The Importance of Assessing Quality of Experience of IPTV and Video on Demand Services in Real-life Environments

Nicolas Staelens, Stefaan Moens, Wendy Van den Broeck, Ilse Mariën, Brecht Vermeulen, Peter Lambert, Rik Van de Walle, and Piet Demeester, *Fellow, IEEE*

***Abstract*—The ever growing bandwidth in access networks, in combination with IPTV and Video on Demand (VoD) offerings, opens up unlimited possibilities to the users. The operators can no longer compete solely on the number of channels or content and increasingly make High Definition channels and Quality of Experience (QoE) a service differentiator. Currently the most reliable way of assessing and measuring QoE is conducting subjective experiments, where human observers evaluate a series of short video sequences, using one of the international standardized subjective quality assessment methodologies. Unfortunately, since these subjective experiments need to be conducted in controlled environments and pose limitations on the sequences and overall experiment duration they cannot be used for real-life QoE assessment of IPTV and VoD services. In this article, we propose a novel subjective quality assessment methodology based on full length movies. Our methodology enables audiovisual quality assessment in the same environments and under the same conditions users typically watch television. Using our new methodology we conducted subjective experiments and compared the outcome with the results from a subjective test conducted using a standardized method. Our findings indicate significant differences in terms of impairment visibility and tolerance and highlight the importance of real-life QoE assessment.**

***Index Terms*—Quality of Experience (QoE), Subjective Video Quality Assessment, Video on Demand, IPTV.**

I. Introduction

IP based networks, such as the Internet, are more frequently used for the delivery of high quality video services, e.g. Internet Protocol Television (IPTV) and Video on Demand (VoD). However, due to the packet-based best-effort characteristics of these IP networks, offering enhanced video services can be a real challenge for service providers. Network impairments such as packet loss and jitter can severely degrade audiovisual quality during playback which, in turn, will influence end-users satisfaction of the video service. According to [1] there is a positive correlation between the willingness to pay and the QoE of the video stream offered to the user. Even while there is large uncertainty on the value of this correlation, the study clearly shows that users are always inclined to pay less if they are offered a video stream with a lower QoE. When users feel they are overpaying their service in regards to the quality they experience, they will react in different ways - reducing their usage of extra paying services such as VoD or specific channels, spread negative publicity, cancel their subscription, repeatedly call your help-desk or customer service department, etc. - all of which will eventually lead to a decrease in revenues for the operator from those customers. Clearly the video service providers will want to ensure and maintain a satisfactory quality at all times (and preferably an even better quality to be sure). Measuring and monitoring audiovisual quality and gaining an insight into the way end-users react on and tolerate audiovisual degradations become as such important fields-of-interest for the service providers.

From a network point of view, quality can be measured, monitored and optimized through Quality of Service (QoS) in terms of packet loss, delay, jitter, available bandwidth, etc. [2]. However, it is generally known that the impact of network impairments on audiovisual quality also depends on the characteristics of the video stream. As such, the same amount of packet loss will result in different audiovisual degradations depending on the video encoding, network packetization and video content. The latter is typically characterized in terms of the amount of motion and spatial detail (e.g., textures, edges, ...). Visual impairments will likely be more perceived in high motion areas or in areas with a low amount of spatial detail [3], [4]. In general, only measuring objective QoS parameters is insufficient to reliably estimate end-users' subjective overall perception of quality, commonly referred to as Quality of Experience [5], [6].

End-users' perception of quality is a highly subjective matter and a lot of research is currently ongoing to construct new quality metrics which try to accurately measure QoE. The experience of users is foremost subjective and rather complex, as it is influenced by different contextual factors (e.g. environmental conditions and social circumstances). QoE is not static, as *"the quality of peoples experience changes over time as it is influenced by variations in these multiple contextual factors"* [7]. Therefore it is important to take these contextual factors into account when measuring QoE. Prior to the construction of such quality metrics, subjective experiments are usually conducted in which human observers are required to provide a visual or audiovisual quality rating for a series of video sequences. The results of these experiments serve as ground truth for the construction and validation of new

N. Staelens, B. Vermeulen and P. Demeester are with Ghent University - IBBT, Department of Information Technology, Ghent, Belgium (e-mail: {nicolas.staelens,brecht.vermeulen,piet.demeester}@intec.ugent.be).

S. Moens, P. Lambert and R. Van de Walle are with Ghent University - IBBT, Department of Electronics and Information Systems, Ghent, Belgium (e-mail: {stefaan.moens,peter.lambert,rik.vandewalle}@ugent.be).

W. Van den Broeck and I. Mariën are with the Free University of Brussels - IBBT, Studies on Media, Information and Telecommunication, Brussels, Belgium (e-mail: {wvdbroec,imarien}@vub.ac.be).

QoE metrics. There already exists a variety of international standardized subjective quality assessment methodologies that describe in detail how such subjective experiments must be set up and conducted.

According to the definition in ITU-T Recommendation P.10/G.100 Amendment 2, QoE includes the complete end-to-end system effects and can be influenced by user expectations and context. Subjective quality experiments are, however, conducted in controlled environments where subjects receive specific instructions on how to evaluate the video sequences. Watching television, on the other hand, is known to be a typical lean-backward social experience where people are watching a movie in their living room with friends or family. Video sequences are watched and evaluated very differently during subjective experiments compared to watching television. But QoE includes more than only the aspect of audiovisual quality; the overall end-user experience of IPTV and VoD also includes factors such as channel zapping delays, application responsiveness, ease of use, content availability, ... which all contribute to the overall acceptance and success of new video services. Therefore, since QoE depends on user expectations and context, measuring and assessing the QoE of a particular video service such as IPTV or VoD should be performed in the most natural environment where these services are typically consumed.

