A Study on Quality of Experience for Adaptive Streaming Service

Liu Yitong[1], Shen Yun[1], Mao Yinian[2], Liu Jing[1], Lin Qi[1], Yang Dacheng[1]
[1]Beijing University of Posts and Telecommunications
[2]Qualcomm Research San Diego
Liuyitong@bupt.edu.cn

***Abstract*—This paper proposes a Quality of Experience (QoE) test methodology for adaptive streaming service. We analyze the factors influencing the QoE of adaptive streaming, and evaluate the QoE of the end-to-end service. Our research gives the evidence that adaptive streaming improves end-users' subjective perception greatly compared with fixed-rate streaming in terms of QoE. Two groups of experiments are conducted, one assesses artificially spliced video samples in variable bitrates, and another evaluates Dynamic Adaptive Streaming over HTTP (DASH) [1] [2] video samples under real network traces. The former experiment focuses on the influence from the bitrate distribution without considering the fluency. And the latter pays more attention to the influence of the fluency on QoE performance. Through a large number of subjective assessments, the QoE results are obtained to illustrate various impact factors on adaptive streaming, such as fluency, bitrate distribution, startup bitrate level, bitrate switching frequency, etc. The results of this study can be used to facilitate the research on mathematical modeling of user subjective experience and the algorithm development for adaptive streaming.**

***Index Terms*—Adaptive streaming, QoE, DASH, Subjective video quality assessment**

I. Introduction

Recently, video streaming over smart phones, tablet PCs and laptops has become more and more popular, and more users are demanding high-quality video delivery even in challenging environments. The traditional way to deliver a fixed bitrate video stream is not suitable for diverse usage scenarios and continuously changing network conditions. New video delivery methods that can provide excellent Quality of Experience (QoE) for end users are much needed. As a result, adaptive streaming has recently been proposed as a form of Internet video delivery and is expected to be developed more broadly over the next few years [3]. According to a recent study by TDG, only 17 percent of the Internet video is supported by adaptive streaming technologies today, however, the adaptive streaming portion is expected to grow at an average rate of 77 percent a year towards supporting 51 percent of Internet video by 2015[4].

In the industry, proprietary adaptive streaming solutions are proposed, such as Smooth Streaming by Microsoft, HTTP Live Streaming by Apple, and Dynamic HTTP Streaming by Adobe. In addition, MPEG and 3GPP also developed an open global standard named Dynamic Adaptive Streaming over HTTP (DASH). Adaptive streaming aims to optimize and adapt the video configurations in order to provide the best quality video to the end users. Adaptive streaming service provides different variants of the same video content, each variant is further split in small segments and encoded at different bitrates. The streaming clients can dynamically choose different bitrate variant to fit the changing network conditions [5].

In traditional streaming service, the fixed bitrate directly determines the quality of the video picture. However, high fixed bitrate may give a worse experience on users' perception due to limited network bandwidth, which leads to frequent playback interruptions, i.e. low fluency, and impacts the video QoE. Adaptive streaming provides a better balance between fluent experience and the definition of video for a better QoE. However, a number of fundamental issues regarding an end-user's perception on adaptive streaming are not well understood. For example, what is the end-user's experience when the bit rate fluctuates but the video fluency improves? What factors influence the user experience the most, and how? Existing researches mainly focus on the understanding the QoE of the traditional fixed bitrate streaming. Developing evaluation methodologies and performance metrics to accurately assess end-user QoE for adaptive streaming is much needed. Such QoE evaluation model can also help optimize the bitrate adaption algorithm effectively.

In this paper, we proposed an adaptive streaming test methodology. Using this test framework we studied the QoE influence by different bitrate distributions. Then we evaluated artificially spliced variable bitrate video samples, as well as DASH video samples under real network traces. Through the analysis of 616 votes from 141 users, we extract helpful insights and arrive at conclusions.

