An interdisciplinary analysis of distortions in QoE measurement results caused by preconceptions and cognitive dissonance

Péter András Kara

YBVGB8

kara@mcl.hu

Supervisors: László Bokor and Sándor Imre

Department of Telecommunications

{bokorl, imre}@hit.bme.hu

2012

Table of contents:

Abstract – Quality of Experience (QoE) is an essential metric of the value of telecommunication services. Provisioning QoE is required to maintain and enhance the satisfactory of the customer, thus the result of such process needs to reflect actual user experience. Unlike in a wired scenario, the Quality of Service (QoS) values during wireless communication can momentarily change independently from service provider efforts, affecting QoE. Though it may sound serious, such effects should not be excluded in order to obtain proper, lifelike measurement results. An improper set of results can be misleading and easily disable further service improvement. Although evaluation techniques are indeed well standardized [1] [2], several distortions may still occur in the Mean Opinion Score (MOS). Such erroneous data can be caused by the preconceptions of test subjects. Any information on the QoS or basically anything regarding the given service can generate expected levels of quality before commencing the monitoring measurement. Regardless whether these preconceptions are close to reality or major misbelieves, they manipulate scoring of the test subjects. First of all, behavior models can be determined by Level of Comprehension (LoC) [3], the prior knowledge of the test subject on the involved technology. While people with less technical competence are more likely to be vulnerable to misbelieves and misunderstandings, those who belong to higher levels of LoC can easily get attached to their previous experiences. In both cases, cognitive dissonance [4] [5] is responsible. It means once an image of an expected quality is created, it is hard to let go, especially for those with the lowest or the highest LoC. The purpose of this study is to give a detailed insight into the world of such QoE distortions, to analyze the impact on the MOS and its seriousness, and to provide recommendations on handling the identified distortion phenomenon. The basis of the analysis is a series of real-time video conference QoE measurements performed on a real-life HSPA wireless testbed.

Section I

Introduction

Quality of Service was introduced by the International Telecommunication Union in 1994 [6]. It defines the term quality by a set of attributes that apply to the given service. Numerical descriptors like packet loss, jitter or delay – also known as latency – are fine examples of how the quality of a service can be described; the higher these values are, the lower the performance, the quality should be.

The word *should* is actually the keyword here. This sole word is the essence of the study that shall follow.

As it has been said, performance or quality was referred with a *should*. It is because of the single fact that Quality of Service cannot solely describe the actual human perception of quality; after all, no external human resources are required to measure QoS values. Since the consumers of the service are not involved in the measurement progress, it cannot reflect their reception of the service, although it is related to it.

Quality in the IT world can also be approached from the angle of the standards of the International Organization for Standardization. For instance, ISO 9126 [7] is the standard for the evaluation of software quality. Usability – one of its characteristic – includes understandability, learnability, operability, attractiveness and usability compliance as sub-characteristics. Although the standard is quite precise, the measurement of these aspects can be done in almost any desired way. Understandability can be measured by any of the following practices in case of a software product with several functions to discover:

- How long time does one need to discover a given percentage of the software?
- How many functions can one discover in a given time?
- Can one discover a specific function in a given time?
- Can one determine whether the software is capable of creating a specific output?

It is not difficult to imagine a vast variety for the aspect of attractiveness. Measuring the usability of services is a little bit different.

The measurement of QoE defined by ITU [1] rather focuses on general satisfactory. Feedback results of QoE monitoring are compressed into the MOS, which provides the average acceptance of the service. The typical scale for such measurements ranges from 1 to 5, where 1 represents poor quality and 5 is excellence. Other scales are usable as well, for instance a scale from 1 to 10, which gives more space for the evaluation of quality. Numbers assigned to intervals are usually positive integers; however, negative numbers also appear on evaluation scales. During comparative measurement, a five-interval scale could range from −2 to +2. Regardless what scale is used, the MOS is a number or series of numbers, reflecting how satisfied the users are. The focus is commonly on general satisfactory, but any component of the given service can be a target of the measurement.

Related work

Due to the importance of QoE monitoring and provisioning, there are plenty of exceptional papers and books in this topic.

One of the first approaches of QoE was the analysis of human-computer interactions; to study the experiences of people with a given technology. The publication of L. Alben [8] is a great example of this concept. Today the primary purpose of the measurement of human-computer interactions is to provide the industry a feedback on service acceptance and user satisfactory.

The book [9] of D. Soldani et al. provides a very wide, extensive look into QoS provisioning and QoE monitoring. It allows the reader to acquire nearly any information regarding QoS/QoE in UMTS networks. From streaming and gaming performance to different monitoring tools, the book has it all.

About assessment techniques, although discrete scales are indeed the popular ones, continuous scales are utilized too. H. Ridder et al. presented the usage of a continuous scale for video quality evaluation in their article [10]. They used a graphical slider as assessment input. In this approach, the correct choice of sample rate is vital; a rate too low would result in the risk of missing crucial data, while a rate too high would require immense memory. We preferred the usage of a discrete scale in our measurement, since our aim was average satisfactory instead of continuous feedback.

The work of P. Brooks et al. [11] focuses on subjective and objective measures of QoE, analyzes and compares the current approaches for measuring network service quality from user perspective, and it is very edifying for those who would like a taste of the world of QoE measurement methods. It is full of descriptive examples which assist understanding the vast variety of what this field of science has to offer. The paper also underlines the significance of quantitative data over qualitative data in subjective measurements. They even made a demonstrative, comparative figure to illustrate the differences between the well-known five-interval ITU quality scale equipped with its traditional labels on each interval, and a ten-interval scale with only labels at the ends; the first one is a qualitative scale while the other is a quantitative scale. In our measurement, we used a scale identical with the ten-interval quantitative scale presented in this paper, as it shall be seen later in our study. Our test subjects only knew that 1 was the lowest and 10 was the highest score that could be given, but score numbers between them were not marked with labels like “poor” or “fair”. Their example of alternative qualitative scales is also appealing, but we decided to stick to global QoE ratings.

The paper of A. Watson et al. [12] argues that ITU recommendations are no longer suitable for the QoE evaluation of many new services. Similarly to the previous paper, the authors criticized the qualitative scales of ITU and found them insufficient for their purposes, especially their inappropriate labels. Due to these labels, evaluations are likely to be concentrated and compressed into the lower half of the scale. They also emphasized the problem with the interpretation of these labels; although as in any linear scale, distances between adjacent intervals are meant to equal. To validate label vocabulary, there were investigations [13] [14] where test subjects had to place the given qualitative terms onto the most fitting interval. Not only did they find out that the scale is not uniform, but they also noted that these results vary a lot in different languages. The binary conversation difficulty scale is also defined by ITU for assessment purposes. The recommended question is the following: “*Did you or your partner have any difficulty in talking or hearing over the connection?*” This question awaits 1 as answer in case of YES, and 0 in case of NO. Of course it is usable to determine whether the given audio service is flawless or not in terms of usability. We decided not to use it in our measurement. Instead, we dealt with the matter in an oral discussion after the measurement. The paper also criticizes that the length of recommended test material is too short. Indeed, ten seconds at most is not much. We chose this value to be eighteen times more, which appeared to be sufficient.

M. Fiedler et al. published a paper [15] dealing with the relationship of QoE and QoS, and the exponential interdependency between them, called the IQX hypothesis. They also analyzed the qualitative relationship between QoE and QoS, and illustrated it with a very motivating mapping curve. Three different areas were distinguished: constant optimal QoE, sinking QoE and unacceptable QoE. The name of the first area defines its properties well; constant, because QoE does not vary with the growth of QoS disturbance, and optimal, since "*the user considers the QoE equivalent to that of the reference*". In our measurement, we had a couple of different QoS disturbance setups, where although the difference in QoE was barely detectable, yet the majority felt an urge to make a difference in scores. We even experienced the opposite phenomenon, where a varying user experience received constant scores.

The tests presented in the study [16] of Y. Lu et al. investigate the choice of client during multi-party video conferencing. According to their work, the selected application has a great influence on the received quality of the service. Due to its importance, the authors provide a recommendation on designing video conferencing applications. To test the minimal bandwidth required to launch a video conference, a specific traffic control and monitoring tool was used to adjust the upload rate at each participant. We used a different application to create QoS disturbance in order to analyze the test subjects' reactions to QoE degradation. At the end of our study, we also make a proposal, in attempt to improve the refining of QoE measurement results.

An excellent example for QoE measurements was published [17] in 2010 by I. Ketykó et al. The authors highlighted how much the current environment of service consumption can affect the experienced quality. In this context, environment refers to the location and the number of surrounding people as well. The results indicate that both can have a significant impact on the QoE. For example, watching the given movie trailer at home achieved a higher sound MOS than watching it during travel and focusing on it was easier without anyone around.