In this article, we investigate the difference between assessing QoE using standardized subjective video quality assessment methodologies and conducting QoE tests in real-life environments using a novel methodology based on full length movies. This new methodology encourages subjects to watch the video sequence in the same environment they normally watch television without actively evaluating audiovisual quality. The results obtained using our new methodology show significant differences from those of the standardized subjective experiments.

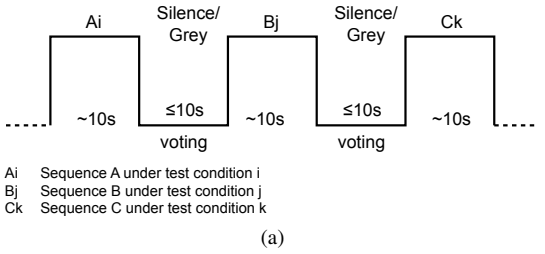
The remainder of this article is structured as follows. We start by motivating the need for a new methodology, in Section II, by describing existing subjective quality assessment methods and highlighting their major drawbacks. Based on these shortcomings, we define the requirements for a new subjective methodology which enables real-life QoE assessment and propose a novel subjective video quality assessment methodology. In Section III, we describe the subjective tests that were conducted in order to validate our new methodology and present the results of this study in Section IV. Finally, we conclude the article.

II. Towards Real-life QoE Assessment

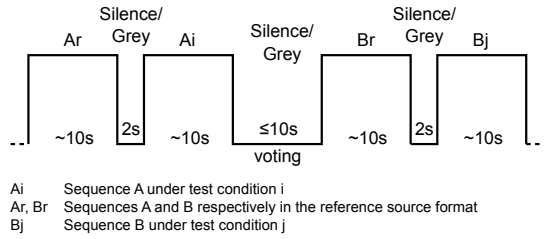
Subjective video quality assessment, which involves real human observers watching and evaluating the (audio)visual quality of a number of short video sequences, is currently the most reliable way of obtaining real quality ratings. Subjective assessment methodologies which describe how to conduct such experiments in controlled environments have already been standardized and are still widely used. However, new methodologies are needed which enable more realistic QoE assessment.

A. The need for a new subjective quality assessment methodology

International standardized subjective video quality assessment methodologies as specified in ITU-T Recommendations P.910 and P.911 and ITU-R Recommendation BT.500-11 include detailed guidelines on how to organize and conduct video quality experiments. These recommendations describe both Double Stimulus and Single Stimulus test methods, which define the overall trail structure. In case of a Single Stimulus test, shown in Fig. 1a, sequences are displayed without an explicit reference and subjects are required to provide a quality rating for each individual sequence. A Double Stimulus test implies that sequences are shown pairwise. The first sequence presented to the subjects is the reference sequence which serves as reference point for optimal quality. Then, an impaired or degraded version of the same sequence is displayed. After each pair of sequences, the viewer is required to rate the degradation/impairment visibility between the second and the reference sequence. This trail structure is depicted in Fig. 1b.



(a)



(b)

Fig. 1. Typical trail structure of a Single Stimulus (a) and a Double Stimulus (b) subjective quality experiment which define the order sequences are displayed and rated by the subjects, as specified in ITU-T Rec. P.910 and P.911.

These assessment methodologies do not only specify the order in which the sequences must be presented but they also limit the maximum duration of the video sequences (typically between 10 and 15 seconds). In case longer video sequences (up to 30 minutes long) must be evaluated, continuous quality evaluation can be used as defined by the Single and (Simultaneous) Double Stimulus Continuous Quality Evaluation methods, standardized in ITU-T Rec. P.911 and ITU-T Rec. BT.500-11. During such a test, subjects continuously have to rate quality during video playback, by means of moving a slider, instead of providing a quality score after the entire sequence has been viewed.

Prior to the start of a subjective experiment, subjects receive specific instructions on how to watch and evaluate the video sequences. As a result, subjects are highly focused on detecting visual degradations during playback. Therefore, in order to

avoid viewer fatigue and boredom, the entire duration of a subjective experiment should be limited to 30 minutes.

Finally, the subjective test methods also specify specific viewing conditions such as the viewing distance as a function of the picture height, monitor resolution, display brightness and contrast and overall test room conditions. More detailed information concerning subjective video quality testing and a comparison of the different methodologies is provided in [8] and [9]. These methodologies have already been widely adopted and are still used in literature to study, for example, the effects of video encoding and video transmission on the perceived quality of end-users [10].

However, the controlled environments in which these subjective experiments need to be conducted are not reflective of a consumers natural environment for watching television. There are significant differences between watching television and participating in a subjective test. Television is primarily watched for leisure and is known to be a typical lean-backward experience [11]. People watch television programs and movies, together with friends or family, for their content or plot. In contrast with conducting subjective tests, people are not primarily concentrated on audiovisual quality evaluation while watching television. When conducting subjective experiments, the duration of both the sequences and the overall experiment are limited. Rating the quality of a sequence which contains visual degradations will depend on the length of that particular video sequence. For example, two impairments in a 10 seconds sequence will be perceived differently in a 30 minutes sequence. The same holds when the two degradations occur during movie playback. It is also clear that the typical home environment, where people watch television, is a highly uncontrolled environment. The stringent test environment demands, which are imposed by the existing methodologies, in terms of viewing distance and screen quality and calibration can thus be very different compared to a living room environment.

As such, user expectations and context differ substantially between subjective video testing and watching television. This will influence perceived QoE, according to its definition. Therefore, a new subjective methodology is needed for measuring the QoE of IPTV and VoD services in real-life environments.