The rest of this paper is organized as follows: in Section II, we present experiment design and methodology. Two groups of experiments, with different emphasis, are conducted to study the user perception in adaptive environment. The subjective test process is introduced as well. In the Section III, the subjective assessment results are presented and analyzed. Then adaptive streaming and fixed-rate streaming are compared with

978-1-4673-5753-1/13/$31.00 ©2013 IEEE

Group1 Aver=384kbps	Seg1	Seg2	Seg3	Seg4	Seg5	Seg6	Seg7	Seg8	Seg9	Seg10	Seg11	Seg12
Rising	R1	R1	R1	R1	R2	R2	R2	R2	R3	R3	R3	R3
Falling	R3	R3	R3	R3	R2	R2	R2	R2	R1	R1	R1	R1
Concave	R2	R2	R2	R2	R1	R1	R1	R1	R3	R3	R3	R3
Convex	R2	R2	R2	R2	R3	R3	R3	R3	R1	R1	R1	R1

Group A

Group2 Aver=768kbps	Seg1	Seg2	Seg3	Seg4	Seg5	Seg6	Seg7	Seg8	Seg9	Seg10	Seg11	Seg12
Rising	R3	R3	R3	R3	R4	R4	R4	R4	R5	R5	R5	R5
Falling	R5	R5	R5	R5	R4	R4	R4	R4	R3	R3	R3	R3
Concave	R4	R4	R4	R4	R3	R3	R3	R3	R5	R5	R5	R5
Convex	R4	R4	R4	R4	R5	R5	R5	R5	R3	R3	R3	R3

Group B

Group3 Aver=768kpbs	Seg1	Seg2	Seg3	Seg4	Seg5	Seg6	Seg7	Seg8	Seg9	Seg10	Seg11	Seg12
Bigteeth	R1	R5	R1	R6	R1	R5	R1	R6	R1	R5	R1	R6
Rising	R1	R3	R3	R3	R4	R4	R4	R4	R4	R5	R5	R6
Falling	R6	R5	R5	R4	R4	R4	R4	R4	R3	R3	R3	R1
Convex	R1	R3	R4	R4	R5	R6	R5	R4	R4	R4	R3	R3
Random	R1	R3	R4	R4	R5	R5	R4	R3	R6	R4	R3	R4

Group C

Group4 Aver=1024kbps	Seg1	Seg2	Seg3	Seg4	Seg5	Seg6	Seg7	Seg8	Seg9	Seg10	Seg11	Seg12
Bigteeth	R1	R6	R2	R7	R1	R7	R2	R6	R2	R6	R2	R6
Rising	R1	R3	R4	R4	R4	R5	R5	R5	R6	R6	R6	R6
Falling	R6	R6	R6	R6	R5	R5	R5	R4	R4	R4	R3	R1
Convex	R1	R4	R4	R4	R6	R6	R6	R6	R5	R5	R5	R3
Random	R1	R3	R4	R5	R6	R6	R6	R4	R5	R6	R5	R4

Group D

Fig. 1. The artificial samples to research QoE performance of the bitrate variation

each other. The characteristics of the variable bitrate distribution are also reported. In the Section IV, our discoveries are discussed and the conclusions are drawn.

II. QoE Evaluation of Adaptive Streaming

Two groups of experiments were designed and conducted to evaluate the performance of the adaptive streaming service's QoE. The first experiment put particular emphasis on time varying video quality using artificially spliced variable bitrate samples. It is designed to mimic the effect of the real-time switching among several different bitrates. The second experiment focuses on the influence of the improved fluency and lower rebuffering. Several DASH samples under different network traces and different average bandwidths are chosen to evaluate the QoE performance.

A. Testing for the Artificial Samples

To explore the influence from the bitrate switching and the distribution of bitrates, we manually created 25 variable-bitrate video samples of different bitrate distribution in lab environment. The counterparts with the same average bitrate are compared together.