Due to the novel nature of our study, our work is, to our knowledge, the first to analyze the effects of test subjects' preconceptions and cognitive dissonance on QoE measurement results.

Motivations

Let us imagine a setup very similar to the previously mentioned measurement. The target objective would be to measure the difference in QoE between watching a given trailer during travel and at home, using a smart mobile client in a real-life 3G network. There would be several test subjects with diverse attributes, such as age or field of work. One of these attributes would be prior technical knowledge and experience on the current telecommunication technology. Diversity in this case would mean that this prior knowledge – let us define it as Level of Comprehension (LoC) [3] – would range from inexperienced user to highly skilled IT engineer. Let us also assume that the test subjects would know nothing about the parameters of the service, only the single fact that one part of the measurement would take place at home, the other during travel. There would be no specific information to gear the subjects with preconceptions; ideas about the upcoming user experience.

Or would there be? First of all, let us try to view the situation from the IT engineer's point of view; the one with the highest LoC. He or she would already have plenty of experience with such devices, not to mention the familiarity with the technological background of 3G networks. Because of the immense information and involvement, one would easily be influenced by thoughts regarding the matter. That would mean that even before commence of the measurement, one would expect that the quality would be less enjoyable during travel due to numerous technical reasons. One of these would be handover; since the mobile client is actually mobile during travel, handovers would occur, which imply temporary QoS/QoE degradation.

When the moment of evaluation would come, these pressing thoughts would be present. Even if there was no or just barely notable difference in the experienced quality between the two use cases, the subject would feel the urge to make a difference in scores; the experienced quality at home would receive a higher score than during travel.

As mentioned earlier, the word *should* is the quintessence here. One with such high LoC could easily think that the experience during travel *should* be worse. More straightforwardly, it has to be worse, due to many reasons. No matter what the genuine perception of user experience would be, the subject would convince oneself that there actually is a difference, according to one's prior thoughts.

What about those who possess no such knowledge at all? Would they be free of this feeling? Since assessments are usually rushed, especially according to the ITU recommendation on listening-quality tests, one would need to make a decision on quality with haste. A hurried decision usually seeks something to rely on, in order to justify its rationality. This could be an incorrect idea about the technology. For instance, such idea could be that quality during travel *should* be better, since the velocity of the device enables the little ones and zeros to flow even faster in the air. Even though a test subject with such a preconception would perceive no noteworthy difference in experienced quality, still there would be a variance in scores in favor of the experience during travel.

Section III will highlight the origins of such evaluation distortions through our approach of LoC separation. But before we make our way to the analysis, we have to get familiar with the measurement which served as the basis of our investigation.

Section II

Aim of the measurement

The previously mentioned line of thought was presented [3] at SAINT2012 in Izmir, Turkey. The paper was based on a series of QoE measurements that were performed at the Mobile Innovation Centre [18]. The primary objective was to detect and analyze the effects of preconceptions in QoE monitoring procedures.

The leading question in this novel topic was the following: Can preconceptions create distortions large enough to have an observable impact on the MOS, or is it ignorable under any circumstances? To answer this, we designed a testbed on a real-life 3G network (see *Figure 1*). The investigated service was a video conference between two people.

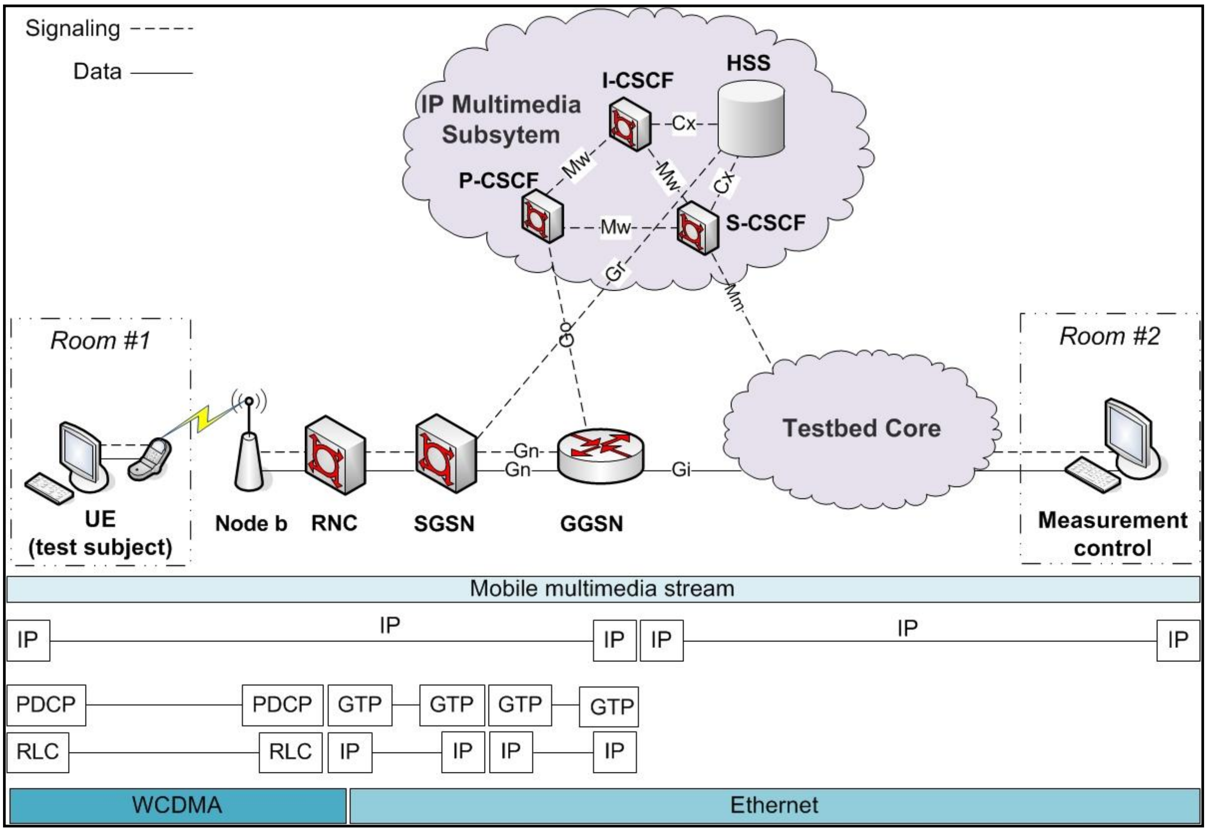

Figure 1 – Testbed at the Mobile Innovation Centre

The participants of the video conference were well separated; they were located in two different rooms of the Mobile Innovation Centre.

Measurement configuration

The location titled as Room #1 was the conference room, which we found ideal for such an evaluation measurement, due to its physical parameters; sufficient room volume, no echo phenomenon, no galling internal noises – for example lurid noise of active machinery – and the noises of the outer world were well blocked. The participant on this end of the communication was the test subject, so it was indeed an important matter to provide a suitable environment. According to the measurement results of the previously mentioned I. Ketykó et al. [17], not only the choice of environment is vital, but it is definitely easier to focus on the evaluation task when there are no other people around. The test subject was isolated; it was made sure that no one would disturb his/her part in the measurement process.

The terminal on the evaluator's side was a desktop computer equipped with a web camera and a head set. The one piece of hardware that needs to be specified is the Huawei 3G HSDPA wireless modem, which was used to connect to the laboratory network.

On the other side of the communication was the measurement guide. His place during the measurement was at Room #2, inside the laboratory. His task was not only to provide a partner in the audiovisual communication, but also to instruct the test subject and control the QoS parameters, which shall be detailed later on.

The terminal on the measurement guide's side had exactly the same hardware and software, including multimedia equipment. To only exclusion to this was its connection to network; while the terminal of Room #1 used a wireless modem to establish connection with the network of the laboratory, the one in Room #2 was connected via Ethernet.

A Linphone 3.2.1 client [19] was chosen for the video conference, running on an Ubuntu 10.04 operation system. As it can be observed on *Figure 1*, the testbed included an IP Multimedia Subsystem (IMS) [20], which was in control of the multimedia traffic over the 3G network.

The configuration of the testbed was untouched during the measurement process. Analyzing the evaluation of a single service would not have been very operative in terms of distortions, since distortions are most likely to be detected during comparisons. We decided to split up the measurement of the given service into 20 subsections. Although the configuration of the end-to-end communication was unmodified, netem [21] was used at measurement control to alter service QoS, thus affecting QoE. By determining 20 unique setups of artificial QoS loads, we virtually generated 20 different service levels of the mobile video conference application to be evaluated. One subsection – from now on referred to as test case – was defined by the amount of extra delay, jitter and packed loss (see *Table 1*), in addition to the genuine QoS parameters (see *Table 2*) of the real-life 3G network.

