Distortions in QoE measurements of ubiquitous mobile video services caused by the preconceptions of test subjects

Péter András Kara, László Bokor, Sándor Imre
Mobile Communications and Quantum Technologies Laboratory (MCL) – Multimedia Networks and Services Laboratory
Department of Telecommunications (HT), Budapest University of Technology and Economics (BME)
Magyar Tudósok krt. 2, H-1117, Budapest, Hungary
E-mail: {kara, goodzi, imre}@mcl.hu

Abstract— **In telecommunication services, alongside QoS, QoE provision has become essential, thus performance and quality evaluation measurement results need to reflect reality as much as possible. Our goal is to enhance QoE evaluation schemes and enable improved QoE provision for video applications and services anytime and anywhere. In order to eliminate potential erroneous conclusions of QoE assessment techniques, our paper reveals a novel topic of distortions caused by preconceptions based on prior technical knowledge of QoE measurement test subjects. In our analysis we introduce the differences from genuine QoE measurement results in 3G ubiquitous mobile video service scenarios where test subjects were aware of the service parameters during measurements. We show how subjects' evaluations were affected and investigate the identified phenomenon in terms of Mean Opinion Score deviations and the overall QoE result distortion.**

Keywords: Quality of Experience, Quality of Service, Mean Opinion Score, performance evaluation, ubiquitous video servcies, 3G HSDPA, Internet-of-Services

I. Introduction

One of the most important pillars of modern society is the provision and consumption of services. The list of properties of a service provides comparable information to the consumer. Although this does seem to be the universal method of comparison between services of the same kind, it must not be ignored that it is not the equivalent of actual user experience. This means that no matter how high such properties score if the service does not satisfy the consumer. For instance, in case of a video chat which uses mobile Internet connection, it is totally irrelevant how staggering the bandwidth is when the two participants of the conversation have a hard time understanding each other. This leads to the conclusion that the true value of a service rather lies in user satisfactory (Quality of Experience – QoE) than pure numerical properties (Quality of Service – QoS). Service providers inevitably require user feedback to reach a cost effective level of QoE. Monitoring this phenomenon primarily benefits for service providers, but on the other hand, it improves reception for the subscriber.

Because of its importance, QoE monitoring is a well defined, standardized process. However, the results of such measurements are affected by the subjects prior knowledge on the present technology (Level of Comprehension – LoC), especially if one is aware of the QoS parameters. In some cases, the awareness of such parameters cannot be avoided; therefore the results are preordained to be shaped and distorted. Several examples can be mentioned from everyday life, where the preconceptions create distortions in user experience. The direction and power of these effects are quite far from triviality, yet it hasn't been circumspectly analyzed so far. Ubiquitous mobile video services demand accurate measurements and the avoidance of such distortions. This article deals with this untended topic: we study how the combination of aforementioned QoS parameters and different LoC levels distort the results of a QoE measurement. The QoE measurements were made on a real-life 3G HSDPA network. The objective of the test subjects was to grade the experienced quality of a video conference, while possessing the parameters of the mobile Internet connection.

The article begins with the introduction of present QoE measurement methods with some up-to-date examples in Section II, followed by the configuration and the results of our experiments in Section III and IV. The last section concludes the paper, containing the possible future directions of this topic.

II. Measurement methodology and related work

As mentioned in the introduction, QoE monitoring plays an essential role in designing, initializing and maintaining services. The standard techniques for such measurements are defined by the recommendation [1] of the International Telecommunication Union. It contains all the important parameters that can be involved in the configuration of a QoE measurement. Subjective determination of transmission quality can be achieved by four different clusters of methods. The most popular ones are considered to be the conversation-opinion tests, since they are designed to replicate actual service usage situations. Listening-opinion tests rather focus on ones perception, which makes them excellent to measure basic usability and acceptance. Interview and survey tests are efficient methods to extract information beyond a numerical judgment. A group labeled other tests is also defined. We decided to use conversation-opinion tests in our measurements, with minor additions from interview tests methodology [2].