The duration of each variable-bitrate video samples is 108s. Each sample has the same structure, with 12 segments spliced together artificially. Each segment lasts 9 seconds. These 25 samples are divided into 5 groups. The video samples in each group have the same average bitrate with different bitrate distributions. To explore different bitrate distribution patterns, for each segment, we chose the bitrate out of a set of bitrates 256kbps, 384kbps, 512kbps, 768kbps, 1024kbps, 1538kbps, and 2048kbps. These bitrates are denoted as R1 – R7 in Fig.1. In Fig.1, 4 groups of counterparts are shown. Only 3 different bitrates are involved in group A and B. The average bitrate for group A and B is 384kbps and 768kbps, respectively. There are 4 patterns designed in group A and B. The "Rising" indicates that the bitrate is climbing from a lower level to a higher level, while the "Falling" is on the opposite. The "Concave" and the "Convex" adjust the sequences of the bitrates to study the contribution to QoE by the initial and final impression.

More bitrates are involved in the group C and D, whose average bitrates increase to as high as 768kbps and 1024kbps respectively. Therefore, a new distribution pattern called "Bigteeth" is added to account for large variation in the streaming quality. A random distribution is added for comparison with the "Bigteeth" pattern. A separate baseline group, group E, which includes 7 different fixed-rate samples, is not shown in Fig.1.

B. Testing for Real Network Video Samples

29 recorded DASH samples, including 11 adaptive video samples and 18 single rate samples, are acquired in a test bed emulating the real network environment. Three network traces representing diverse video delivery environments are chosen. These network traces are obtained from real-world 3G wireless carrier networks via vehicle field test. Therefore there are frequent and large temporal fluctuations in terms of available bandwidth in these network traces. Every trace is scaled to three different average bandwidths, including 768kbps, 1024kbps, and 1538kbps. In addition, each trace incurs a constant 100 msec RTT delay.

The following characteristics are recorded and tracked for all of the DASH samples [6]:

--Bitrate distribution;

--Average bitrate;

--Number of bitrate changes;

--Startup bitrate.

Note that the startup bitrate of all DASH video samples is fixed at 256kbps.

Besides, for all the samples the following characteristics are observed as well:

--Stall time: total stall time, average stall duration, and stall distribution;

--Stagnant number.

We take three samples in Fig.2, Fig.3 and Fig.4 as examples to describe the fixed-rate and adaptive video samples. All of these samples are under the same network trace (Trace III) with the 1538kbps average network bandwidth. The purple line represents the actual network bandwidth, the orange line for the playback rate and the green line for the download rate.

In Fig.2, the playback rate is fixed on as high as 1024kbps. It can be found that there exists significant stagnation, even when the average bandwidth is higher than the playback rate. This is because the bandwidth is fluctuating around the average bandwidth severely. In Fig. 3, the video is fluently played when the rate is fixed on 512kbps. However, the download sometimes is stopped because the client buffer is easily full in the 512kbps case.