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 1 – Variable QoS parameters of the service

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

Table 2 – Genuine QoS parameters of the service

Test subjects and Level of Comprehension

From the people who volunteered to participate in the measurement process, twenty were selected, with the central aim of diversity in prior technical knowledge. The youngest participants were between eighteen and twenty, and the oldest were in their mid-thirties. There were completely no intentions to have equal numbers from the different genders, so in the end there were a bit more male test subjects than female ones.

Education, occupation and personal interests were the attributes that were the basis of test subject selection. Some people had minimal prior experience with telecommunication services and nearly zero familiarity with the subject, while some others had day-to-day contact with such services due to profession or daily routines. Of course there were also those who were in between these two radical ends of comprehension, since variety was essential.

Although these attributes were proven to be of assistance during test subject selection, but far from sufficient to determine their relations in terms of Level of Comprehension. We introduced this term to distinguish different levels of how test subjects refine the service related inputs; one with immense knowledge and experience considering the subject has a higher capability to extract information about the service. Though this term is already present in other contexts, it has a rather different meaning; referring to education and skills, for instance comprehension as in extracting information from articles and books, understanding their messages.

As it has been said, just some data on the test subjects could not distinguish different levels of LoC. It required its own procedure. This shall be detailed later on in the phases of the measurement.

Indeed diversity was the main aspect in selecting the test subjects. However, there was one criterion to take into consideration: we had to select subjects who make disjoint sets in terms of relationship. That means people could not know about each other, so affecting each other's evaluation was impossible.

Phases of the measurement

For each and every test subjects, the measurement was divided into four phases. The different phases followed each other in numerical order, without delay. The test subjects received the same conditions in all phases.

Phase 1 – LoC determination

The purpose of the initial phase was to gather the information required to determine the test subject's LoC level; this is the aforementioned procedure. It was a real-life conversation between the test subject and the measurement guide, performed in Room #1, recorded on camera for further use. Of course there were several other options as well how to determine one's LoC, but we chose the conversation-option for various reasons. It is a fact that it would have been much faster to complete and easier to process to use a multiple choice survey based on questions involving the current technology. The problem in this case would have been the danger of overestimation; people with lower prior knowledge could have obtained a higher score though guessing, thus one would have been allocated into an inappropriate, higher LoC level. Other written surveys or tests seemed ineffective because of their limitation in information amount. Making a recorded conversation was beneficial because of the opportunity of the test subject to answer a question as detailed as he/she preferred. We used a fix set of questions, all waiting answers about the size of at least three or four sentences.

About the test subjects, not only did they not know about each other, but they didn't have any detailed information on the measurement either. The only thing they knew that they were going to participate in some kind of a measurement. These two precautions were very important to preserve. If one would have known specific details about the measurement, or even about the nature of the measurement, one could have gathered some information on the subject, making serious overestimation in terms of LoC. Having two test subjects known each other would have jeopardized the purity of results even more. For example, if one would have shared his/her experiences about the measurement with another test subject who had not participated yet. The duration of this phase was around twenty minutes. The data gathered in this phase was processed after all twenty test subjects were ready with the measurement, since there was no rush to determine the LoC before commencing the video conference.

Phase 2 – Prior user experience discussion

Similarly to Phase 1, Phase 2 was a recorded real-life conversation between the test subject and the measurement guide, also performed in Room #1. However, the goal in this phase was to gather some information on prior experiences with services and devices; of course with a special focus on disappointments regarding services. The results of Phase 2 were not used in the determination of LoC levels; LoC itself was solely determined by Phase 1. Instead, it provided metadata which helped during the analysis. Duration was approximately ten minutes.

Before advancing to Phase 3, it needs to be mentioned that the first two phases are acceptable because of the number of test subjects. Scalability is in fact a serious issue of this approach; while twenty test subjects required massive yet supportable effort, two hundred test subjects could not have been manageable. Just for example, our approach of LoC determination would necessitate a much more cost-effective method in case of numerous test subjects. Even with this tolerable number, almost twenty hours of recorded conversation awaited profound analysis.

Phase 3 – Usage and evaluation of the service

After the recorded conversations were completed, began the usage and evaluation of the service. Before commencing the video conference, the test subject received basic instructions and was given the parameter matrix, including all the genuine and variable parameters of the service. In all four phases of the measurement, the same measurement guide was present, which was the most important at Phase 3, in order to provide the same conditions for all test subjects.

The usage and evaluation took nearly three minutes for each test case, so the entire phase was approximately one hour. Three minutes for such a procedure does not seem too long, but we needed to keep the complete length at bay, since perception varies over time due to the physical limitations of the human body and mind. Although the average length of the attention span in case of adults is not more than forty minutes [22], the task did not require exceptional focus – for example handling fragile objects – and there were no distractions during usage and evaluation.

The video conference was a fluent and natural audiovisual conversation. The test subject was provided a three minute sample of the service, and made a score when the incoming sample was sufficient, typically around the middle of the third minute. The only unnatural elements of the conversation were the sentences announced by the measurement guide to differentiate test cases; for instance "*This concludes test case one, commence of the case two.*"

Evaluation was done on a sheet of paper; the test subject had to mark the cells of the corresponding scores. The idea of handwritten scores was of course out of the question, since scores had to be unambiguous. A quantitative, label-free, ten-interval scale was used for the assessment of audio and video quality. At the end of Phase 3, when the conversation was already over and the measurement guide returned from Room #2 to Room #1, the validity of the results was checked. The sheet needed to contain one and only one score for each and every test case in terms of video and audio quality; a missing or multiple score would have resulted in an instant abortion of the measurement, without the commence of Phase 4.

Phase 4 – User experience discussion

Phase 4 was done in a similar manner to Phase 1 and 2; it was a recorded conversation in Room #1 between the test subject and the measurement guide. The topic was the experienced quality of Phase 3, especially negative experiences. The goal of this phase was to understand the motivations behind the given scores.

As it shall be seen later, the recorded conversations of Phase 4 were indeed a lot of assistance during the analysis of the results.

Before the arrival of the first test subject, preliminary measurements were made to ensure no design errors were left in the phases of the measurement, especially in Phase 3. The results of these measurements were not included in the final result set. We also thought about the possibility of measurement errors – such as device or network malfunctions, or even health issues with the test subject – during the measurements, so of course we prepared extra test subjects in case of the abortion of a measurement.

Section III

Mean Opinion Score analysis

After the results of the twentieth measurement were validated and found acceptable, we obtained the raw data set required to generate the MOS (see *Figure 2*).

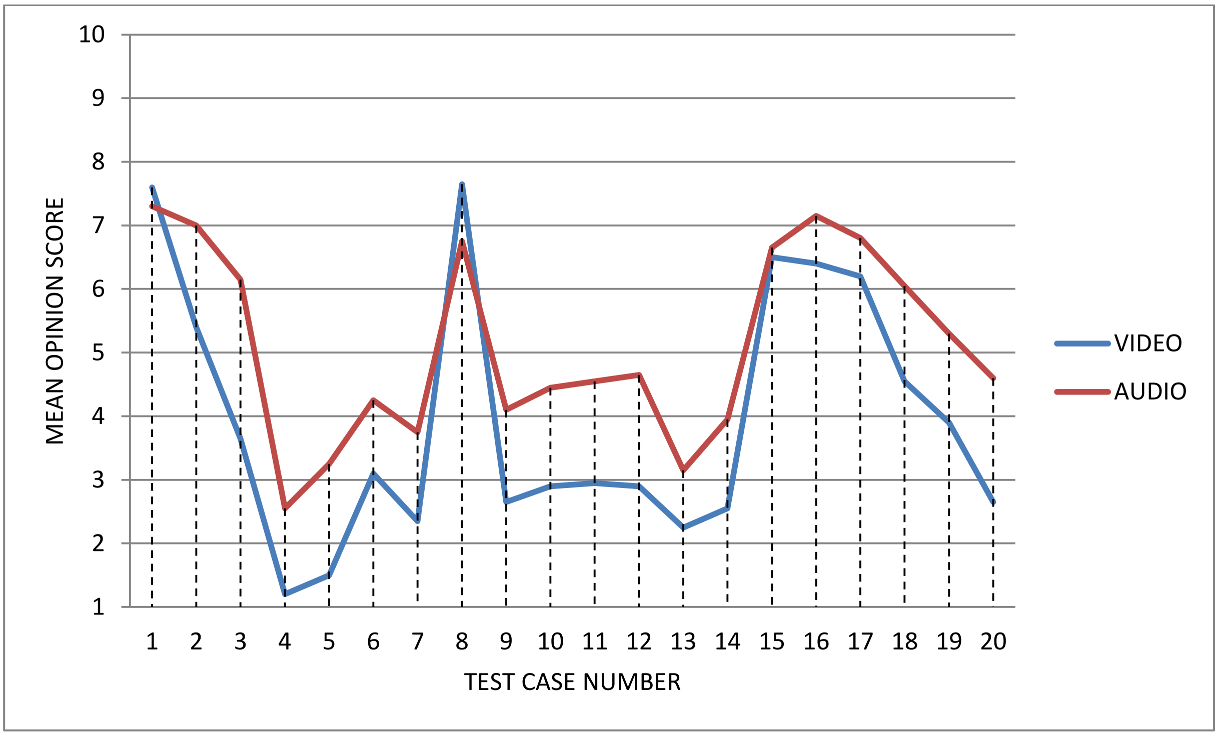

Figure 2 – Average Mean Opinion Score of the measurements in terms of video and audio

Before refining the video footage of Phase 1 and making an inspection of the results based on the different LoC levels, our initial approach was to analyze the MOS and to be on the lookout for anomalies. Such irregularities were quite easy to find.

