A Subjective Evaluation using Crowdsourcing of Adaptive Media Playout utilizing Audio-Visual Content Features

Benjamin Rainer
Alpen-Adria-Universität Klagenfurt
Multimedia Communication (MMC) Research Group
Institute of Information Technology (ITEC)
Klagenfurt, Austria
Email: benjamin.rainer@itec.aau.at

Christian Timmerer
Alpen-Adria-Universität Klagenfurt
Multimedia Communication (MMC) Research Group
Institute of Information Technology (ITEC)
Klagenfurt, Austria
Email: christian.timmerer@itec.aau.at

***Abstract*—Inter-Destination Multimedia Synchronization (IDMS) pushes social interactions to a new level. IDMS allows the users to experience multimedia together with friends, colleagues, or the family while having a real-time communication at the same time. The actual challenge of synchronizing the playout of each participant to a reference playout time is a tough task in terms of Quality of Experience (QoE). A possible solution for carrying out the synchronization is Adaptive Media Playout (AMP) where the playout speed of the multimedia is increased or decreased. In this paper we evaluate the impact of the playout variations on the QoE by adopting a crowdsourcing approach. In particular, we investigate the impact of randomly selecting content sections for adapting the playout rate compared to our approach that exploits audio-visual features of the content in order to minimize the impact on the QoE.**

I. Introduction

Social networks have become pervasive and have found their way into our daily life. Many of our social activities are nowadays handled via social platforms like Facebook, Twitter, and Google+. In this context, watching TV together while being geographically distributed enriched with the possibility of having a real-time communication with each other (e.g., via text, voice, or even video) has become a new social event.

One major challenge to allow this type of social interaction is the synchronization of the multimedia playout among the geographically distributed clients. In the literature, this type of synchronization is referred to as Inter-Destination Multimedia Synchronization (IDMS) [1]. Existing IDMS schemes mainly deal with the signaling of timing information such that the media playout of geographically distributed clients can be synchronized by utilizing the signaled timing information. Additionally, the assumption that there is a real-time communication channel between the users is a key aspect of IDMS and, thus, puts an upper bound on the asynchronism between the clients. Not only the signaling of timing information and control information will have to fulfill the new demands but we also have to take a look at how the synchronization is actually carried out at the clients.

In our research we focus on how the synchronization is carried out at each participating user in an IDMS system. Therefore, we suggest the use of Adaptive Media Playout (AMP). Previously, AMP was designed to control the buffer fill state and, in particular, to avoid buffer underflows and buffer overflows by increasing or decreasing the media playout rate (e.g., [2], [3]). A simple way of achieving synchronization among geographically distributed clients is to pause or skip audio/video frames but with an increased amount of pauses the Quality of Experience (QoE) degrades exponentially [4].

In this paper, we focus on how the synchronization is carried out at each client by employing Adaptive Media Playout (AMP) and its impact on the Quality of Experience (QoE). Therefore, we subjectively evaluate the impact of two AMP algorithms on the QoE adopting a crowdsourcing approach. The first AMP algorithm randomly chooses content sections for increasing and decreasing the playout rate. The second is referred to as the QoE- and Context-Aware Adaptive Media Playout (QoECAMP) algorithm presented in [5] which exploits audio-visual features for identifying appropriate content sections where the playout rate may be increased or decreased.

The remainder of this paper is organized as follows. Section II provides an overview of the related work in IDMS, AMP, and subjective quality assessments using crowdsourcing. The two AMP algorithms that are compared in this paper are briefly described in Section III. Section IV introduces the methodology of the subjective quality assessment including stimuli, subjects, and the selected crowdsourcing platform. Furthermore, we briefly discuss the acquisition of the data and which measures are collected for conducting a posterior cheat detection. The results of the subjective quality assessment are provided in Section V. In Section VI the results are discussed and an outlook on future work is given.

II. Related Work

IDMS can be achieved in various ways. The common assumption of most of IDMS solutions is that clocks are already synchronized by any existing clock synchronization protocol (e.g., NTP). Thus, most of the schemes only deal with the signaling of timing information and control information to achieve IDMS among the participating clients [6], [1].

The synchronization thresholds for IDMS under different communication possibilities between users were investigated in

978-1-4799-0913-1/14/$31.00

[7]. The results stated that depending on the type of real-time communication (eg., video chat with audio or text chat) the upper bound on the asynchronism varies. For example, when users are provided with a voice communication tool, asynchronism is only subjectively perceived above two seconds.

In addition to the selection of the reference and the type of the control scheme, there is ongoing work on how the synchronization should be carried out at each client. Currently, the common denominator of the mentioned schemes and solutions is that compensating the identified asynchronism is done by skipping or pausing media units. In [4] the effect of stalls during media playout was subjectively assessed. The results indicate that the Mean Opinion Score (MOS) degrades with an increase in stalls during media playout. As a result, using skips and pauses to overcome asynchronism may lead to a low QoE for the users. To overcome these shortcomings, AMP was introduced and the approach described in [8] deals with simply increasing or decreasing the playout rate in order to avoid buffer underflows or overflows without considering the influence on the QoE of the user.

