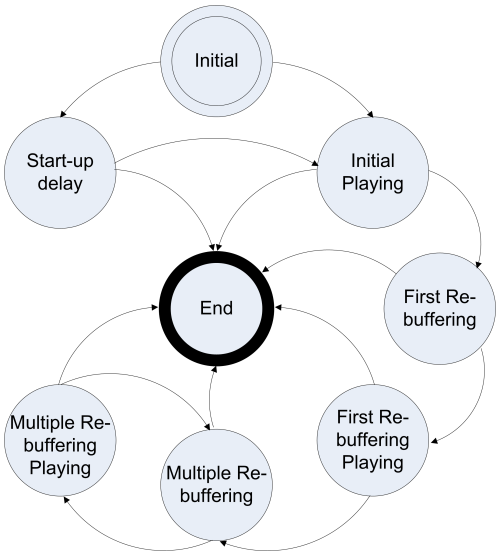

Fig. 1. DQS model state machine.

delay is tolerated in the initial state. Also, the "end" state represents not only the completion of a media session, it can be reached from all stall and playback states as a result of early termination. The DQS model is built on the premise that viewer satisfaction decreases during playback interruptions and increases during uninterrupted playback, as a function of time. The shape of these decreases and increases for a variety of events is captured by the following general parametric, function

$$f(t) = \begin{cases} 0 & 0 \le t < T_1 \\ \frac{a}{2}[1 + \cos(\frac{\pi(t-T_2)}{T_2-T_1}) & T_1 \le t \le T_2 \\ a + m(t - T_2) & t > T_2\,, \end{cases} \quad (1)$$

The above function is a combination of a raised cosine in the interval of T_1 to T_2, and a ramp with slope m after T_2. This function can model a wide variety of waveforms from a ramp to a logarithmic-like function. Note that other functions, additional intervals and additional states may be introduced to further improve model accuracy and/or to support additional delivery techniques (e.g. adaptive streaming) at the cost of added complexity.

2.1. Modeling Dissatisfaction During Interruptions

Using (1), we model viewer satisfaction decreasing over time as a result of interrupted playback, sometimes referred to as a frustration region. This can occur during startup delay, first re-buffering, and multiple re-buffering states. Let Q_{0-} be the DQS right before starting a re-buffering event. A decrease in human satisfaction may then be expressed as

$$\mathrm{DQS(t)} = \begin{cases} \mathrm{Q}_{0-} & 0^+ \le t < T_1 \\ (\mathrm{Q}_{0-}) - \frac{a}{2}[1 + \cos(\frac{\pi(t-T_2)}{T_2-T_1}) & T_1 \le t \le T_2 \\ (\mathrm{Q}_{0-}) - [a + m(t - T_2)] & t > T_2\,, \end{cases} \quad (2)$$

It can be seen from (2) that DQS(t) is monotonically decreasing and there is no lower bound defined for this function. However, the output of the model is meant to be in the range of 1 to 5, thus we let 1 be the lower bound of the model.

2.2. Modeling Satisfaction During Playback

The same reasoning can be applied to model viewer satisfaction that gradually increases over time as a result of uninterrupted playback, sometimes referred to as a recovery region. Using (1), we may express this as

$$\mathrm{DQS(t)} = \begin{cases} \mathrm{Q}_{0-} & 0^+ \le t < T_1 \\ (\mathrm{Q}_{0-}) + \frac{a}{2}[1 + \cos(\frac{\pi(t-T_2)}{T_2-T_1}) & T_1 \le t \le T_2 \\ (\mathrm{Q}_{0-}) + [a + m(t - T_2)] & t > T_2\,, \end{cases} \quad (3)$$

where Q_{0-} refers to the DQS value before resuming playback. Equation (3) is a monotonically increasing function without any upper bound, while the maximum MOS is typically 5, thus we let 5 be the upper bound of the model. Combining Equation (2) and Equation (3) for different states leads to a novel framework that generates a single score reflecting the delivery QoE for a media stream. The remaining task is to determine the function parameters for the different types of events and associated regions