As seen in the table of variable parameters, test case 1 was free of any artificial load; it suffered no additional delay, jitter or packet loss. Since all test cases were unique, and this test case had no load at all, then this should have received the highest score, or at least there should have not been any other test case with a higher score. This statement is pretty much reasonable. However, as mentioned in the introduction, *should* is the essential keyword here.

Let us take case 8 in contrast with case 1. Test case 8 suffered 800 milliseconds of additional delay; that means data coming from Room #2 to Room #1 was delayed for almost one second. The video and audio scores of case 1 should definitely be better than case 8, shouldn't them? Although the actual relationship between QoS and QoE is still not perfectly clear, this situation seems rather trivial.

Yes, it does seem trivial. However, reality can be a bit disappointing sometimes. In terms of video quality, while test case 1 received an average score of 7.6, test case 8 got 7.65. Even though the difference in scores is quite minor, we cannot escape the single fact that case 8 somehow achieved a better MOS value than case 1. The relation in scores should be the opposite. This phenomenon at this point of the analysis cannot be clarified.

This was not the only piece of the MOS result set which waged war versus expectations. From case thirteen to sixteen, the jitter was reduced in every step. That would mean an improvement in the experienced quality; MOS should be increasing or at least it should not be decreasing. The relation between the first three cases of this quartet was formed as expected; however, the fourth one received a slightly lower score than the third. Similarly to the previous phenomenon, this one cannot be explained either by relying solely on the MOS.

Later on we shall see that these were not the only abnormalities in the results. Already at this point it can be felt that the MOS is unquestionably distorted, but this simple approach is insufficient to discover its origins.

Before we move on to a more detailed analysis involving LoC, we could also examine the evaluation of the test subjects separately. In turns out that the majority of the test subjects did not think that test case 8 would be better than test case 1. In fact, there were quite some who thought that case 1 is indeed the better one. Only a few subjects gave case 8 a better score, yet it received a higher MOS score. Shouldn't the MOS reflect the evaluation of the majority instead of the minority? Fortunately our novel approach gave an explanation to phenomena like this one.

Analysis using LoC levels

By analyzing the recorded conversations of Phase 1, we managed to categorize the test subjects into ten different levels of LoC in a uniform manner. From this point forward, we were able to look at the results from another point of view, which helped us detect and analyze the distortion phenomena. Fully understanding them required the user observations from Phase 4.

Evaluation of the reference test case

Test case 1 endured no additional delay, jitter or packet loss, thus it was the reference test case of the measurement; an opportunity to witness the operation of the service with its own genuine QoS parameters. As it has already been proclaimed, this test case was meant to have the highest score. Before investigating its relation with case 8, let us examine what the test subjects thought.

Evaluation was performed on a discrete, label-free scale from 1 to 10. Although it allows the test subjects to distinguish ten different scores, it also discourages them from using the lowest and the highest score. On a scale from 1 to 5, subjects are somewhat forced to choose 1 and 5 to express their thoughts of quality in case of many test cases. Using a scale from 1 to 10 gives them the opportunity to preserve 1 and 10 for the real extremes. As heard in some conversations of Phase 4, subjects suffered from the fear of limitations in evaluation. If one used 10 to evaluate a specific test case, but later met an even better one, it would not be possible to express the relationship between them; one would actually be forced to evaluate it with a 10 since there is no 11. The same postulation applies to 1.

Due to this matter at hand, many test subjects denied to give test case 1 an exceptionally high score. This behavior is essentially related to LoC differentiation. Those who realized the nature of test case 1 at the beginning gave it a higher score compared to the evaluation of those who did not grasp this thought in time or at all. While subjects of the lower levels gave a score of 6 or 7, those with higher levels gave a higher a score. Not surprisingly, the highest LoC level produced the highest score.

The problematic issue with this is that if a subject gives a reference test case a low score, but realizes later on that this was meant to be the best, then the subject compresses his/her thoughts into a smaller scale; for example it would easily transform a scale from 1 to 10 into a scale from 1 to 7.

It was heard during Phase 4 that quite a lot of test subjects – typically from the upper half of the LoC scale – realized that there should be nothing better than test case 1, but this realization was too late and it was not possible to go back and change the evaluation; once a score was made and the measurement advanced to another, modification was forbidden.

Anomaly between test case 1 and 8

The difference in video quality between the two given test cases was noticeable, but rather small. As seen during the initial analysis without LoC levels, only a lesser percentage of the test subjects claimed test case 8 to be better. After inspecting the results again with the LoC approach, it was made clear that these subjects actually belong to the lowest levels of LoC. If we subtract their scores from the MOS (see *Figure 3*), the distortion and its origin becomes visible.

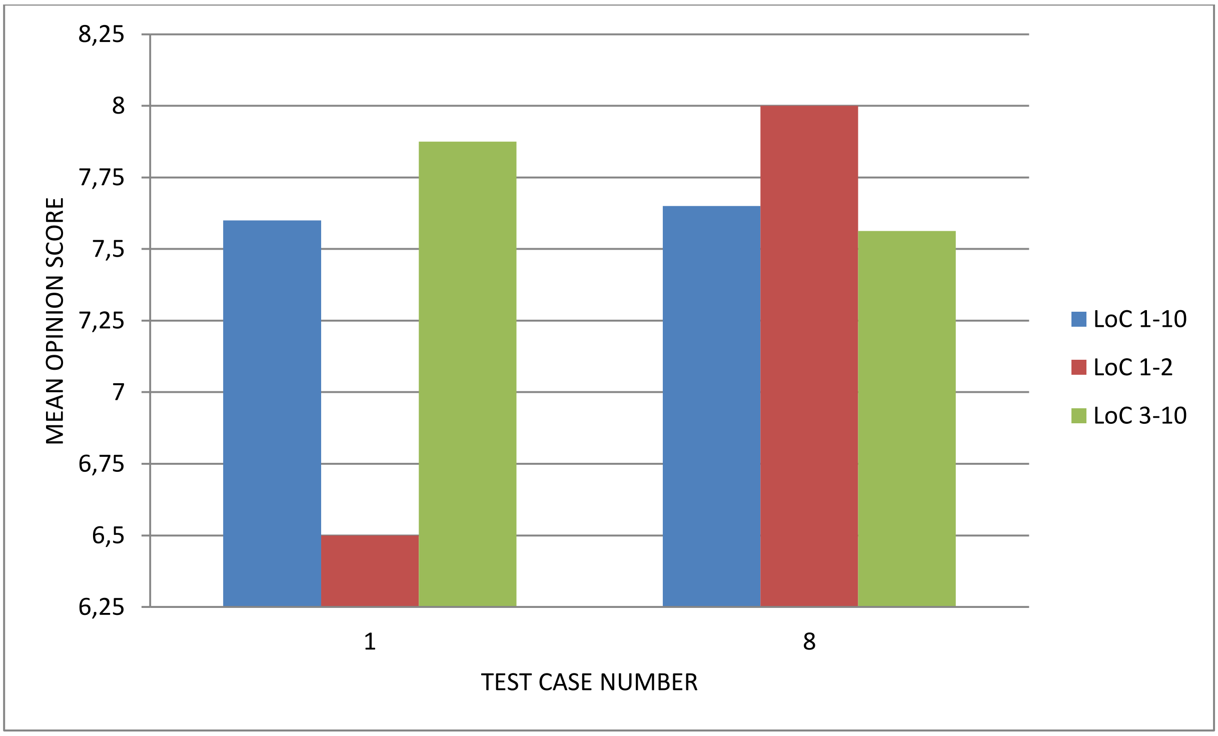

Figure 3 – Comparison of the video quality of test case 1 and 8

The lowest two levels made such a significant difference in scores – 1.5 points – that it was able to derail the whole MOS in this relation, resulting distorted results. How could such a distortion occur?