B. Full Length Movie Quality Assessment Methodology

In order to enable real-life QoE assessment we are particularly interested in a new subjective methodology which (1) encourages subjects to watch the video sequences in the same environment they normally watch television and (2) which is capable of mimicking the lean-backward TV viewing experience. By doing so, we want to avoid that subjects are already biased and try to actively evaluate audiovisual quality. Since evaluating the influence of audiovisual degradations, caused by failures (e.g., packet loss) in the delivery chain from service provider to end-user, is an important aspect of QoE assessment, (3) the new methodology must also allow the controlled insertion of such degradations during playback. Since we want to test IPTV and VoD services (4) the quality of the content, which is shown to the test subjects, must be representative and comparable with current IPTV content. Finally, the new methodology must also (5) provide a means for collecting feedback from the subjects concerning the (audio)visual quality of the video sequence(s).

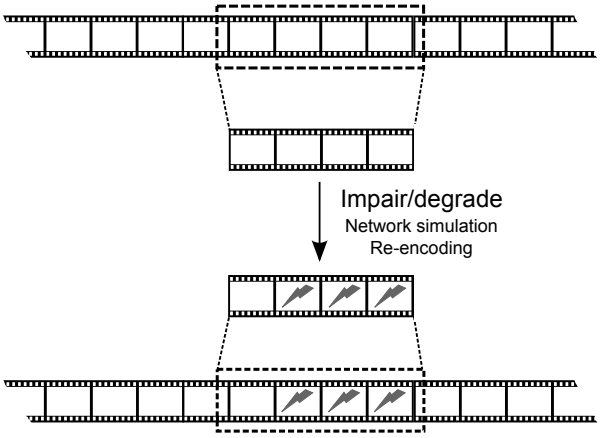


Fig. 2. Toolchain for creating impaired full length DVD movies. Visual impairments are inserted in selected parts or scenes of the movie. The impaired scenes are then re-inserted into the full movie.

Taking into account the different requirements imposed on the new subjective methodology, we propose the use of full length DVD movies for real-life QoE assessment. By using DVDs, subjects can take the disc home where they will watch it most likely in the same environment and under the same conditions they watch television. Using full length movies we also encourage subjects to watch the movie for its content, not for its audiovisual quality[1]. As such, we want to create a realistic lean-backward TV viewing experience. Furthermore, subjects tend to take DVD quality as reference when judging the QoE of video services such as VoD [12].

Evaluating the influence of audiovisual degradations during movie playback can be performed by impairing the movie and writing the resulting damaged movie onto a new DVD. This enables the controlled insertion of impairments in the video and/or the audio track of the movie. Our methodology can thus be used for conducting auditive only, video only or audiovisual subjective quality assessment tests. In order to insert visual degradations into the movie we used the toolchain depicted in Fig. 2. First, we extracted the part or the scene from the movie which we wanted to impair. To inject network impairments into the fragment we used our in-house developed modular multimedia streamer xStreamer [13]. Other different tools can be used to degrade the selected part. Coding errors can be inserted by re-encoding the selected part at a lower bit rate or using different encoders. It is important to mention that when converting the movie or the selected fragments back into a compliant DVD movie, the encoding bit rate was set as high as possible in order to avoid additional coding impairments.

Since our new methodology is focusing on real-life QoE assessment, we do not want to inform our subjects about any possible degradations during playback but we still want to

[1]During subjective tests in a controlled lab environment, viewer fatigue appears after 30 minutes. This implies that subjects are not capable of actively assessing audiovisual quality of full length movies.

collect user feedback immediately after watching the movie concerning perceived quality. Therefore, we use questionnaires which are provided together with the DVD to the subjects. The questionnaire is enclosed in a sealed envelope and we strongly insist that none of the subjects opens the envelope prior to watching the entire movie. This approach ensures that the test subjects are not biased when they start watching the movie. Subjects were asked to answer different questions which enable us to track the number of impairments/degradations perceived, impairment annoyance and the overall quality rating for the movie. Amongst others, the following questions were included in the questionnaire:

1) Personalia (age, gender, video expertise level)
2) Did you perceive any visual artefacts during playback? If yes, how many? Which types: faltering image, blockiness, green blocks, other?
3) Describe the scenes or the locations where the degradations occurred.(e.g. the scene with the fire place at the beginning of the movie, when they are talking in close-up at the end of the movie).
4) Indicate on a scale from 1 to 5 the annoyance of the impairment (1 = not annoying at all, 5 = very annoying).
5) Which types of impairments are the least disturbing (choice between faltering image, blockiness, green blocks, other)?.
6) Which types of impairments are the most annoying? (faltering image, blockiness, green blocks, other)?
7) On a scale from 1 to 5 (1= very bad, 5= very good), rate the overall visual quality of the movie.

For the overall quality rating, a 5-grade Absolute Category Rating (ACR) scale was used as specified in ITU-T Rec. P.910.

As such, by using DVDs and providing a questionnaire to the subjects, we are able to meet all requirements for our new subjective video quality assessment methodology.

III. Subjective test setup for comparing real-life QoE assessment with a standard methodology

Using our novel proposed subjective video quality assessment methodology, described in the previous section, we conducted two subjective experiments.

In [14], we used full length DVD movies to assess the influence of packet loss and frame freezes on the perceptual quality of end-users. This subjective test was primarily conducted to investigate whether visual impairments are less noticed during real-life QoE assessment and to check which kind of impairments interrupt the viewing experience the most. Seven different movies of different genres (action, adventure, comedy and drama) were used and impaired with frame freezes and random blockiness caused by packet loss. The length of the movies varied between 120 and 170 minutes. Only the video track of the movie was impaired, the audio track remained unchanged. The duration of the impairments varied between 320ms and 400ms. According to a study of Pastrana-Vidal *et al.* [15] concerning the effect of motion jerkiness on quality perception, a single frame freeze with a duration of more than 200ms is detected by 100% of the test subjects. This result was obtained by conducting a standard subjective test with video sequences of 10 seconds long. For our test, several different impaired versions of the movies were created which contained up to three visual impairments, equally spread over time. The impairments were inserted into scenes with varying temporal and spatial complexity. However, no impairments occurred in the first and last half hour of movie playback. After watching the movie, subjects were asked to provide immediate feedback concerning the visual quality by completing a questionnaire which was provided together with the movie. It was stressed that subjects did not look at the questionnaire prior to watching the movie. The same questions were asked as described in the previous section. A total number of 56 non-expert viewers, of which 32 males and 24 females, participated with this subjective experiment. Some of them watched more than one movie. All subjects were aged between 14 and 49 years.