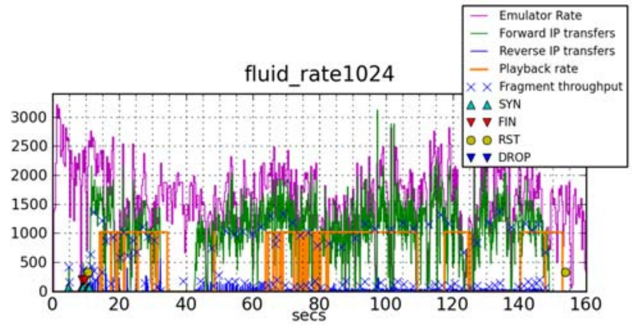

Fig.2. Fixed 1024kbps sample with 1538kbps bandwidth under trace III

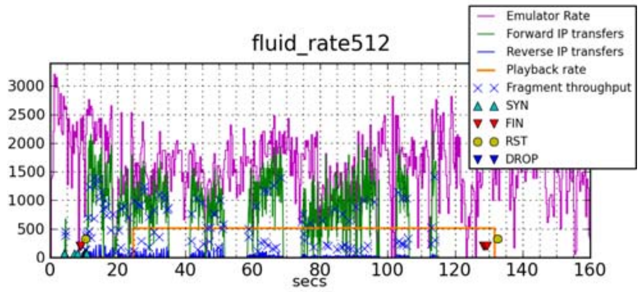

Fig. 3. Fixed 512kbps sample with 1538kbps bandwidth under trace III

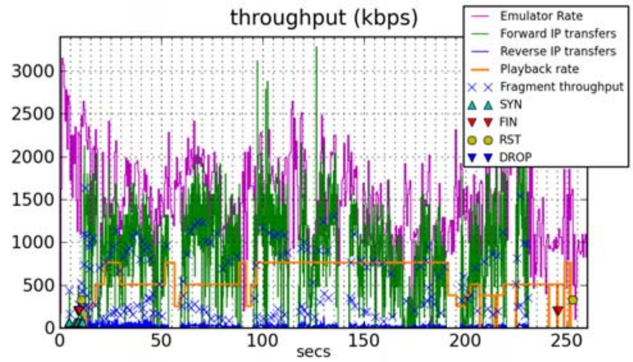

Fig. 4. DASH sample (a) (Avg.rate=613kbps) with 1538kbps bandwidth under network trace III

In Fig.4, an adaptive video with 613kbps average playout bitrate is shown to compare with the above two fixed-rate samples. In this sample the playout bitrate changes according to the network bandwidth variation. In Fig.5, the percentages of the participated bitrate are summarized. The total stall time of this sample is 0.63s, and the average stall duration is 0.05s. The stagnation between 0.2s and 0.5s happens once, and stagnations shorter than 0.1s are 10 times.

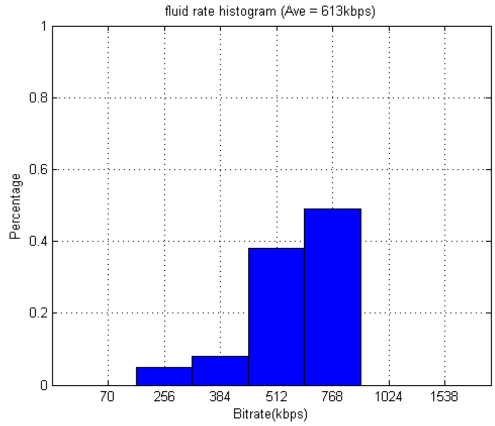

Fig. 5. The multiple bitrate percentage of the DASH sample (a)

C. Subjective Assessment Scheme

We recruited 141 viewers to assess the QoE of all of video samples above. Their average age is 27, among which the youngest is 18 and the oldest is 62. In addition, most of them were students and teachers.

The test environment was bright and non-interfering. The same playback equipment and the same resolution were required. According to characteristic of the adaptive streaming service, the formal test was performed by using no-reference (without any "original video") and single stimulus method [7][8][9].

Before the beginning of the subjective test, the rating options and experiment rules were explained and described to each volunteer in details. Then they were asked to accept a training using some examples.

TABLE I
TESTING INDEX

Vote standard	Quantity	Impairment
MOS	5(Excellent)	Imperceptible
	4(Good)	Perceptible but not annoying
	3(Fair)	Slightly annoying
	2(Poor)	Annoying
	1(Bad)	Very annoying
Definition	3	Clear
	2	Tolerable
	1	Very fuzzy
Fluency	3	Smooth
	2	Tolerable
	1	Very rough
Response speed	2	Tolerable
	1	Intolerable

Each volunteer rated 4-6 samples independently. Table I gives the test index particularly. The volunteers took part in the first Experiment should rate MOS [10] (the quality of the entire video) using 5-point scales, and the definition using 3-point scales. The volunteers took part in the second Experiment should rate MOS, definition, fluency, response speed.

In total, 295 votes were received for the artificial samples mentioned in the First Experiment, while 321 votes were obtained for the real network DASH samples and fixed-rate samples mentioned in the Second Experiment.

III. Result Analysis

Through collecting and analyzing 616 votes from 141 users, a number of indicators to QoE can be discovered.