The answer lies in false preconceptions and cognitive dissonance. By the time test subjects belonging to medium and high LoC levels reached test case 8, they knew that nothing was supposed to be better than test case 1, especially a severely delayed case like 8. Subjects of the lowest levels were not bound by such ideas; nothing prevented them from making an evaluation in this relation. However, the lack of information alone would not create such a distortion.

As heard during the conversations of Phase 4, most of these subjects were led by the fabricated presumption that delay is beneficial to video quality; they actually thought that a higher delay would result in a better QoE. Misguided expectations like this one probably would not be a problem when unaided; when the test subject experienced the opposite, a reconsideration of ideas would occur.

The real problem is that such ideas were in fact aided. It was supported by cognitive dissonance. In this case, there were two conflicting cognitions [4]. One is the prior idea about delay, the other is the perception of quality. Since they were in contradiction, a state of dissonance [4] was created.

A healthy human mind immediately demands dissonance reduction [4] in case of state of dissonance. It is an automatic safety mechanism with the purpose of protecting internal harmony. Without it, one would be defenseless against undesirable, depressing feelings.

In this case, it was the battle of an idea versus a perception. In order to eliminate dissonance, one of them needed to change. If the idea had changed, it would have been the recognition of defeat, failure, which is not a fitting outcome. As mentioned earlier, the experienced difference between the two test cases was rather small, almost insignificant. Due to this insignificance, the test subjects were able to alter perception instead. They convinced themselves that not only case 8 was not worse, but in fact it was better.

On the other hand, the evaluation of some test subjects belonging to higher levels was distorted as well due to similar reasons. There were subjects at these levels who did not experience any significant dissimilarity between these two cases, but felt the urge to make a difference in scores, at least with 1 point. The condition of cognitive dissonance was very similar to the previous one. There was the idea that test case 8 should have been worse due to the QoS load, and the perception that the difference was not as noteworthy as expected. Although the influence of conflicting cognitions was not that intense in this case, but still it came with a state of dissonance which required reduction.

Uniform evaluation

Let us stay focused a little more on the evaluation of test subjects belonging to medium and high LoC levels. These subjects were more resistant against false preconceptions; they were less likely to show vulnerability towards misbelieves.

The first four test cases together demonstrate a general degradation of a service quality; all QoS loads rise until they reach unbearable heights. Let us now take a closer look at the evaluation and use the previous approach of result separation. The bottom levels of LoC shall be separated from the rest (see *Figure 4*).

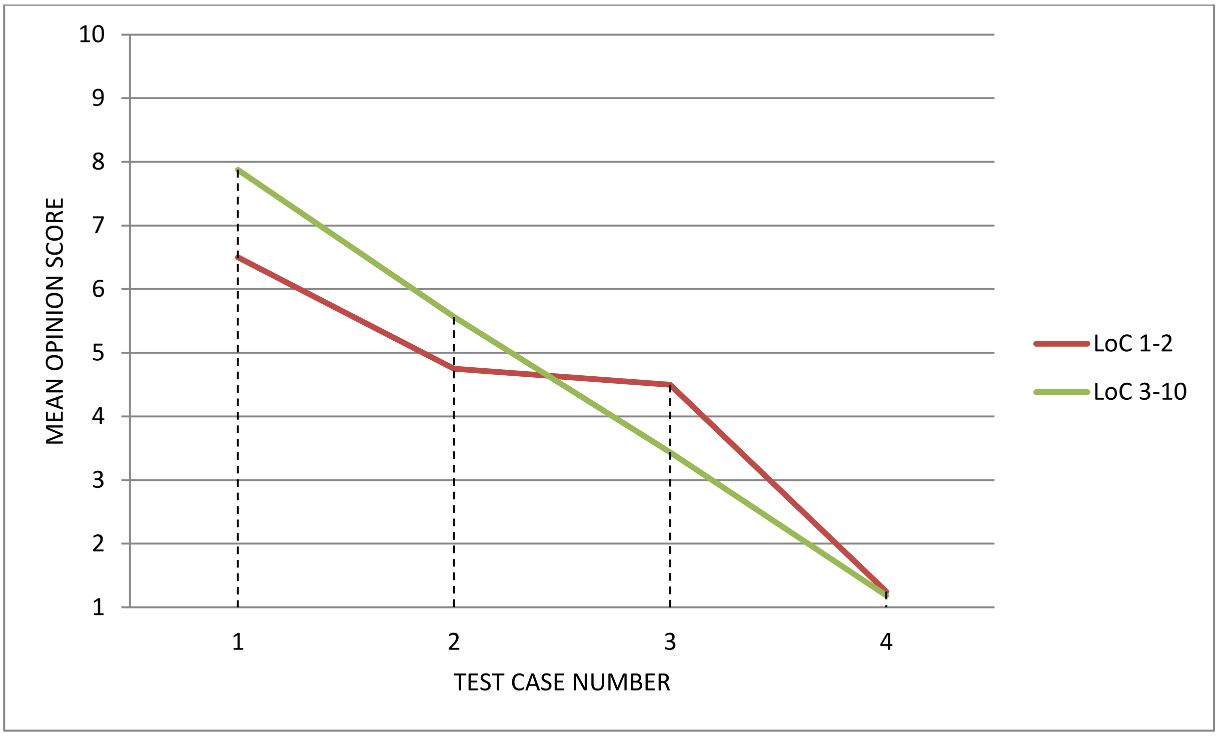

Figure 4 – Uniform evaluation in video quality

What we have encountered here is simply marvelous. It seems that test subjects were able to recognize the intention of degradation in this quartet of test cases; they have foreseen the transformation of QoE from best to worst. This process was considered to be somewhat uniform, so their evaluation became uniform as well.

Some could say that the evaluation performed by the lowest levels was closer to undistorted reality. It could be said, however, in their case cognitive dissonance was more involved. Test subjects of medium and high levels did not need to convince themselves much about their perceptions, since uniformity could be experienced without a state of dissonance, even though the actual difference between case 2 and 3 was slightly smaller than between case 1 and 2, and case 3 and 4. The judgment of test subjects belonging to the lowest levels was not clouded by the idea of uniformity. However, instead, they were deluded by the idea of beneficial delay. Since the difference between case 2 and 3 was indeed a little smaller, they thought it was because of the rising delay. To preserve this thought, they made an overestimation of test case 3 in order to minimize the difference in scores.

Evaluation during jitter reduction

Alongside the distortions considering test case 1 and 8, four cases of jitter reduction were mentioned during the MOS analysis; from test case 13 to 16.

The first thing that became visual after the determination of LoC levels was the relation between test case 1 and 15 (see *Figure 5*). As a reminder, test case 1 was a reference test case since its QoS parameters were genuine, while test case 15 suffered moderate loads in both delay, jitter and packet loss. While the test subjects of LoC levels 1 and 2 gave case 15 a better score, everyone else thought the opposite.

This was also caused by the lack of control provided by factual preconceptions. According to the configuration tests performed before launching the series of measurements with test subjects, while test case 14 barely reached the lower edge of usability, test case 15 came with a rather decent view. Due to the colossal difference between the two adjacent test cases, test subjects of the lowest LoC levels were convinced that test case 15 provided outstanding video quality.