A second subjective test was conducted using the same full length DVD movie methodology in [16]. During this test, the influence of H.264 Scalable Video Coding (SVC) on the perceptual quality was studied. In the case of SVC, a video sequence is encoded using one base layer and one or more enhancements layers. The base layer offers the sequence at minimal quality. By decoding the base layer and additional enhancements layers, better perceived quality can be achieved. As such, SVC encoded video can be streamed over heterogeneous networks to a wide variety of end devices without the need for transcoding. H.264/SVC supports temporal, spatial and Signal-to-Noise Ratio (SNR) scalabilty. Temporal scalability implies that the sequence can be played at different frame rates whereas spatial scalability refers to the possibility of playing the sequence at different spatial resolutions. SNR scalability can be used to increase the overall picture quality. As such, in case a video needs to be scaled down, different techniques can be used. This experiment was set up to get a better understanding of the preferred scalability solution in case of video downscaling. One full length movie of 130 minutes long was used which contained six degradations, each 8 seconds long, equally spread over time but not in the first and last 15 minutes of movie playback. Each impairment was created by temporal and SNR downscaling that particular part of the original movie. For this, the movie was first encoded with a base layer and a Medium Grain Scalability (MGS) enhancement layer. In order to make a fair comparison between temporal and SNR (quality) scaling, the temporal reduced bit rate was used as a target for the quality scaling as illustrated in Fig. 3.

The latter experiment was also carried out as part of a wider social-scientific qualitative user research, investigating practices, motivations and trends regarding the use of video-in-the-home in a broad definition. A sample of 38 households, accounting for a total number of 100 subjects, participated in this test. The participants were asked to watch the movie on their preferred device (TV or PC). They were not aware there were any errors included in the movie, they thought the content of the movie would be the subject of discussion. Each subject was also asked to complete the questionnaire immediately after watching the DVD in preparation of a face-to-face interview the next day. During the interview, the questionnaire was used as a starting point for the discussion

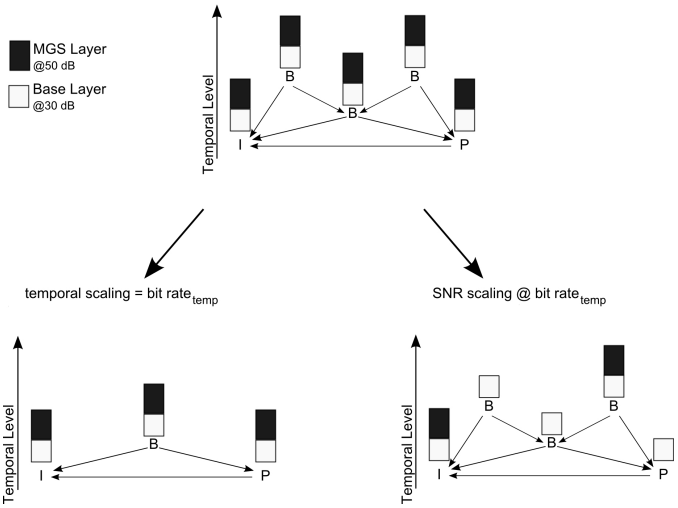


Fig. 3. Impairment creation by applying temporal downscaling, measuring the bit rate of the temporal reduced SVC sequence and targeting this bit rate for the SNR scaling.

and the alternative fragments were shown and discussed on the same device they watched the movie.

In this article, we now compare the results obtained during real-life QoE assessment with results obtained using a standardized methodology. Our goal is to highlight the difference in results and the importance of our novel subjective test methodology. Therefore, we conducted a third subjective test using the Single Stimulus ACR method as described in ITU-R Rec. BT.500-11. We opted for an SS test which implies that only the degraded videos were shown to the test subjects without the presence of an explicit reference which also corresponds more with watching television, where viewers can only evaluate the received video. The video sequences to be evaluated and rated during this experiment were taken from exactly the same impaired movies used in [14] and [16] and were created as follows. Using the movies from [14], we created 56 short video sequences with a duration of 15 seconds that contained either a frame freeze or random blockiness. These sequences corresponded to the scenes in the movies where the impairments occurred. For each impairment, a short sequence was created in which the visual degradation appeared in the first 5 seconds and a second sequence was created in which the same error appeared in the last 5 seconds of playback. As for the movies from [16], 18 sequences of 15 seconds long were created. For each of the six visual degradation caused by downscaling, three sequences were created which respectively contained the original sequence, the temporal downscaled version and the SNR downscaled version. As a result, each of the short video sequences to be shown to the subjects contained at most one visual impairment.

At the beginning of the subjective test, training sequences were used to indicate the type of impairments subjects could expect. Furthermore, specific instructions were given so that the subjects knew they had to evaluate the visual quality of the video sequences. After each sequence was displayed to the subjects, they were first asked to indicate which type of impairment they perceived (or none when they did not perceive any) and to rate the visual quality of the video using a 5-grade ACR scale. The video sequences were all displayed in full screen on a 19" flat panel TV display and the test subjects were seated at a viewing distance of 7 times the screen height. In total, 25 non-expert viewers participated with this subjective experiment.