As discussed previously, viewers do not react to startup delay with the same level of dissatisfaction as they do to re-buffering, nor do they react to the first re-buffering event with the same level of dissatisfaction as they do to multiple re-buffering events. This suggests using different parameters for two regions (frustration and recovery) of each of these events: initial buffering, single re-buffering and multiple re-buffering. Accordingly, six sets of four parameters (T_1, T_2, a, m) remain to be determined. Finding the parameters is essentially a regression problem. We used an exhaustive search to find the best parameters which predict subjective scores accurately. The details of tuning as well as the subjective study discussed in the next section.

3. TUNING THE DQS MODEL

3.1. Subjective study

In order to observe human responses to start-up delays and re-buffering events, and to tune and validate the DQS model against a large set of content and impairments, a subjective study was performed. This study was conducted in collaboration with the Laboratory for Image and Video Engineering (LIVE) at the University of Texas at Austin. One hundred and eighty videos were divided into two sets (set A and set B) where each set contains about 1.5 hour of footage where the durations of the shortest and the longest videos were 29 and 134 seconds, respectively. Fifty three subjects were then divided into two groups, and each group was required to rate only one set of videos. The subjects were asked to score the perceived quality from 1 to 5 for each clip, corresponding to the worst and the best quality, respectively. The subjective scores were recorded using the Absolute Category Rating (ACR) method [19, 20], as shown in Table 1.

Table 1. 5-level quality scale.

Score	Assessment
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 2. The mean and the standard deviation of subjective scores given to different stalling patterns with different stall duration

Stall duration in seconds	Start-up delay		First re-buffering		Two re-buffering		More re-buffering	
	μ	σ	μ	σ	μ	σ	μ	σ
5 - 10	4.57	0.03	3.94	0.1	–	–	–	–
10 - 20	4.52	0.22	3.89	0.33	3.54	0.11	3.4	0.28
20 - 30	–	–	3.83	0.37	3.44	0.25	3.2	0.41

It is known that the ACR method is susceptible to sequence effects, and thus the presentation sequence of the videos was changed randomly to mitigate that issue. Moreover, assessment results when using the ACR method are strongly affected by the scope of variation in the quality of test videos. Therefore, an ACR with hidden reference (ACR-HR) method was employed in our study. Typically, the assessment results obtained by the ACR-HR method are processed by calculating the difference in scores between the assessment video and the hidden reference video. The assessment results are thus recorded as difference mean opinion scores (DMOS):

$$\text{DMOS} = (\text{assessment video score}) + (5 - \text{reference video score}) \quad (4)$$

The perceptual quality of the reference video was judged by each subject to be either "5: Excellent" or "4: Good" by experts in the field of video quality. Finally, a statistical analysis as specified in the ITU-R BT.500 [19] was performed to remove outlier subjects from the set.

Table 2 provides the mean and standard deviation of the subjective scores for different stalling patterns and durations. This table shows a clear divide between two categories of impairment events: start-up delay events, and one or more re-buffering events. This phenomenon is also noted in other works [17] [5]. Moreover, given similar stall durations, there is a significant difference in scoring between sessions with one and two re-buffering events. However, the difference in the mean between sessions with two re-buffering events and sessions with more than two re-buffering events is much smaller. Specifically, the mean ± 1 standard deviation of the scores for sessions with more than two re-buffering events is almost within the mean ± 1 standard deviation of the scores for sessions with two re-buffering events. These observations support selecting start-up delay, first re-buffering, and multiple re-buffering as the important events for defining the model states.