A. Impact of Fluency to QoE

The fluency of video service influences QoE greater than the definition. DASH gives more fluent experience than fixed-rate streaming mode to users. For example, the fixed rate video shown in Fig.2 stalls for 55s in all. This 1024kbps sample is only scored 1.18. Meanwhile, under the same network Trace III the DASH sample (a) in the Fig.4 with just 613kbps average bitrate obtains MOS score as high as 3.18. This DASH sample stalls only 0.63s totally due to its adaptive mode. Its “Fluency” is scored 2.64. The fixed 1024kbps sample under the network trace I in the Fig.6 is graded up to 4.36, while its “Fluency” score is as high as 2.91, because the total stall time of this sample is 0.41s.The same conclusion can be observed in the Fig.7 under the different bandwidth.

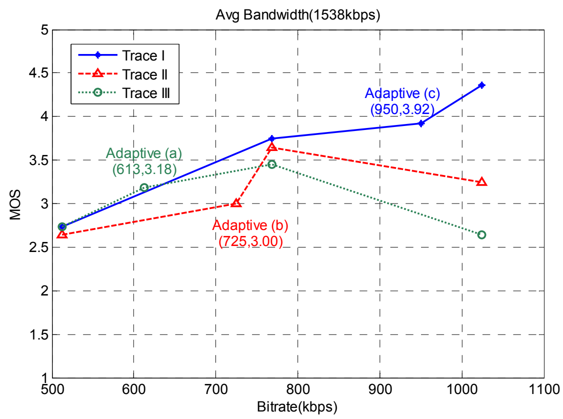

Fig. 6. The QoE performance with 1538kbps bandwidth

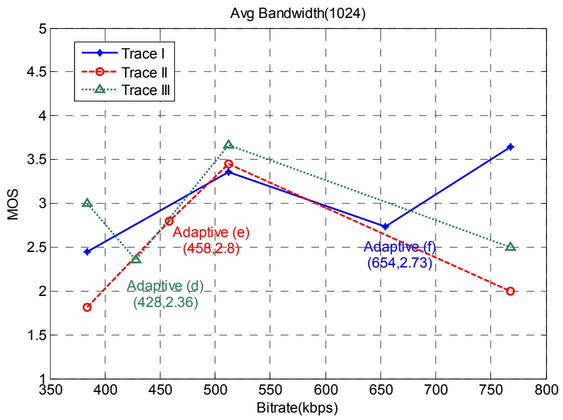

Fig. 7. The QoE performance with 1024kbps bandwidth

B. Impact of Startup Bitrate

The initial impression plays a crucial role in the experience of video viewers. It is observed that low startup bitrate and slow bitrate ramp-up clearly degrade QoE in the evaluation of DASH samples.

In Fig.6, a notable question is that although the DASH sample under Trace III has an average bitrate of 725kbps, higher than that of sample (a), it was given a lower MOS score. The detailed information of this sample (called DASH sample (b)) is revealed in the Fig.8. Its fluency score is 2.55, not much lower than the sample (a), because it takes 40s to increase from initial 256kbps to 384kbps at the beginning.

In our DASH samples, startup rate of adaptive video is limited at 256kbps. Most of time, resource assignment is conservative in the beginning of the service in order to limit the streaming bandwidth consumption and ensure the whole network performance. Then the bitrate climbs up gradually as the client adjusts to the network environment. However, the user perception may be influenced negatively during this phase. According to our test results, moderate startup bitrate and more aggressive ramp-up mode will improve the QoE drastically.

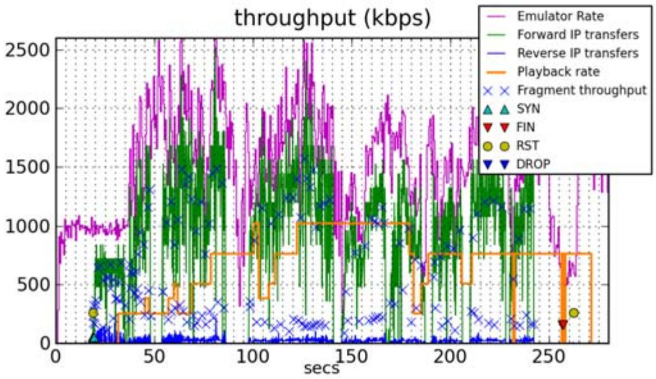

Fig. 8. DASH sample (b) (Avg.rate=725kbps, MOS=3, Fluency=2.55)