An excellent example for conversation-opinion tests is the measurement [3] achieved by Yue Lu and his colleagues in the Netherlands. They focused their work on the astute

The current paper is also submitted to Symposium under submission ID 33.

choice of the video conference client. For listening-opinion tests, an absolutely noteworthy example is the measurement series [4] done by István Ketykó and his colleagues in 2010. Their topic was user experience and service quality acceptance in different environments. The field of subjective determination of transmission quality has well defined standards, intensely detailed recommendations, and countless of exceptional papers sharing the experiences of researches and measurements. The ITU-T P series [5] provide a wide range of recommendations relating to the topic. For those who would rather generally approach the subject matter, the publication [6] of Telenor is advised, or the book [7] of David Soldani, Man Li and Renaud Cuny, which also gives a view of great extent into world of UMTS networks. When mentioning measurement examples involving not only audio, but video quality evaluation, it is meritorious to separate those configured on a wired [8-11] and a wireless [12-15] network. It needs to be mentioned that while the preceding instances were based on human evaluation, there are automated evaluation methods [16] as well, excluding the human factor.

III. MEASUREMENT CONFIGURATION

The evaluation measurement was a video conference between the measurement guide and the test subject, such emulating a typical mobile video service in the world of ubiquitous computing. The tests were performed on the test network (see *Figure 1*) of the Mobile Innovation Centre [17]. Twenty subjects participated in the series of measurements with different levels of prior technical knowledge, ranging from simple inexperienced user to engineer with PhD degree. Just before the measurement itself, the Level of Comprehension of each subject was revealed by asking a set of questions related to the technical background of telecommunication activities. These conversations, each taking approximately thirty minutes, were recorded for further analysis to precisely determine the LoC of the subjects. Ten different levels were distinguished, which means two subjects represented a level. Level one represents the lowest, while level ten represents the highest level of technical comprehension. To preserve the purity of LoC determination, the subjects were given no information about the nature of the measurement before it had begun. The variety of technical competence was not the only aspect during the selection of the test subjects, but it was also necessary to only select people who have never seen each other before in order to prevent information leak between measurements. The subjects haven't even met each other during the series of measurements, because of the different dates of the measurements. If any subject had received even the slightest information about the measurement before its date, it could have and probably would have resulted in LoC overestimation.

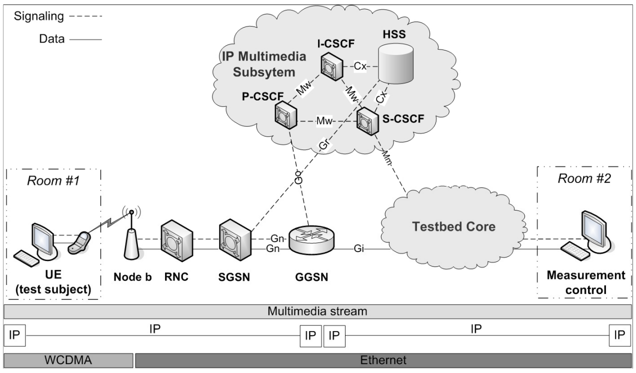

Figure 1. Testbed of the measurements.

The complete process of a measurement was divided into four sections, following each other without delay. The first part is considered to be the LoC level determination conversation, as mentioned before. This was followed by questions on general user behavior, involving the quality of previously experienced video conferences. After the basic instructions, began the third and most important part of the process, the mobile video conference and its evaluation. This was concluded by an oral evaluation of the experienced quality, which was also recorded like the first two conversations. The measurement guide was the same in each and every part of the process and for all subjects.

During the video conference, the guide used a terminal in the laboratory of the Mobile Innovation Centre (Room #2 in Figure 1), while the subject was isolated in the conference room of the laboratory. The audiovisual connection was established by *Linphone 3.2.1* client [18] on an *Ubuntu 10.04* operation system. Both end terminals shared the same hardware and software, including multimedia equipment such as web camera and headset. Connection to the test network, however, was different. While the terminal at the laboratory connected via Ethernet, the computer at the conference room (Room #1 in Figure 1) used a Huawei 3G HSDPA wireless modem. IP Multimedia Subsystem (IMS) [19] was in control of the mobile multimedia traffic over the UMTS network.