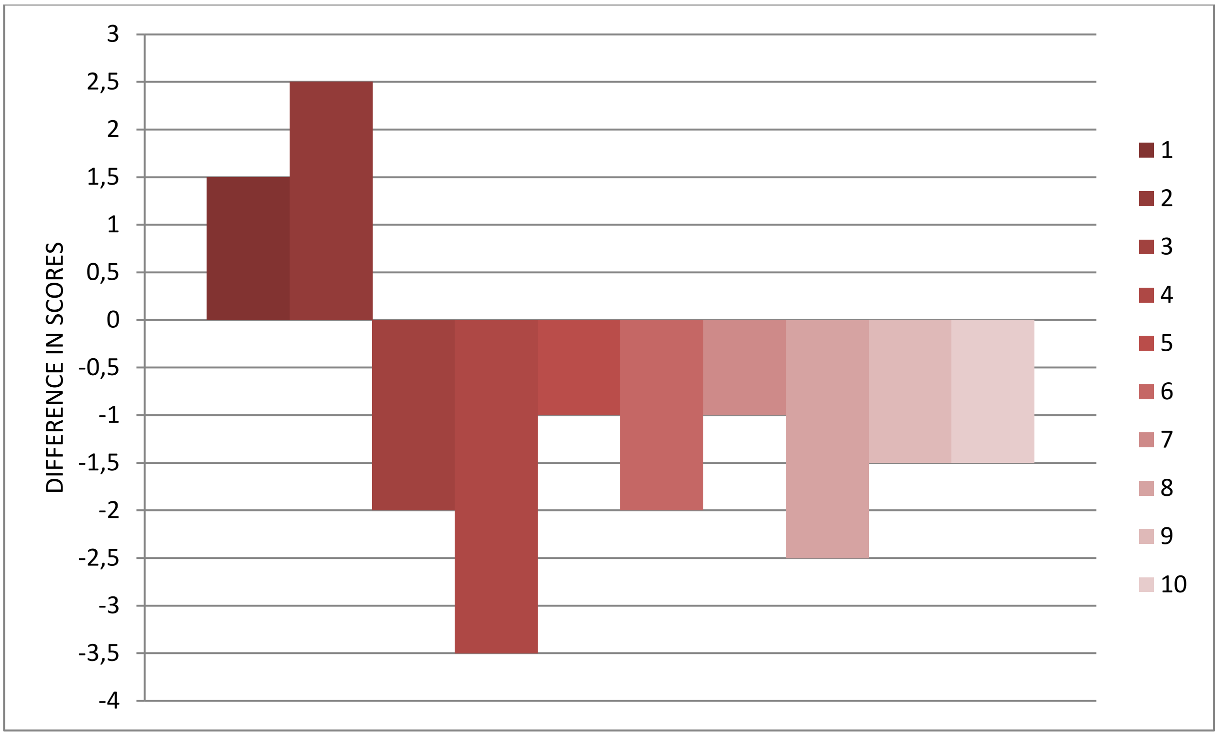

Figure 5 – Difference in video quality between test case 1 and 15

A **positive** value means that test case 15 received a **higher** score than test case 1.
A **negative** value means that test case 15 received a **lower** score than test case 1.
Distinct colors represent the evaluation difference of different LoC levels.

They thought that the quality of test case 15 was absolutely superior; better than any other prior test case, including test case 1. However, these test subjects were not affected by cognitive dissonance, since there were no conflicting cognitions.

In this scenario, cognitive dissonance rather played a part in the evaluation of others. The assumption was that nothing could deliver a quality better than what test case 1 had to offer. However, perception did have a little conflict with the assumption, because of the sudden advancement in quality. They needed to make a notable alteration to satisfy the condition of dissonance reduction.

The distance between the evaluation difference of LoC level 2 and 4 is simply astonishing. While level 2 provided results declaring that test case 15 is better by 2.5 points, level 4 made assessed in the opposite direction, by 3.5 points. That is a distance of a staggering 6 points.

This was not the only occurrence of cognitive dissonance in these four test cases. It took intense measures in the evaluation of test case 13 and 14. Although there was indeed a significant lessening in jitter, it did not make a noteworthy difference in video quality.

For the test subjects of the lowest LoC levels, the phenomenon of jitter was quite challenging to comprehend. Because of this, they were unable to create preconceptions based on jitter adjustment alone. This resulted in an unvarying yet honest evaluation, since both test cases represented nearly the same poor quality.

Unlike these test subjects, almost everyone belonging to higher levels claimed test case 14 to be better. This was a very clean case of cognitive dissonance. The preconception was that such an improvement in QoS needed to result in a better quality, and the perception was that there was nearly no difference. Not surprisingly, perception was the cognition to be altered.

As mentioned earlier in the analysis which relied only on the MOS, test case 15 somehow obtained a slightly higher score than case 16. A few test subjects of medium and high levels rated test case 16 with a better score, while some others gave them the same score. The decision of those who claimed case 16 to be better could be explained with conflicting cognitions, but what is more important here is how the subjects of the bottom levels evaluated.

The explanation is quite simple. Test case 15 was so shockingly superior compared to the six prior test cases – especially test case 13 and 14 – that it was considered to be the best. It even obtained a better average score than the previously praised test case 8; while case 8 received an average score of 8 from these LoC levels, case 15 got 8.5. Due to the admired quality of test case 15, even though case 16 was more or less the same, it was given a lower score. Yet again, the evaluation of the lowest LoC levels was powerful enough to have an observable influence on the MOS.

Constant evaluation

The focus of the analysis so far has been on video quality, and most of the target test subjects were the ones with lesser prior technical knowledge. Audio evaluation results hold an excellent example how much the other end of LoC can be distorted.

From test case 9 to 12, the amount of additional jitter was reduced. In audio quality, that is quite beneficial, since a high delay can easily create bothering discomfort in a conversation. The most common way is mutual speech interruption. In a regular, real-life, face-to-face conversation, the channel is typically used by one participant at a time; when one finds the channel empty, one can begin transferring vocal data, and the other waits until the transfer is completed. In case of a channel with nearly one second of delay, both participants can begin since the channel was found empty, which results in collision. To avoid this, participants sometimes manually insert pauses after the recognition of channel quality, but if both participants do it, then they are back to tile one. Delay reduction alone does improve audio quality; however, a constant high jitter was present in these test cases, which meant the escalation of the jitter/delay ratio.

If we have a look at the audio MOS, we can see that quality was assessed to be rising. At first sight, it could be assumed that the majority of different LoC levels shared this way of evaluation. However, there was only one level with a matching pattern of assessment (see *Figure 6*).

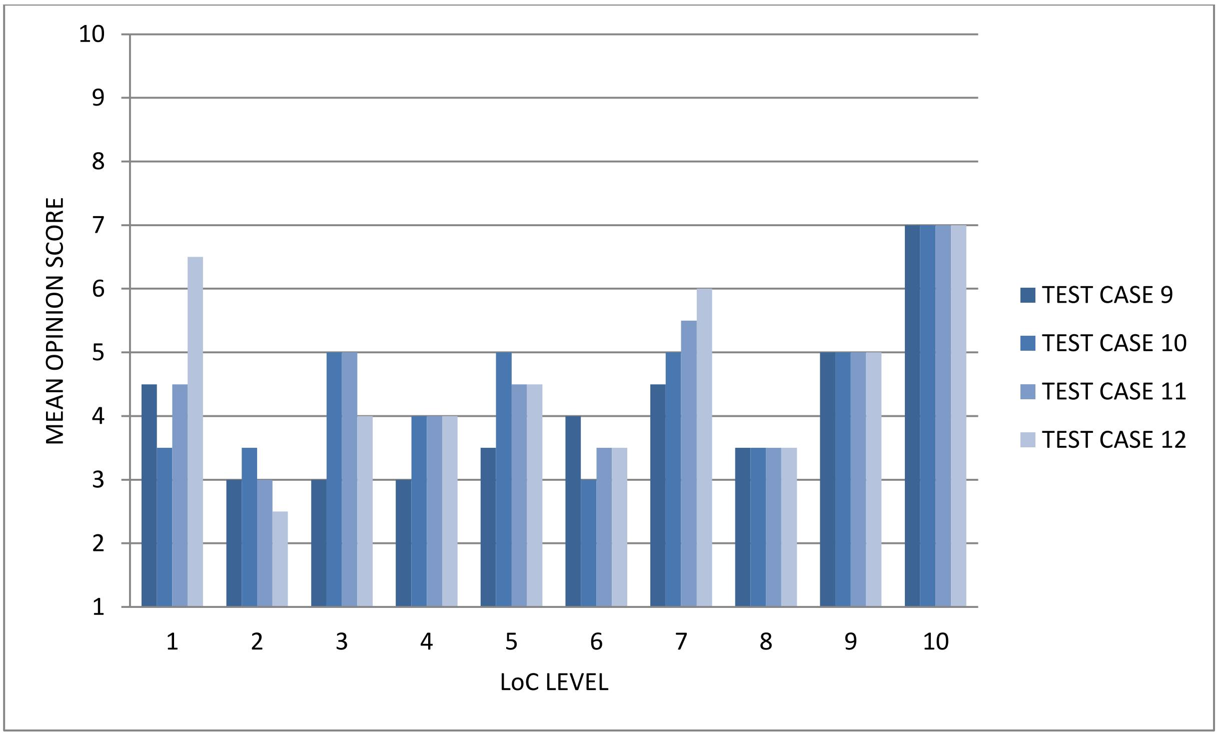

Figure 6 – Audio evaluation patterns from test case 9 to 12

As it can be seen, only level 7 produced matching test case relations to the MOS. From level 1 to 6, evaluation seems totally random. The reason is the complexity of the situation. These test subjects did not have exact expectations towards the quality, no preconceptions were generated, and thus the cognition of perception was left without a conflicting opponent. Test subjects scored according to their own personal experience of quality.