In the following section, we present the results for this third subjective test and compare them with the results obtained using our full length movie subjective quality assessment methodology.

IV. Results

The results obtained during our research can be classified in two categories. First, we present the results which study the difference in impairment visibility. Then, we investigate the tolerance towards visual degradations. A comparison is always made between our novel methodology and the standard SS ACR test method.

A. Impairment visibility

Results from our previous research [14] showed that 42% of the subjects detected the frame freezes during movie playback and that 91% of them noticed the blockiness impairments. Blockiness impairments were created by injecting packet loss in that particular part of the movie. Packet loss rates were set high enough so that almost 50% of each individual image in the fragment contained the impairment. In the case of frame freezes, one single frame was frozen for the entire degradation period. By this, we wanted to create degradations which would be clearly visible during a standard test and check whether these would also be visible during real-life QoE assessment. It is clearly visible from the graph in Fig. 4 that blockiness is more often detected, even during movie playback. Blockiness impairments result in scrambled pictures whereas frame freezes leave the pictures intact which is less visible. Using the short video sequences in the standardized test, both frame freezes and blockiness were detected by 98% of the subjects, also shown in Fig. 4. While watching full length movies, subjects are concentrated on the content instead of on audiovisual quality. Before the start of the standard SS ACR test, subjects received specific instructions on the type of visual impairments they could perceive. As a consequence, subjects' primary focus was on visual quality evaluation. This shows that error visibility highly depends on whether users are focusing on content or quality and that impairments are less visible while watching television. This result also indicates that impairments can be masked using frame freezes as an efficient and simple error concealment technique when applied to full length movies. However, it is important to ensure that the natural flow of the movie is not interrupted. This will be discussed later.

Fig. 5a and 5b depict for each scene the percentage of the subjects that perceived the temporal or SNR downscaling during respectively our full length movie and our SS ACR test. As

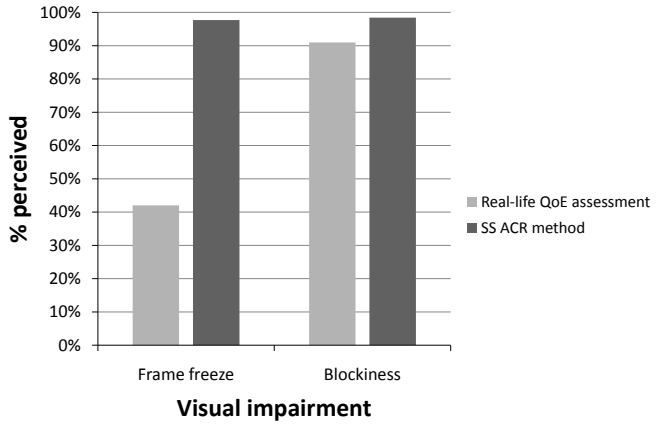


Fig. 4. Percentage of the subjects that perceived the frame freezes and blockiness during real-life QoE assessment and during our standard subjective test.

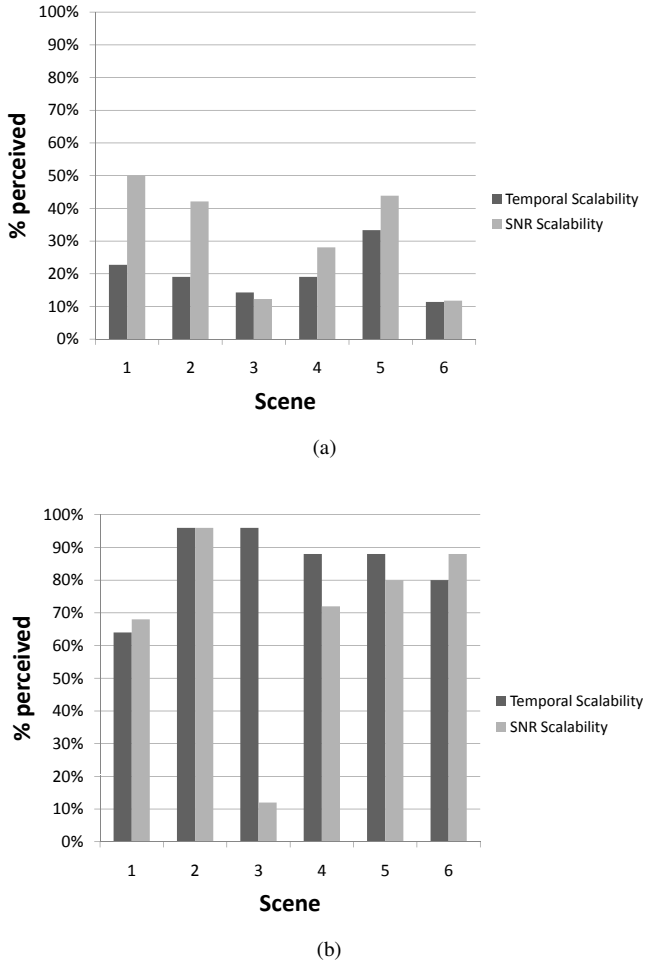


Fig. 5. Percentage of the subjects that noticed the temporal or SNR downscaling in each scene while watching the full length movie (a) and while evaluating the short video sequences during our SS ACR test (b).

can be seen, the detection threshold is content dependent and in general much more impairments were noticed during the standard test since subjects are, as explained earlier, focused on quality evaluation. In slight contrast with the full length movie, there is not much difference between the perceptibility of SNR and temporal scalability in the standard test, except for scene 3. This scene consisted of diagonal panning and contains a lot of spatial detail, making it a rather trivial case where SNR scalability would be preferred[2].