The video conversation took approximately one hour. Although it was divided into twenty subsections (test cases), the conversation itself was fluent and natural. Every subsection had a different artificial one-way QoS parameter load in terms of delay, jitter and packet loss, in addition to the real QoS values of the network. To achieve this, we used the command line based *netem* application [20] in order to change the output traffic of the laboratory terminal without the interruption or pause of the video conversation. The parameter values were given to the subject before commencing the conversation, in a form of a QoS parameter matrix (see *TABLE I*), together with the fix parameters of the measurement (see *TABLE II*), such as video resolution. The objective of the subject was to separately evaluate the audio and video quality of the twenty different test cases on a scale from one to ten.

TABLE I. QOS PARAMETER MATRIX VARIABLE VALUES

The current paper is also submitted to Symposium under submission ID 33.

Test case	**QoS parameters**		
	Delay	***Jitter***	***Packet loss***
1	0 ms	0 ms	0 %
2	50 ms	10 ms	0.5 %
3	200 ms	40 ms	2 %
4	800 ms	180 ms	8 %
5	0 ms	180 ms	8 %
6	0 ms	0 ms	8 %
7	0 ms	180 ms	0 %
8	800 ms	0 ms	0 %
9	800 ms	100 ms	1.2 %
10	400 ms	100 ms	1.2 %
11	200 ms	100 ms	1.2 %
12	100 ms	100 ms	1.2 %
13	100 ms	180 ms	0.5 %
14	100 ms	100 ms	0.5 %
15	100 ms	40 ms	0.5 %
16	100 ms	20 ms	0.5 %
17	200 ms	20 ms	0.5 %
18	200 ms	20 ms	2 %
19	200 ms	20 ms	4 %
20	200 ms	20 ms	8 %

TABLE II. QoS PARAMETER MATRIX FIX VALUES

Delay: 133 ms	Resolution: 640x480
Jitter: 30 ms	Video codec: MPEG4
Packet loss: 0%	Audio codec: speex

IV. MEASUREMENT RESULTS

The processing of the measurement results began after the last measurement had been finished. The first step was to match all the subjects to their most relevant LoC levels. This was achieved by analyzing the conversations recorded during the first section of the measurement process. The completion of the LoC based subject list generated the preferred results from the raw data set (see *Figure 2*) of the measurements.

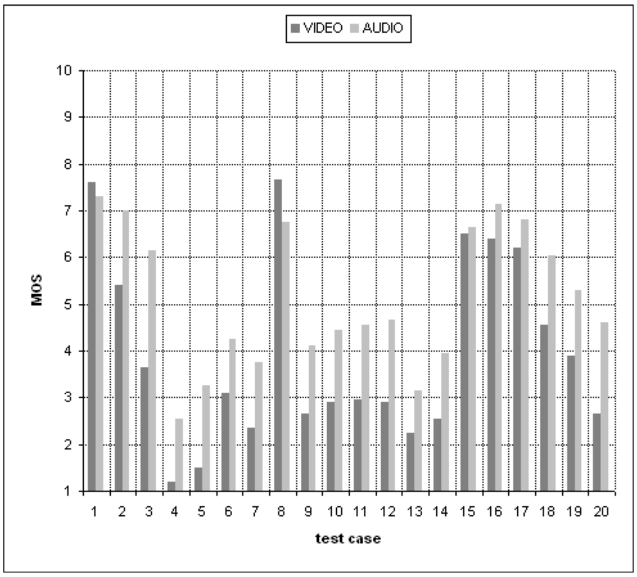

Figure 2. Measurement results.