On the other hand, the evaluation of the top three LoC levels was as distorted as possible. Since they considered the opposing effects to be equal in intensity, they made a constant evaluation.

Indeed these test cases were not the same; while the user experience of test case 9 and 10 was ruined by mutual speech interruptions and pointless pauses, test case 11 and 12 had a lesser sound quality. It was up to the test subject to decide which of these effects was more irritating.

Test subjects with the highest levels of LoC suffered a great deal of cognitive dissonance. The idea of an unvarying overall quality had to face a perception providing feedbacks on altering audio experience. Although they did experience the aforementioned differences, they convinced themselves that their global level of satisfactory was not affected.

About the increment of score from LoC level 8 to 10, these levels followed a similar evaluation pattern during the measurement. Scores were sometimes different in power, but the relations between cases were comparatively the same. As mentioned earlier, the higher the level was, the higher the score of the reference test case was because of the aforesaid realization; test case one was ought to be the best due the genuine QoS parameters. Because of this, the evaluations performed by level 8 and 9 were compressed compared to level 10 since the scale was limited.

On the assessment of level 7, direction and uniformity shows a clear case of simple dissonance reduction, where one effect was considered to be superior to the other.

Cognitive dissonance of LoC levels

This is the point where it needs to be understood that in general test subjects belonging to the highest levels have to bare a much greater weight of cognitive dissonance than the ones of the bottom levels. Since they possess an immense extent of prior technical knowledge, they are the ones who can easily generate preconceptions to nearly everything. Because of their experience and familiarity with the topic, it is even harder to imply dissonance reduction with an undesirable conclusion.

Cognitive dissonance does not always conclude with a "happy ending". Sometimes perception can be strong enough to prove its validity. Such dissonance reduction can occur among those with lower levels of LoC. Since they are not or less involved in the current technology, no serious harm would come from accepting misbelieves. As we have seen in the examples of distortions, in case of minor differences, this automatic defense mechanism works perfectly.

However, low LoC level test subjects sometimes face penetrating conflicting cognitions. Because of the lack of experience and involvement, they are bound to have misbelieves at some extent. If perception is in a complete contradiction with preconception, and the rationality of perception cannot be questioned, preconception will have to surrender.

For instance, during the last four test cases, where packet loss was rising, only level 1 gave the second one – test case 18 – a better score than the previous one. By the time packet loss reached 4%, perception grew stronger than preconception, and the direction of evaluation was changed.

Choosing preconception over perception in cognitive dissonance comes with a threshold; if a specific level of difference is breached, the protection of idea will not apply. This threshold can be slightly increased by making an evaluation based on the same preconception during a prior test case, but with an even smaller difference. By the time the test subject reaches the other evaluation, cognitive dissonance is not only fuelled by the regarding preconception and the protection of idea, but also by the protection of decision.

As we've seen on the previous example, even though packet loss was escalating, level 1 evaluated audio quality in a positive way, but abandoned the idea after witnessing a greater difference. Since the decision was at 4% of packet loss, the actual threshold was between 2% and 4%. We assume, however, that the genuine threshold was increased by the protection of decision; because of the already existing decision supporting the preconception, it was even harder to let go of the idea.

The phenomenon of the aforementioned protection of decision is also known as post-decision dissonance [23] [24]. In a QoE evaluation measurement like ours, it evolves when the assessment becomes irreversible. Since we forbade the modifications of the results of elapsed test cases, when the measurement proceeded to the next test case, a form of post-decision dissonance would be formed if cognitive dissonance was involved in the assessment. Just as we assumed, this justifying response would establish a bond with the decision, and through it, this bond would empower the preconception, changing the threshold of perception validation. Although its initial purpose is to minimize the regret of this irreversible choice, when the threshold is passed, not only will the test subject face the error of the once justified preconception, but will also need to handle a misguided choice.

The theory of cognitive dissonance [4] states that conflicting cognitions cannot remain in a state of dissonance in case of a healthy human mind and soul. Dissonance reduction can be achieved by changing a dissonant cognition, lowering the importance of a dissonant cognition or by adding new elements [4].

In case of lower LoC level test subjects, lowering the importance of perception is not likely to be possible, since their preconceptions are not supported enough. Instead, they make minor alterations in perception and persuade themselves about their ideas.

In case of higher level test subjects, they have the opportunity to lower the importance of perception, due to their experiences and knowledge. As the example of the last four test cases has shown, sometimes perception does prevail over preconception. The problem with subjects of the highest levels is that their preconceptions are more or less accurate. They do not have false preconceptions to be forcefully confronted with perceptions. This leads to a small difference between ideas and reality. And as we have seen, nothing really prevents small differences from being ignored.

If we approach cognitive dissonance from attitudes and thoughts, there is one major difference between the edges of LoC. It lies in the structure of preconceptions. The idea of test subjects of the lower ones builds up as “*I **think** component A **should** result in quality B*.”, while in case of the highest levels it goes as “*I **know** component A **must** result in quality B*.”.

The medium levels of LoC are bound to be the ones with the least occurrence of cognitive dissonance. On one hand, they are less likely to be overwhelmed by false preconceptions and misbelieves, on the other hand, their decisions do not get enslaved by their high level of prior technical knowledge.

Cognitive dissonance of the different LoC levels is nearly impossible to evade. Because of the nature of QoE assessments, it would eventually appear in the evaluation of test subjects. But since QoE measurement results are such essential inputs of the industry, it should be at least attempted to be reduced or minimized.

One approach would be to eliminate the sources of preconceptions. Usually, test subjects are not provided the genuine QoS values or disturbances, or basically any technical information regarding the service. The problem is that some information cannot be cloaked. Let us imagine a hypothetical scenario where test subjects have to compare a given service in a stationary and a mobile environment. However, let’s add that the stationary one receives QoS disturbances to have exactly the same QoS parameters as the mobile one. The test subjects would not be informed about the synthetic alteration of QoS. Although it would probably result the exact same user experience, test subjects belonging to the highest levels of LoC would rate the stationary one better, and there could be some low level subjects who evaluate in the opposite direction due to some false ideas.

Another approach would be to measure and distinguish the different LoC levels, exclude the data generated by extremely high and low levels, and compare it to the genuine MOS. This idea was composed in the upcoming proposal.

Section IV

Proposal

Because of the high rate of occurrence of cognitive dissonance in the lowest and highest levels, their evaluations can be considered to be outliers. In attempt to generate a less distorted evaluation, we can take a look at the results without the extreme LoC levels (see *Figure 7*).

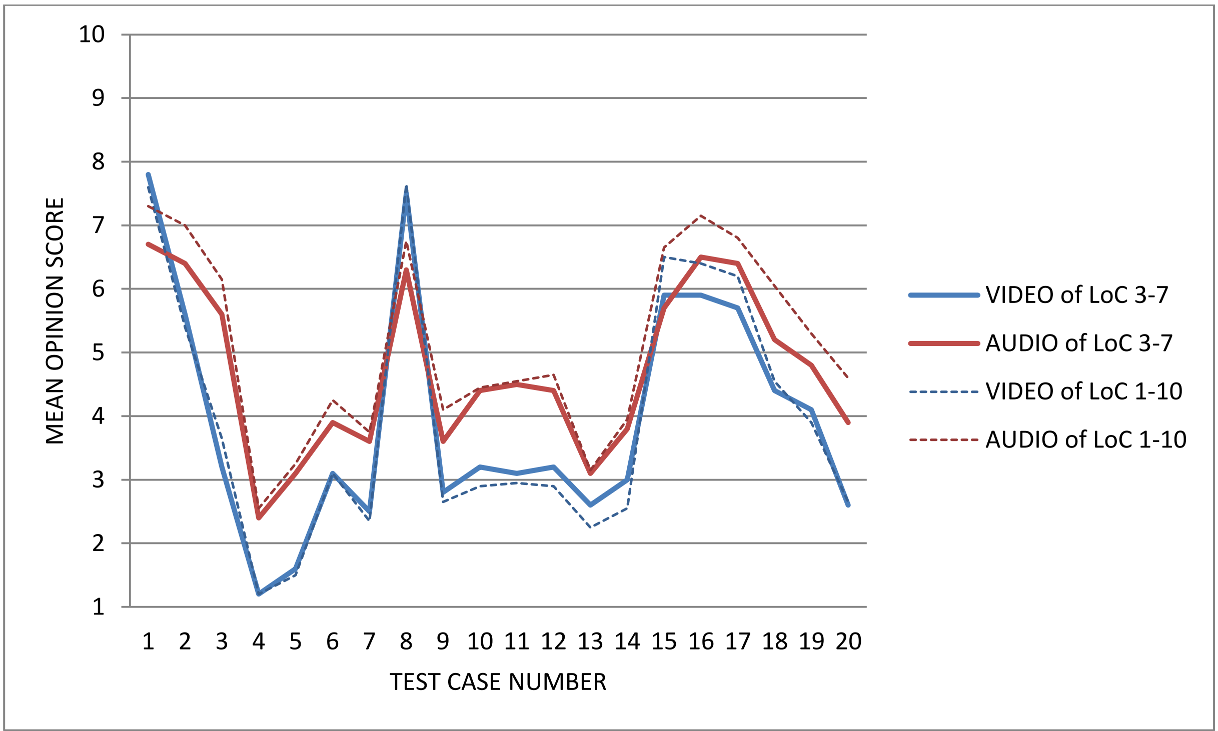

Figure 7 – Average Mean Opinion Score of LoC levels 3-7

By doing so, we eliminated the misleading anomalies created by false preconceptions, and removed unnatural scoring from the components of the MOS.