Leaving scene 3 aside for reasons described above, it should be noted that the SS ACR test does not yield big differences between the two scalability methods. From a service provider point of view these results are not of practical use in order to improve the QoE of their service. The results from the proposed full length movie quality assessment however show larger differences. Thus, they can be of more use for the network provider in order to target a certain type of scalability. Furthermore, the provider can also rest assured that this assessment methodology leans closer towards the real perceived quality of the consumption of their service.

B. Impairment tolerance

As explained in section III, our standard SS ACR test included 28 sequences which contained a visual impairment (either blockiness or a frame freeze) occurring in the first 5 seconds and 28 sequences with the same degradation in the last 5 seconds of playback. Hands *et al.* [17] studied the influence of primacy and recency effects on television picture quality assessment. Primacy and recency effects occur when the earliest, respectively the latest information has a higher impact during quality rating. Their study indicated the existence of a recency effect when using the Double-Stimulus Continuous Quality-Scale (DSCQS) method from ITU-R Rec. BT.500-11. This means that subjective quality ratings were lower for sequences which contained a visual degradation towards the end. We used a standard Student t-test for testing the significance between the Mean Opinion Scores (MOS) of our sequences which contained an impairment in the beginning and our sequences with an impairment towards the end. The calculated probability values are listed in Table I, significant values are placed in italic. In correspondence with the findings from [17], our results also indicate the occurrence of recency effects during the SS ACR test. However, the recency effect only holds in the case of blockiness impairments. The position of a frame freeze has no significant impact on the MOS of subjects.

TABLE I
CALCULATED P-VALUES FOR TESTING THE EXISTENCE OF RECENCY EFFECTS DURING A STANDARD SS ACR TEST AND REAL-LIFE QOE ASSESSMENT.

Test method	Frame freeze begin ↔ end	Blockiness begin ↔ end
SS ACR	0.0859	*0.0351*
Real-life QoE	0.4843	0.08876

Table I also indicates that no primacy nor recency effects occur during real-life QoE assessment [14]. This was calculated using the MOS scores of the subjects who perceived exactly one impairment during movie playback.

[2]This is because SNR will mostly remove spatial detail, which is typically less noticeable, while temporal scaling will be less fluent and hence more annoying during a panning.

As already stated, the impairments lasted between 320ms and 400ms. The relative portion of a sequence that is impaired depends on the total sequence duration. One visual degradation in a period of 120 minutes is perceived very differently compared to the same impairment in a sequence of 15 seconds long.

When subjects perceived both frame freezes and random blockiness impairments during the full length movie, they were asked to indicate which type of impairments they found the most and the least annoying. The results of this question showed that 83% of the subjects rated frame freezes more annoying than blockiness. Comparing the MOS for the short sequences which contained a visual impairment indicated that there is also a significant difference between the quality ratings for blockiness and frame freezes. In contradiction with the results from [14], subjects now rated frame freezes better than blockiness during the standard test. Sequences with frame freezes were rated on average 3.13 whereas blockiness was rated 2.52 on average with standard deviations of respectively 0.79 and 0.93. This difference in MOS is statistically significant to the 0.001 level. This means that, for the short video sequences, frame freezes are less disturbing to the subjects compared to random blockiness.

During the face-to-face interviews in [16], subjects indicated that the natural flow of the movie is very important and that they do not like impairments that break down the fluidity of the movie playback. Although frame freezes are less often noticed, when perceived they are considered more annoying. Blockiness impairments do not introduce jerky motion and therefore do not break playback fluidity. The importance of maintaining the natural flow of the movie also shows that special attention must be given to the use of frame freezes as concealment strategy. These results also indicate that a real TV viewing experience can be created with our novel methodology which is not the case during a standard subjective test. The social-scientific research clearly showed that when people watch TV in their natural setting on their preferred screen, not only the environment is important (the actual setting: screen size, distance to screen, the comfort of their home), but also the experience of the content that is watched. When people relate to the content and are really engaged in the flow of the movie, they might have a different idea about the disturbance of errors (e.g. in a face-to-face conversation blocks will be less disturbing than a disruption of the fluidity of the movie). The equipment people use to watch TV does influence the expected quality of the content. People who watch a movie on a PC are less error prone than people who watch it for example on their HDTV. This can be explained by the fact that people are used to see video content of lesser quality on the computer (e.g. streamed content, user generated content) than on TV and the people who bought themselves a HDTV screen have higher quality demands, as they expect the quality to be better than their old CRT screen.

After watching the full length movies from [14], subjects were required to provide a quality rating using a 5-grade ACR scale. Fig. 6 shows the average MOS scores together with the standard deviations of subjects who perceived none, one, two or three visual impairments during playback. Here we also used a Student t-test to see whether there is a significant difference between the MOS scores as a function of the number of perceived visual errors. The results listed in Table II, with significant values placed in italic, show that there is no statistically significant difference in MOS when subjects detect up to two visual impairments during movie playback. Both the graph and the table also clearly show that 3 visual impairments are no longer tolerated by the subjects, resulting in a significant drop of MOS to 3.13 (quality rated as 'fair'). Furthermore, these overall quality ratings are higher compared to the MOS scores given to the short sequences evaluated during our SS ACR test, which only contained one visual degradation.

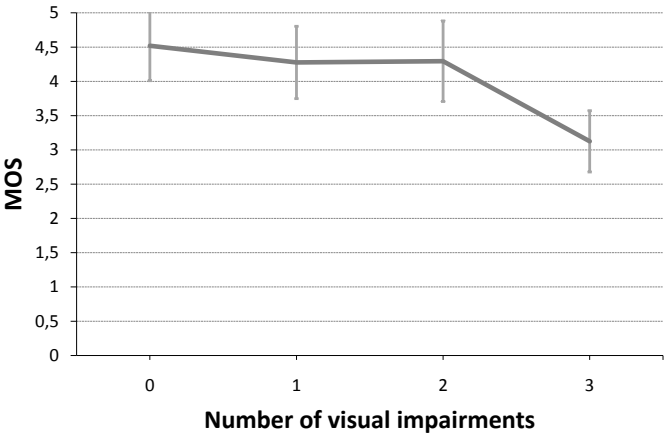


Fig. 6. Mean Opinion Scores, with standard deviations, as a function of the number of perceived visual impairments during real-life QoE assessment [14].