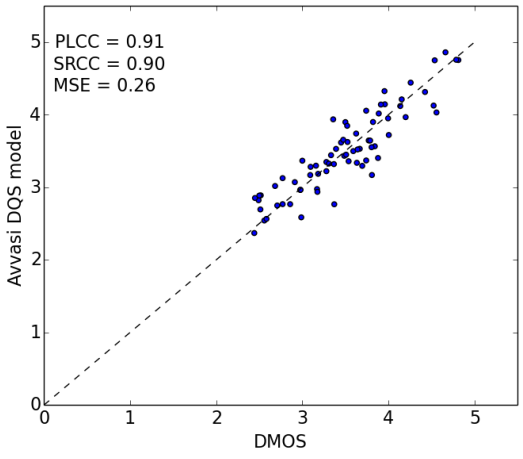

Fig. 2. Performance on the training set

3.2. Parameter tuning

The results on set A were regressed to tune the parameters of the DQS model. The best parameters were the ones which led to a minimum root mean squared error (RMSE) between subjective scores and the DQS model. Figure 2 illustrates the scatter plot of the DQS versus DMOS on the training set. It may be observed that exploiting the tuned parameters for set A resulted in a very high Pearson correlation (PLCC = 0.91) and very small deviation from DMOS (RMSE = 0.25). Note that all results exclude scores on the reference videos. Including the reference videos invariably increases correlation and decreases RMSE.

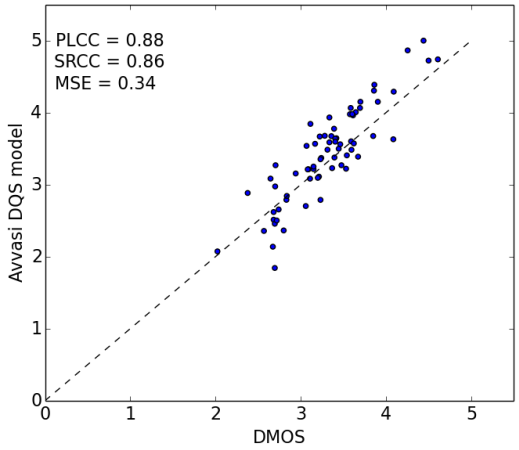

Fig. 3. Performance on the validation set

4. VALIDATION

Although the model predicts DMOS very accurately on set A, the important step remains to validate the model on set B which the DQS model was not exposed to. Adopting the tuned parameters and running the model on set B demonstrates that the DQS model correlates very well against the subjective data, with high Pearson correlation (PLCC = 0.88) and small deviation from DMOS (RMSE = 0.34). Figure 3 shows the scatter plot of the DQS model versus subjective scores on the validation set. Using the tuned parameters, we further analyzed the performance of the DQS model by inspecting different stalling patterns and by observing how well it behaves over time for these scenarios. Figure 4 shows several stalling patterns, using a red dashed line to represent playing or stalled. The blue solid line is the continuous DQS model output. The final DQS as well as DMOS scores are also included in the figures. Figure 4(a) illustrates the performance of the model for a media session with a 16 seconds start-up delay. Figure 4(b) illustrates the performance of the model for a media session with 2 seconds of start-up delay, 34