First we took a closer look at the MOS, independently from LoC levels. As can be seen on the parameter matrix, while the first case was free of any additional load, case number 8 suffered eight hundred milliseconds of further delay. It would be expected for number 1 to have a better MOS score, however, the outcome showed the opposite. Even though the difference is rather minor, it cannot be denied that test case number 8 achieved a higher score in terms of video quality. By relying only on the MOS results, it would be quite exigent to give an accurate explanation to this phenomenon. The source of these MOS values can be found by dividing the results into two sets based on LoC levels: the first eight and the last two levels (see *Figure 3*). This subtracts those from the results who possess the lowest prior technical knowledge and provides an unequivocal explanation. While the subjects of the first set were commonly controlled by the fact that delay is noxious to experienced quality and thus such measurement case cannot achieve a better score, the rest was not aware of this. In fact, as heard on the recorded summary conversations, some people even think delay is beneficial and produces a higher level of quality. The subjects of the first set were not affected by such misbelieves so not even a single subject gave number 8 a better score, even though there is no major difference between the video quality of the two cases. The opinions of the people in the bottom two levels on the quality of these two cases were enough to create a distortion large enough to significantly affect the overall MOS results. It needs to be noted that the evaluation set of the top eight levels is also distorted, since there were people who barely distinguished the quality of the two cases but made a difference in evaluation because of their preconceptions.

The current paper is also submitted to Symposium under submission ID 33.

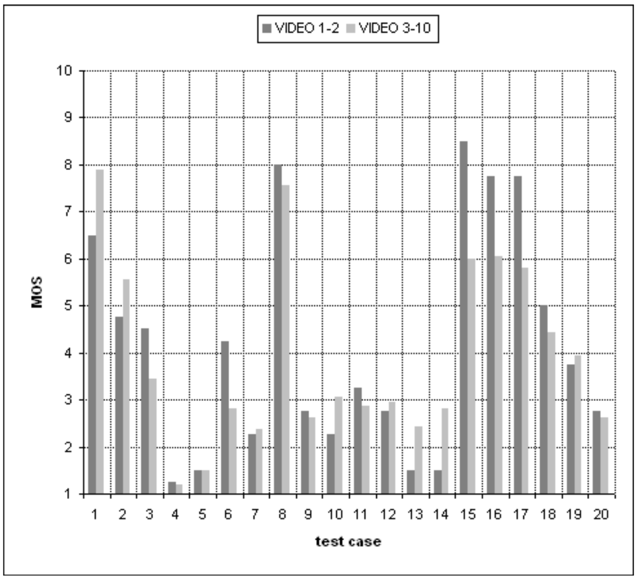

Figure 3. Video measurement results of LoC level 1-2 and 3-10.

Another noteworthy phenomenon in overall video MOS is the progress from case thirteen to sixteen. Jitter reduction describes these four cases as seen in the QoS parameters. Here we would expect an obvious rise in MOS, but the last case decreases instead. The previous explanation applies to this problem as well. If we take a look at the separated MOS diagram (see *Figure 3*), it can be seen that the subjects in the lowest two LoC levels are responsible again. In addition, the score of case 1 should be compared to case 15 on these lowest levels. There is indeed an immense difference between the quality case 14 and 15. While on 14 the video images are barely recognizable, 15 presents a quite enjoyable view. Because of the lack of evaluation control provided by prior knowledge and preconceptions, case number 15 received a stupendous score compared to number 1, the case where no additional load was applied.

The first four measurement cases represent a general decrement in QoS values, both delay, jitter and packet loss increases. In this progress, the overall MOS results provide nothing unusual, experienced quality decreases as it should. However, after we removed the lowest two LoC levels, we obtained something that was beyond our expectations. The average evaluation of the remaining subjects is mathematically equable, the scores show perfect uniformity, the line connecting these cases is straight. In some ways, this can be considered as one of the greatest distortions we have experienced so far. Before even encountering these cases, most subjects had a prior idea of how they should evaluate the quality, because they were already aware of the upcoming changes. This does not mean that the lowest two levels represent actual experienced quality values, since they are less likely to be controlled by preconceptions. As mentioned before, they are just as controlled. It is just enough to compare the audio scores of the first two cases on these levels. Number 2 suffers loads in all three QoS parameters, but still received a higher score because of technical misbelieves.