C. User Perception of Bitrate Switching

In both of the two experiments, it is found that users are sensitive to frequent bitrate switching. In the first experiment, among the artificial samples, the “Bigteeth” samples perform worst, even worse than the randomly variable bitrate samples, as shown in Fig.9. In addition, frequent switching with large rate differences deteriorates user perception. This can be seen in Fig.7 with the 1024kbps bandwidth. The DASH sample (f) with 654kbps average bitrate in Fig. 9 is scored 2.73. The DASH sample (e) with 458kbps average bitrate in Fig. 10 is scored 2.8. There are 1.2% of 1024kbps and 44% of 768kbps bitrates in DASH sample (f), while all of the involved bitrates of the DASH sample (e) are equal to or lower than 512kbps. However, DASH sample (f) has 28 rate switches. The MOS score of the sample (f) is lower than the sample (e), as well as lower than the fixed-rate 512kbps samples.

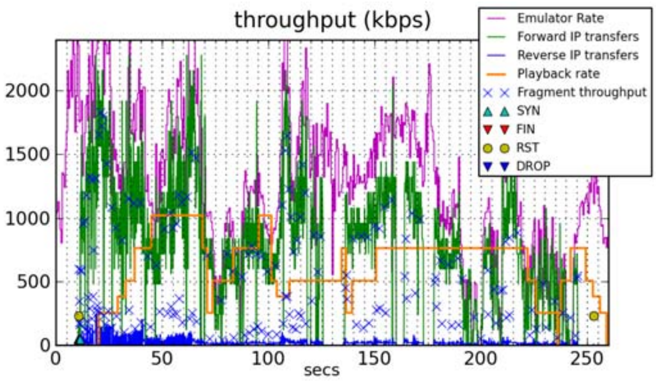

Fig. 9. DASH sample (f) (Avg.rate=654kbps, MOS=2.73, Fluency=2.64)

The subjective ratings indicate that short-term bitrate spikes (bitrate switching by a large delta), although consuming more network resources, degrades QoE. The QoE payoff for increasing the bitrate for a short time is negative. Meanwhile, short-term bitrate dip is harmful to the QoE, too, but cannot be avoided in most cases without a stall.

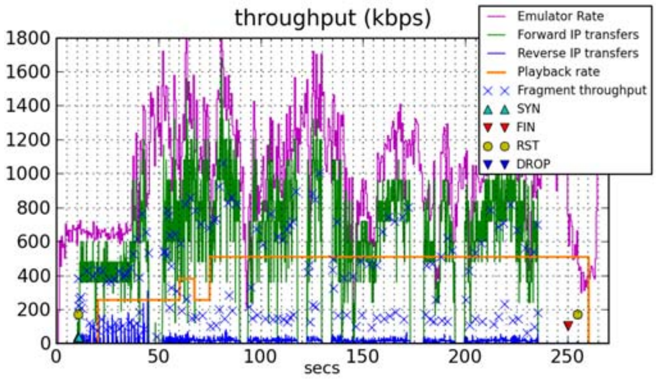

Fig. 10. DASH sample (e) (Avg.rate=458kbps, MOS=2.8, Fluency=2.6)

D. *Impact of Bitrate Distribution*

In the first experiment, artificially spliced multiple bitrate samples are assessed in order to study the influence from the bitrate distribution without considering the fluency. Fig. 11 shows that under the same average bitrate, "Falling" distribution obtains the highest MOS. It verifies that the startup bitrate is an important factor in another way.