End-to-end recommended minimum QoE requirements for triple-play services, as defined by the DSL Forum [12], specify that in the case of Standard Definition (SD) video one visible impairment per hour of video playback is tolerable to maintain a satisfactory service level quality. For HDTV, only one visible impairment per four hours is allowed in order to provide a satisfactory QoE towards the end users. Our results show that users tolerate up to two visual impairments during movie playback, which corresponds with the recommended QoE requirements for SD. Issa *et al.* [18] showed that in the case of HDTV, the recommended mean time between visible distortions can be relaxed to some extent.

TABLE II
CALCULATED P-VALUES FOR TESTING WHETHER THERE IS A SIGNIFICANT INFLUENCE OF THE NUMBER OF PERCEIVED VISUAL IMPAIRMENTS ON MOS SCORES (DURING REAL-LIFE QOE ASSESSMENT).

Impairment count comparison	**p-value**
$0 \leftrightarrow 1$	0.0903
$0 \leftrightarrow 2$	0.2070
$1 \leftrightarrow 2$	0.9166
$0 \leftrightarrow 3$	*1.138e-07*
$1 \leftrightarrow 3$	*1.535e-06*
$2 \leftrightarrow 3$	*3.826e-06*

As a final result, we tested the significance between the MOS scores of SNR and temporal downscaled short sequences. This indicated that SNR scalability is rated better quality compared to temporal scalability for scenes 2, 3 and 4. For the other scenes, which contained less motion, there is no clear preference for a certain type of scalability. From the results in [16], it could be concluded that subjects in general favour temporal over quality scalability as it was less noticeable.

V. Conclusion

In this article, we presented a novel subjective video quality assessment methodology which enables real-life QoE assessment. Our methodology, based on full length movies, encourages subjects to watch and evaluate a video sequence in the same environment they usually watch television. By providing a questionnaire, feedback can be collected concerning the visual impairments and degradations, which were inserted in the movie.

Our results show some significant differences concerning impairment visibility and acceptability when conducting subjective experiments using one of the standardized methodologies and our novel proposed methodology. Using our methodology, we are able to mimic the typical lean-backward TV viewing experience which cannot be created using a standardized methodology. By doing so, test subjects are mainly focused on movie content instead of actively evaluating visual quality.

As a result, we showed that exactly the same visual impairments are perceived very differently when our full length movie methodology is used. In general, impairments are less visible during real-life QoE assessment. This is especially true in the case of frame freezes. Furthermore, we highlighted the importance of the flow experience of a movie which, in turn, influences users' acceptability of visual degradations during playback and users' preferred scalability solution. In the case of video downscaling, subjects always prefer that solution which does not break the natural flow of the movie.

Our work shows that results obtained using one of the standardized subjective quality assessment methodologies do not always hold on the case of real-life QoE assessment and that user expectations and context indeed influence end-users QoE. Visual quality metrics which are constructed using subjective data from short video sequences should therefore be reconsidered when these are used for measuring and monitoring end-users QoE of IPTV and VoD services.

Acknowledgment

The research activities that have been described in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT) and the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT). This paper is the result of research carried out as part of the Video Q-SAC project funded by the IBBT. Video Q-SAC is being carried out by a consortium of the industrial partners: Alcatel-Lucent, Telindus, Televic and fifthplay (Niko Group) in cooperation with the IBBT research groups: IBCN & MultimediaLab (UGent), SMIT (VUB) and IMEC.

The authors would also like to thank dr. ir. Koen Casier from Ghent University - IBBT and Philip Corriveau, Principal Engineer and Director of the User Experience Research Group at Intel Corporation for their invaluable contributions to this work.

References

[1] K. Yamori and Y. Tanaka, "Relation between willingness to pay and guaranteed minimum bandwidth in multiple-priority service," in *The 2004 Joint Conference of the 10th Asia-Pacific Conference on Communications, 2004 and the 5th International Symposium on Multi-Dimensional Mobile Communications Proceedings.*, vol. 1, 2004, pp. 113–117.

[2] J. Asghar, F. Le Faucheur, and I. Hood, "Preserving video quality in IPTV networks," *IEEE Transactions on Broadcasting*, vol. 55, no. 2, pp. 386 –395, june 2009.

[3] A. R. Reibman, S. Kanumuri, V. Vaishampayan, and P. Cosman, "Visibility of individual packet losses in MPEG-2 video," in *Proceedings of the International Conference on Image Processing*, vol. 1, October 2004, pp. 171–174.

[4] A. R. Reibman and D. Poole, "Predicting packet-loss visibility using scene characteristics," in *Packet Video 2007*, November 2007, pp. 308 –317.

[5] S. Kanumuri, P. Cosman, A. Reibman, and V. Vaishampayan, "Modeling packet-loss visibility in MPEG-2 video," *IEEE Transactions on Multimedia*, vol. 8, no. 2, pp. 341–355, April 2006.

[6] S. Winkler, "Quality of Experience (QoE) - an important measure of success for IP-based video services," in *SCTE 2008 Conference on Emerging Technologies*, January 2008.

[7] M. Buchenau and J. F. Suri, "Experience prototyping," in *Proceedings of the 3rd conference on Designing interactive systems*, 2000, pp. 424–433.