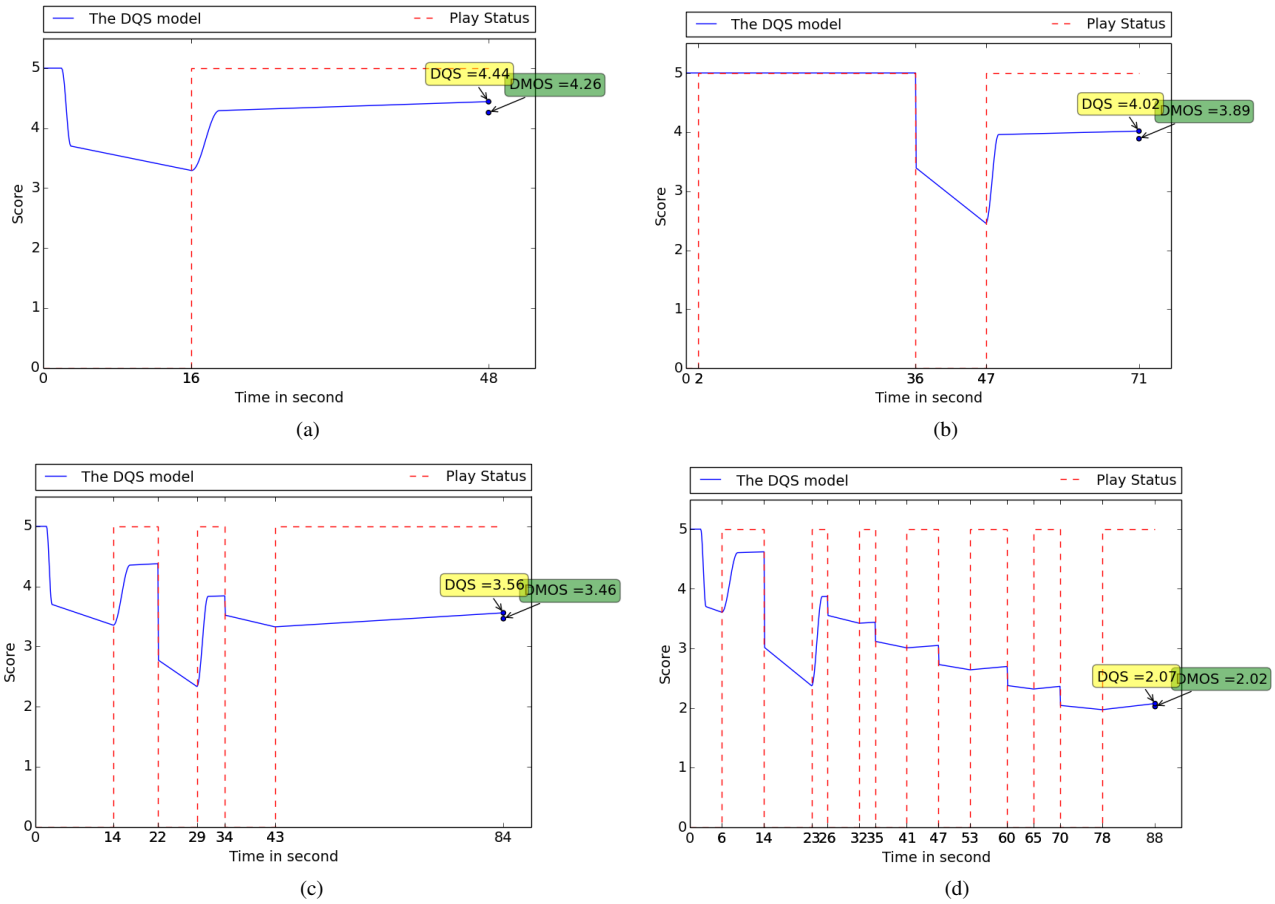

Fig. 4. Media sessions with different stalling patterns including (a) start-up delay (b) single re-buffering event (c) multiple re-buffering events(d) multiple re-buffering events.

seconds playback, 11 seconds of stall and 24 more seconds of playback. Figures 4(c) and 4(d) depict two media sessions with multiple stalling events. It can be seen that in all of these scenarios the DQS model accurately predicts subjective opinion. One recent objective model that has been proposed to predict the quality of experience was developed by the Telecommunications Research Center Vienna (FTW) [17]. The FTW model of [17] assumes that human perception is influenced by two major factors: the number and length of the stalls. It also assumes that the QoE can be expressed as an exponential function of these two major parameters. To further validate the DQS model, the FTW parameters were also tuned using set A in the same way as was done for the DQS model. FTW was then run on set B. Table 3 shows a comparison between the DQS and FTW models for both training set (set A) and validation set (set B). It can be observed that the FTW performs well in our experiments, but the DQS model outperforms the FTW model on both sets.