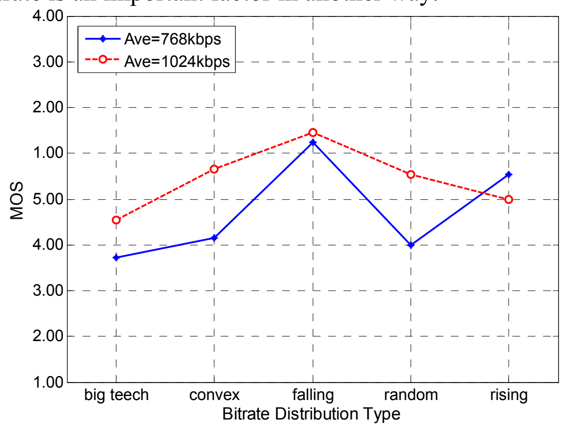

Fig. 11. QoE performance of multiple bitrate samples (involving 5 rates)

Fig.12 shows that under the same average bit rate, users perceive the fixed-rate samples to be better than most of the variable bitrate samples. Only the "Falling" distribution is rated on par with the fixed bitrate sample. This result indicates that under the same average bitrate and without considering the fluency factor, the variation of the video picture quality degrades the user experience. The similar indication that large-scale and frequent bitrate variations degrade QoE has also been demonstrated in the real network DASH tests.

IV. Conclusions

In this paper, we have studied the QoE of adaptive streaming in comparison with fixed-bitrate streaming. We have conducted two groups of experiments using manually spliced video samples as well as DASH samples under real network environments. Our experiments revealed a number of important influencing factors to the QoE of adaptive streaming, including the start-up bitrate, the distribution of bitrate, the frequency and scale of rate switching, and the fluency of presentation. Such studies are among the first to explore user perceptions of adaptively streamed video, and pave way to answer crucial questions regarding design criteria for adaptive streaming algorithms, and policies for network resource utilization.

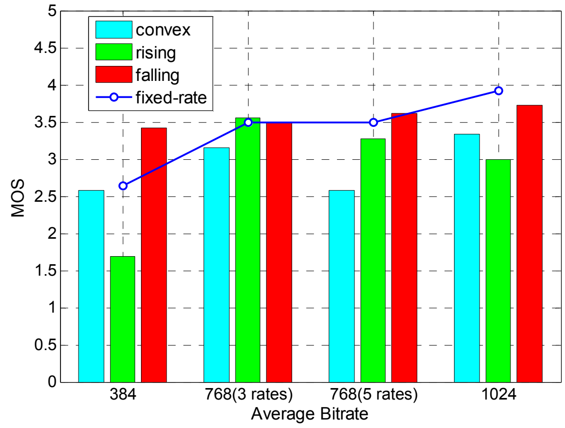

Fig. 12. QoE performance in several average bitrate

Acknowledgement

This work is a joint effort between BUPT WT&T Lab and Qualcomm Research's Adaptive Streaming Project [11]. The authors would like to thank Qualcomm Research for providing research grant, adaptive streaming expertise, and various test materials for user experiment.

References

[1] ISO/IEC 23009-1:2012 "Information technology — Dynamic adaptive streaming over HTTP (DASH) - Part 1: Media presentation description and segment formats".

[2] Thomas Stockhammer. "Dynamic Adaptive Streaming over HTTP –Standards and Design Principles," *MMSys'11*, pp.133-143, February 23 –25, 2011

[3] Ozgur Oyman and Sarabjot Singh, "Quality of experience for HTTP adaptive streaming services," *IEEE Communications Magazine*, pp.21-27, Apr. 2012.

[4] TDG Research, http://www.Tdgresearche.com

[5] Selim lckin, Katarzyna Wac, and Markus Fiedler, "Factors lnfluencing Quality of Experience of commonly used mobile applications," *IEEE Communications* Magazine, pp.48-56, Apr. 2012.

[6] Bouten Niels, Famaey Jeroen, Latre Steven, "QoE optimization through in-network quality adaptation for HTTP Adaptive Streaming," *Network and Service Management (CNSM), 2012 8th International Conference on* , pp.336-342, Oct. 2012

[7] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality", International Telecommunication Union, Geneva, 1996.

[8] ITU-R Recommendation BT.1788, "Methodology for the subjective assessment of video quality in multimedia applications," 2007.

[9] ITU-T Recommendation BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," Jan. 2012.

[10] ITU-T P.800.1, "Mean Opinion Score (MOS) terminology," July. 2006.

[11] Http://www.qualcomm.com/research/projects/dash