The **proposal** is the following: *We define a QoE observation as an outlier if it changes the relation between particular evaluation scores. We recommend determining the test subjects' Level of Comprehension and analyzing the MOS without the lowest and highest levels. If this subtraction of scores identifies outliers in the MOS, abandoning the scores of radical LoC levels could be indeed beneficial. In case subtraction does not reveal any outlier, initial results should be left untouched.*

Conclusions

The study has shown that preconceptions and cognitive dissonance can have a serious impact on the MOS. Such impacts are evaluation distortions.

Temporary setbacks in QoS due to the nature of the service – for instance a wireless environment – and its consequences on QoE are not distortions, since they will be present during everyday service usage. However, modified evaluation is indeed a distortion, because it is only present during the measurement. While the test subjects need to make an evaluation after receiving a limited sample of the service, users are not rushed to make an assessment. Besides, opinion can change over time; users have plenty of time to discover the service.

Because of the limitations and the rushed decisions, test subjects need to briefly prove themselves that their evaluation is correct. The basis of such proof can be the prior expectation about the quality of the service.

It is easy to say that if we want to prevent preconceptions from being created, let us avoid providing information like QoS parameters. The solution is not that simple, since even the smallest, most basic information can generate preconceptions; for example whether the consumption of the service is mobile or not.

The real problem is that preconceptions are aided by cognitive dissonance. Regardless whether these thoughts are actually accurate or complete nonsense, they get validated through dissonance reduction if the distance between the conflicting cognitions is small enough.

A restriction of our approach is the scalability issue of LoC determination. The method we used was rather time-consuming yet minimized the chance of errors. A measurement with a vast test subject number would require the reconsideration of the balance between efficiency and accuracy.

Our study proposes a method which could reduce the distortion of the MOS. Although it could even mean the disposal of a high percentage of the measurement results, but if it makes the results of the measurement more precise and accurate, it could be really worth it.

Possible continuations

The subject is far from being closed or depleted. Now that new technologies are spreading – for instance 3D video services – it would be interesting to see how distorted the MOS of a service evaluation can be where it is common that test subjects have not had the opportunity to previously intensely familiarize themselves with the technology. 3D video is indeed an interesting topic, since up to this date most people still prefer 2D experiences.

It is also a motivating challenge to design, test and compare methods for LoC determination in terms of efficiency and accuracy. Results could help overcome the scalability issue of our method.

Not to mention that a deeper analysis of distortions could be possible by separating the LoC from the initial configuration. That would mean the measurement of QoE with and without provided information on QoS. By adjusting the information available to test subjects, we could see how each specific piece of information modifies evaluation.

It is also possible to use this approach for equipment evaluation instead of service evaluation; to witness how information regarding the equipment alters its assessment.

Modeling threshold alteration is another possible continuation of this topic. Right now we can only assume for sure that a threshold T of assessment X_n will be increased to $T+\varepsilon$ through post-decision dissonance if exists a foregoing X_{n-1} with the same preconception. It would be interesting to see how the value of ε changes with the increment of preceding assessments.

The difference between various handover techniques using this approach is also worth analyzing. It could also highlight how mobility-awareness itself distorts QoE measurement results.

Last but not least, the usage of a pseudo-continuous scale during evaluation would be exciting. Test subjects would rely on a visual approach rather than qualitative labels or quantitative numbers. It would be interesting to investigate the scale's effect on the presence and influence of cognitive dissonance.

List of figures and tables

References

[1] International Telecommunication Union. Methods for subjective determination of transmission quality. ITU Recommendation P.800 (08/96), August 1996.

[2] International Telecommunication Union. P series: Terminals and subjective and objective assessment methods. ITU-T Recommendations, P series. http://www.itu.int/rec/T-REC-P/en, last visited: 2012-10-25.

[3] P. A. Kara, L. Bokor, S. Imre. Distortions in QoE measurements of ubiquitous mobile video services caused by the preconceptions of test subjects. In: IEEE/IPSJ International Symposium on Applications and the Internet SAINT2012. Izmir, Turkey, 2012.07.16-2012.07.20. IEEE, pp. 409–413.

[4] Festinger, L. (1957). A theory of cognitive dissonance. Stanford, CA: Stanford University Press.

[5] Aronson, E. (1969). The theory of cognitive dissonance: A current perspective. In L. Berkowitz (Ed.). Advances in Experimental Social Psychology, Vol. 4, pp. 1–34. New York: Academic Press.

[6] International Telecommunication Union. Terms and definitions related to quality of service and network performance including dependability. ITU Recommendation P.800 (08/94), August 1994.

[7] ISO/IEC 9126-1:2001 Software engineering — Product quality — Part 1: Quality model.

[8] L. Alben. Quality of Experience: Defining the criteria for effective interaction design. ACM Interactions, vol. 3, pp. 11–15, 1996.

[9] D. Soldani, M. Li, and R. Cuny. QoS and QoE Management in UMTS Cellular Systems. John Wiley and Sons, August 2006.

[10] H. Ridder, R. Hamberg. Continuous Assessment of Image Quality. SMPTE Journal, pp. 123–128, February 1997.

[11] P. Brooks, B. Hestnes. User Measures of Quality of Experience: Why Being Objective and Quantitative Is Important. Network, IEEE, Vol. 24, No. 2. (March 2010), pp. 8–13.

[12] A. Watson, M. A. Sasse. Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications. In MULTIMEDIA '98: Proceedings of the sixth ACM international conference on Multimedia (September 1998), pp. 55–60.

[13] Narita, N. Graphic scaling and validity of Japanese descriptive terms used in subjective-evaluation tests. SMPTE Journal, July 1993, pp. 616–622.

[14] Jones, BL. & McManus, P.R. Graphic scaling of qualitative terms. SMPTE Journal, November 1986, pp. 1166–1171.

[15] M. Fiedler, T. Hossfeld, P. Tran-Gia. A Generic Quantitative Relationship Between Quality of Experience and Quality of Service. IEEE Network, Vol. 24, No. 2. (March 2010), pp. 36–41.

[16] Y. Lu, Y. Zhao, F. Kuipers, and P. Van Mieghem. Measurement study of multi-party video conferencing. In Proceedings of the 9th IFIP TC 6 international conference on Networking (NETWORKING 10), pp. 96–108, 2010.

[17] I. Ketykó, K. De Moor, W. Joseph, L. Martens, and L. De Marez. Performing QoE-measurements in an actual 3G network. In IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB'10), pp. 1–6, March 2010.

[18] BME-MIK. Budapest University of Technology and Economics - Mobile Innovation Centre, Official Website. https://www.mik.bme.hu/home/aboutus/, last visited: 2012-10-25.

[19] Linphone. Official Website. http://www.linphone.org/,
last visited: 2012-10-25.

[20] 3GPP TS 23.228. IP Multimedia Subsystem (IMS); Stage 2. Rel-8, 2008.

[21] netem. Linux Network Emulation OfficialWebsite.
http://www.linuxfoundation.org/collaborate/workgroups/networking/netem,
last visited: 2012-10-25.

[22] Dianne Dukette; David Cornish (2009). The Essential 20: Twenty Components of an Excellent Health Care Team. RoseDog Books, pp. 72–73.

[23] Knox, R. E., & Inkster, J. A. (1968). Postdecision dissonance at post time. Journal of Personality and Social Psychology, 8(4), pp. 319–323.

[24] P. Rosenfelda, J. G. Kennedyb & R. A. Giacalonec. Decision Making: A Demonstration of the Postdecision Dissonance Effect. The Journal of Social Psychology, Volume 126, Issue 5, 1986.