Audio evaluation holds a few interesting phenomena as well (see *Figure 4*). The focus is now on the progress from case number 9 to 12. These four cases endure delay reduction while preserving a notable constant jitter. Presuming the experienced quality tendencies in the progress is not a trivial task. It is beneficial to have a smaller delay, however, the ratio of jitter and delay increases. The audio MOS shows a definite raise in these four cases, even though none of the subjects thought it that way. After analyzing the LoC levels from one to seven, we could not find any obvious behavior pattern. In fact, there weren't even two levels showing the same relations between the adjacent cases, since there was no major overall difference in experienced audio quality. On one hand, mutual speech interruptions were fewer, but on the other, audio quality was less enjoyable to some extent. The scores given by the subjects were based on the personal decision whether the first or the second effect was more dominant. However, what we've discussed so far deals only with the lowest seven levels of LoC. The top three levels produced a marvelous result. The scores on these four cases were constant. It means that preconceptions had such a high level impact on evaluation that these subjects ignored any lesser differences that they experienced between cases. They considered the opposing effects nearly equal, which supposes an unvarying overall experience.

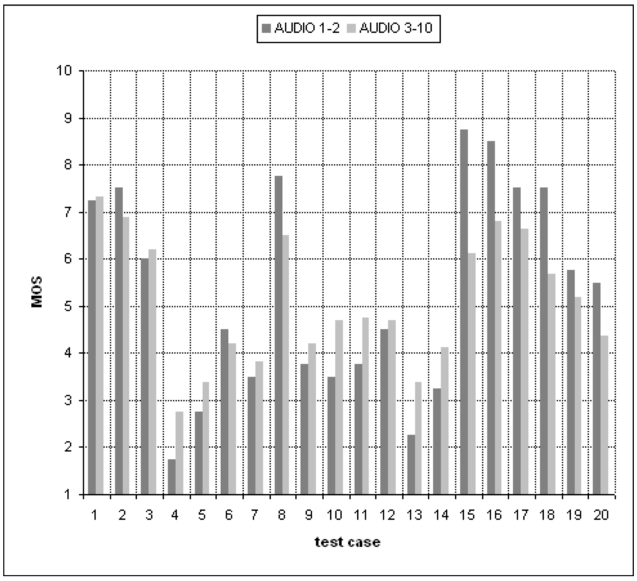

Figure 4. Audio measurement results of LoC level 1-2 and 3-10.

V. CONCLUSIONS

Audio and video quality evaluation in our measurements achieved more detailed and accurate information considering the distortions of MOS due to our novel approach. We managed to explain phenomena with the help of LoC that would have been unaccountable and ignored otherwise. The paper highlights the hazard that lies within preconception based distortions, the level of impact it can attain on the overall evaluation values. The results showed us that similar

The current paper is also submitted to Symposium under submission ID 33.

further measurements could improve QoE monitoring of ubiquitous video services, especially in this rapidly developing, progressing world of mobile Internet.

A possible continuation of this topic would be to separate the LoC from the initial configuration. In this case the same subjects would first evaluate the experienced audio and video quality without access to the QoS parameters, and then repeat the measurements with parameter awareness. This could result in a greater insight to distortions and would make it possible to define more detailed behavior models. It would be also interesting to use this approach for modem or end terminal equipment comparison, to involve a full duplex delay load in the measurements, or to analyze different solutions of handover. The number of varying parameters, test cases, or test subjects could be increased, but these are bounded by the cost-effectiveness of the measurements.

ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Union's Seventh Framework Programme ([FP7/2007-2013]) under grant agreement n° 288502. This work was also supported by the Mobile Innovation Centre Hungary (MIK). The authors are grateful to the many individuals whose work made this research possible.

REFERENCES

[1] International Telecommunication Union: „Methods for subjective determination of transmission quality", Aug 1, 1996.

[2] International Telecommunication Union: „Methods for evaluation of service from the standpoint of speech transmission quality", Apr 5, 1989. (Was deleted on 2009-06-30 since the four-point MOS scale that it describes is obsolete and replaced by the five-point scales defined in ITU-T P.800, P.800.1 and P.805)

[3] Yue Lu, Yong Zhao, Fernando Kuipers, Piet Van Mieghem: „Measurement Study of Multi-party Video Conferencing", ISBN: 978-3-642-12963-6_8, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands, 96-108, 2010.