[8] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," T. Ebrahimi and T. Sikora, Eds., vol. 5150, no. 1. SPIE, 2003, pp. 573–582.

[9] P. Corriveau, *Video Quality Testing*, ser. Digital Video Image Quality and Perceptual Coding. CRC Press, 2006, ch. 4, pp. 125–153.

[10] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 660–668, September 2008.

[11] W. Van den Broeck, J. Pierson, and B. Lievens, "Confronting video-on-demand with television viewing practices," in *Innovating for and by users*, J. Pierson, E. Mante-Meijer, E. Loos, and B. Sapio, Eds. Opoce, 2008.

[12] T. Rahrer, R. Fiandra, and S. Wright, "TR-126: Triple-Play Services Quality of Experience (QoE) Requirements," DSL Forum, Tech. Rep., December 2006.

[13] A. Rombaut, N. Staelens, N. Vercammen, B. Vermeulen, and P. Demeester, "xStreamer: Modular Multimedia Streaming," in *Proceedings of the seventeenth ACM international conference on Multimedia*, 2009, pp. 929–930.

[14] N. Staelens, B. Vermeulen, S. Moens, J.-F. Macq, P. Lambert, R. Van de Walle, and P. Demeester, "Assessing the influence of packet loss and frame freezes on the perceptual quality of full length movies," *Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM-09)*, January 2009.

[15] R. Pastrana-Vidal, J. Gicquel, C. Colomes, and H. Cherifi, "Sporadic frame dropping impact on quality perception," *Human Vision and Electronic Imaging IX*, vol. 5292, 2004.

[16] N. Staelens, S. Moens, W. Van den Broeck, I. Mariën, B. Vermeulen, P. Lambert, R. Van de Walle, and P. Demeester, "Assessing the perceptual influence of H.264/SVC signal-to-noise ratio and temporal scalability on full length movies," *First International Workshop on Quality of Multimedia Experience (QoMEX 2009)*, July 2009.

[17] D. Hands and S. Avons, "Recency and duration neglect in subjective assessment of television picture quality," *Applied Cognitive Psychology:15*, pp. 639–657, 2001.

[18] O. Issa, W. Li, H. Liu, F. Speranza, and R. Renaud, "Quality assessment of high definition tv distribution over ip networks," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting,*, 2009.

Nicolas Staelens obtained his Master's degree in Computer Science at Ghent University (Belgium, 2004). He started his career in 2004 as a R&D engineer at Televic (a Belgian company that develops, designs and manufactures high-end network systems and software applications for the healthcare market). In 2006, he joined the IBCN-group (INTEC Broadband Communications Network Research Group) at Ghent University as a PhD student. His research focuses on studying the effects of network impairments on the perceived quality of audiovisual sequences. As of 2007, he is also actively participating within the Video Quality Experts Group (VQEG).

Stefaan Moens is a researcher at the Multimedia Lab of Ghent University - IBBT. He obtained his Master's degree in computer science from Ghent University in 2004 and his Master's degree civil engineer in computer science from Ghent University in 2007. His research is mostly related to video coding and efficiency.

Wendy Van den Broeck is researcher at SMIT (Studies on Media, Information and Telecommunication) - part of IBBT (Interdisciplinary institute for BroadBand Technology) and located at the Vrije Universiteit Brussel. She is preparing her PhD on user aspects of interactive digital television. Her main expertise and research interest is in the domestication and QoE of new media technologies in the home context, with a focus on TV-related services.

Ilse Mariën is junior researcher at SMIT (Studies on Media, Information and Telecommunication) - part of IBBT (Interdisciplinary institute for BroadBand Technology). Her research focuses on policy and user aspects of new media. As of 2008 she is preparing a Phd on policy aspects related to the digital divide.

Brecht Vermeulen received his Electronic Engineering degree in 1999 from Ghent University, Belgium. In June 2004, he received his PhD degree for the work entitled 'Management architecture to support quality of service in the internet based on IntServ and DiffServ domains.' at the department of Information Technology of Ghent University. Since June 2004, he is leading a research team within the IBCN group of Prof. Demeester which investigates network and server performance and quality of experience in the fields of video, audio/voice and multiple play. Since the start of IBBT (Interdisciplinary institute for BroadBand Technology) in 2004, he leads also the IBBT Technical Test Centre in Ghent, Belgium.

Peter Lambert is a Technology Developer at the Multimedia Lab of Ghent University (Belgium). He received his Master's degree in science (mathematics) and in applied informatics from Ghent University in 2001 and 2002, respectively, and he obtained the Ph.D. degree in computer science in 2007 at the same university. His research interests include multimedia applications, (scalable) video coding technologies, multimedia content adaptation, and error robustness of digital video.

Rik Van de Walle received his M.Sc. and PhD degrees in Engineering from Ghent University, Belgium in 1994 and 1998 respectively. After a visiting scholarship at the University of Arizona (Tucson, USA), he returned to Ghent University, where he became professor of multimedia systems and applications, and head of the Multimedia Lab. His current research interests include multimedia content delivery, presentation and archiving, coding and description of multimedia data, content adaptation, and interactive (mobile) multimedia applications.

Piet Demeester received the Master's degree in Electro-technical engineering and the Ph.D degree from the Ghent University, Gent, Belgium in 1984 and 1988, respectively. He is a full-time professor at Ghent University where he teaches courses in communication networks. He is the head of the Broadband Communication Networks group. His research interests include: multi-layer IP-optical networks, mobile networks, end-to-end quality of service, grid computing, network and service management, distributed software and multimedia applications. He has published over 500 papers in these areas in international journals and conference proceedings. In this research domain he was and is a member of several program committees of international conferences, such as: OFC, ECOC, ICC, Globecom, Infocom and DRCN. He is a fellow of the IEEE.