Table 3. Performance comparison on set A and set B excluding reference videos

Sets	Metrics	FTW model	DQS model
Set A	PLCC	0.85	0.91
	SRCC	0.83	0.90
	RMSE	0.42	0.26
Set B	PLCC	0.84	0.88
	SRCC	0.83	0.86
	RMSE	0.52	0.34

5. CONCLUSION

Delivery-related impairments play an important role in forming users' quality of experience (QoE) when viewing internet video services. Without having a reliable objective model to measure the delivery quality, content and service providers are unable to automatically predict users' quality of experience (QoE). We have made a substantive and demonstrably successful attempt to create an objective method that can accurately predict the QoE of internet video. Experiments using subjective data demonstrate that promising predictive power of the DQS model. Future work will include more careful parameter calibrations, and we plan to conduct more subjective studies on longer duration content.

6. REFERENCES

[1] Cisco Inc, "Visual networking index: Global mobile data traffic fore-cast update, 2012 to 2017," http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.htm, Feb. 2013.

[2] A. C. Bovik, "Automatic prediction of perceptual image and video quality," in *Proceedings of IEEE*, 2013, vol. 101, pp. 2008–2024.

[3] YouTube Statistics, ," http://www.youtube.com/yt/press/statistics.html.

[4] S. S. Krishnan and R. K. Sitaraman, "Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs," in *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC*, 2012, pp. 211–224.

[5] X. Tan, J. Gustafsson, and G. Heikkilä, "Perceived video streaming quality under initial buffering and rebuffering degradations," in *MESAQIN Conference (June 2006)*, 2006, vol. 90.

[6] V. Menkovski and A. Liotta, "Qoe for mobile streaming," *Mobile Multimedia–user and Technology Perspectives*, p. 31, 2011.

[7] R. K. P. Mok, E. W. W. Chan, and R. K. C. Chang, "Measuring the quality of experience of http video streaming," in *Proceedings of the 12th IFIP/IEEE International Symposium on Integrated Network Management, IM 2011*, 2011, pp. 485–492.

[8] W. Song, D. W. Tjondronegoro, and M. Docherty, "Understanding user experience of mobile video: framework, measurement, and optimization," *Mobile Multimedia: User and Technology Perspectives*, pp. 3–30, 2012.

[9] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens, "Quantifying subjective quality evaluations for mobile video watching in a semi-living lab context," *IEEE Transactions on Broadcasting*, vol. 58, no. 4, pp. 580–589, 2012.

[10] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens, "Quantifying the influence of rebuffering interruptions on the user's quality of experience during mobile video watching," *IEEE Transactions on Broadcasting*, vol. 59, no. 1, pp. 47–61, 2013.

[11] L. K. Choi A. L. Moorthy, A. C. Bovik, and G. De Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE Journal on Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, 2012.

[12] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[13] S. Jumisko-Pyykkö and J. Häkkinen, "Evaluation of subjective video quality of mobile devices," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 535–538.

[14] Z. Wang, H. R. Sheikh, and A. C. Bovik, "Objective video quality assessment," *The handbook of video databases: design and applications*, pp. 1041–1078, 2003.

[15] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121–132, 2004.

[16] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang, "Developing a predictive model of quality of experience for internet video," in *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*. ACM, pp. 339–350.

[17] T. Hofeld, R. Schatz, E. Biersack, and L. Plissonneau, "Internet video delivery in youtube: From traffic measurements to quality of experience," in *Data Traffic Monitoring and Analysis*, vol. 7754 of *Lecture Notes in Computer Science*, pp. 264–301. Springer Berlin Heidelberg, 2013.

[18] S. Ickin, L. Janowski, K. Wac, and M. Fiedler, "Studying the challenges in assessing the perceived quality of mobile-phone based video," in *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*. IEEE, 2012, pp. 164–169.

[19] ITU-T, "Methodology for the subjective assessment of the quality of television pictures," Recommendation BT.500, International Telecommunication Union, Geneva, January 2012.

[20] ITU-T, "Subjective video quality assessment methods for multimedia applications," Recommendation P.910, International Telecommunication Union, Geneva, September 1999.