[4] István Ketykó, Katrien De Moor, Wout Joseph, Luc Martens, Lieven De Marez: „Performing QoE-measurements in an actual 3G network", ISBN: 978-1-4244-4461-8, Shanghai, 24-26 March 2010.

[5] http://www.itu.int/itu-t/recommendations/index.aspx?ser=P
Last visited: 2012-02-13

[6] Bjørn Hestnes, Peter Brooks, Svein Heiestad: „QoE (Quality of Experience) – measuring QoE for improving the usage of telecommunication services", ISBN: 978-82-423-0620-3 / 1500-2616, R21/2009, Sept. 2008.

[7] David Soldani, Man Li, Renaud Cuny: „QoS and QoE Management in UMTS Cellular Systems", John Wiley & Sons, Ltd. ISBN: 0-470-01639-6, 2006.

[8] Ricky K. P. Mok, Edmond W. W. Chan, and Rocky K. C. Chang, „Measuring the Quality of Experience of HTTP Video Streaming", Proc. IEEE/IFIP IM (Pre-conf Session), May 2011.

[9] Arum Kwon, Joon-Myung Kang, Sin-seok Seo, Sung-Su Kim, Jae Yoon Chung, John Strassner, and James Won-Ki Hong: „The Design of a Quality of Experience Model for Providing High Quality Multimedia Services", Modelling Autonomic Communication Environments - 5th IEEE International Workshop, MACE 2010, Niagara Falls, Canada, October 28, 2010.

[10] David Rodrigues, Eduardo Cerqueira, Edmundo Monteiro: „Quality of Service and Quality of Experience in Video Streaming", in Proc. of the International Workshop on Traffic Management and Traffic Engineering for the Future Internet (FITraMEn2008), EuroNF NoE, Porto, Portugal, 11-12 December, 2008.

[11] R. Serral-Gracia, E. Cerqueira, M. Curado, M. Yannuzzi, E. Monteiro, and X. Masip-Bruin: „An Overview of Quality of Experience Measurement Challenges for Video Applications in IP Networks", Wired/Wireless Internet Communications, 8th International Conference, WWIC, Luleå, Sweden, June 1-3 2010.

[12] Dialogic: „Quality of Experience for Mobile Video Users", White Paper, December 2009.

[13] Amir Mehmood, Sachin Agarwal, Cigdem Sengul, Anja Feldmann: „Mobile Video QoE in Future Mobile Communications", TU Berlin / Deutsche Telekom Laboratories, Germany, 2010.

[14] Vlado Menkovski, Georgios Exarchakos, Antonio Liotta, Antonio Cuadra Sánchez: „Measuring Quality of Experience on a commercial mobile TV platform", Second International Conferences on Advances in Multimedia (MMEDIA'10), ISBN: 978-1-4244-7277-2, 13-19 June 2010.

[15] István Ketykó, Katrien De Moor, Toon De Pessemier, Adrián Juan Verdejo, Kris Vanhecke, Wout Joseph, Luc Martens, Lieven De Marez: „QoE Measurement of Mobile YouTube Video Streaming", ISBN: 978-1-4503-0165-7, New York, 2010.

[16] Daniel De Vera, Pablo Rodrıguez-Bocca, Gerardo Rubino: „Automatic Quality of Experience Measuring on Video Delivering Networks", ACM SIGMETRICS Performance Evaluation Review Volume 36 Issue 2, September 2008.

[17] https://www.mik.bme.hu/home/aboutus/
Last visited: 2012-02-13

[18] http://www.linphone.org/
Last visited: 2012-02-13

[19] http://www.3gpp.org/ftp/Specs/html-info/23228.htm
Last visited: 2012-02-13

[20] http://www.linuxfoundation.org/collaborate/workgroups/networking/netem
Last visited: 2012-02-13

The current paper is also submitted to Symposium under submission ID 33.