Initially, AMP was thought to be used to compensate for buffer underflows or overflows by decreasing or increasing the media playout rate to allow the stabilization of the playout buffer. The authors of [2] modeled the adaptation of the media playout rate depending on the buffer variance. In [3] the buffer fill state was used to decide whether the playout rate should be increased or decreased. The authors of [9] use the motion intensity of video scenes in order to decrease or increase the playout rate for the whole scene but the authors do not consider the impact of these playout rate variations on the QoE. All these schemes model the underlying error prone channel by a series of random variables that follow the Markov property.

Using crowdsourcing for subjective quality assessments has received a lot of attention in the past years with the uprise of crowdsourcing platforms like Mechanical Turk [10] and Microworkers [11]. In [12] a Web-based platform for subjective quality assessments using crowdsourcing is presented. This platform supports the pair comparison method and provides a very intuitive user interface where the participant uses the space bar for a binary continuous rating.

In [13] another crowdsourcing framework which is called QualityCrowd is presented that directly interacts with Mechanical Turk and allows the definition of various test methodologies. Additionally, they discuss various challenges that arise when subjective quality assessments are conducted by the use of crowdsourcing platforms, i.e., conceptual, technical, motivational, and reliability challenges.

In [14] cheat detection mechanism for subjective quality assessments using the pair comparison evaluation method are proposed. These cheat detection mechanism are evaluated on top of a conducted subjective quality assessment. The authors present how cheat detection can be achieved for pair comparison. The proposed methods are not applicable to all evaluation methods. Thus, for our subjective quality assessment we had to come up with a cheat detection that fits the evaluation methodology we used.

A very interesting compilation of best practices for crowdsourcing subjective quality assessments is presented in [15]. In particular, different outlier screening methods are compared (including the one presented in [14]). The results clearly state that, detecting outliers can not only relied on taking user ratings into account. Thus, additional mechanism should be introduced in order to detect outliers.

III. Adaptive Media Playout Algorithms

In this section we briefly describe the algorithms used in our subjective quality evaluation. The algorithms are used to select the content sections for which the playout rate is increased or decreased.

A. Random Selection of Content Sections

AMP algorithms discussed in [8], [2], [3] decrease or increase the playout rate according to the buffer fill state. The fill state of the buffer is influenced by an error prone channel which introduces transmission errors according to a series of random variables X_k that follow the Markov property. In order to mimic these AMP algorithms we determine the point in time for increasing or decreasing the playout rate randomly. This allows us to simulate transmission errors on which the above algorithms would react by increasing or decreasing the playout rate. The duration of these randomly determined content sections have the same duration as the content sections determined by our QoECAMP algorithm. This allows us to compare the random selection of content sections to our QoECAMP algorithm in terms of QoE.

B. QoE- and Context-aware Adaptive Media Playout

The second algorithm tries to postpone the increase or decrease of the playout rate until a suitable content section is identified that may reduce the impact of increasing or decreasing the playout rate on the QoE. Therefore, audio-visual features of the current buffer contents are used to determine these content section.

For the video feature we select the average length of motion vectors of each frame n depicted as $f_v(n)$. The set of motion vectors is denoted as V_n, for $1 \leq n \leq N$, $n \in \mathbb{N}$ and N represents the maximum number of frames. For each frame $f_v(n)$ is calculated as defined in Equation 1.

$$f_v(n) = \frac{\sum_{i=1}^{N} \|\mathbf{v}_i\|_2}{N} \tag{1}$$

where $\mathbf{v}_i \in V_n$ and $\|\mathbf{v}_i\|_2$ is the L_2-norm of $\mathbf{v}_i$.

For the audio feature we select the Root Mean Square (RMS) of the envelope of each audio frame n depicted as $f_a(n)$ (cf. Equation 2). We use a resolution of signed 16-Bit for each audio sample (a_i) and a sampling rate of 44.1 kHz. Furthermore, we use a hamming window of 1024 samples which corresponds to an audio frame and a half overlapping window, thus, resulting into 512 overlapped samples per window.

$$f_a(n) = \sqrt{\frac{\sum_{i=1}^{|A_n|} a_i^2}{|A_n|}} \tag{2}$$

where $a_i \in A_n$ and A_n is the set of audio samples for an audio frame with $|A_n|$ as the cardinality of A_n.

These features are measured over time and their mean and standard deviation are used to approximate their future

behavior. Therefore, the mean of each feature $f_i(n)$ within a time window is calculated (expressed as frames) by the use of an discrete moving average filter (or low-pass filter) with a windows size in frames given by ω depicted by Equation 3.

$$M_{i_\omega}(n) = \begin{cases} \frac{1}{\omega+1} \sum_{j=n-\frac{\omega}{2}}^{n+\frac{\omega}{2}} f_i(j), & n \geq \frac{\omega}{2} \\ \frac{1}{n+1} \sum_{j=0}^{n} f_i(j), & n < \frac{\omega}{2} \end{cases} \quad (3)$$

These averages are normalized by the current maximum of feature i which is depicted by $\widehat{M}_{i_\omega}(n)$. $\overline{\widehat{M}}_{i_\omega}(n)$ depicts the average of the low-pass filtered f_i (cf. Equation 4) for a given window size κ. During media playout we do not know the overall maximum of a given feature because not all media units may be present at the client. Thus, we cannot avoid using local maxima for normalizing M_{i_ω}. Nevertheless, finding a new maximum does not invalidate previous decisions and calculations.

$$\overline{\widehat{M}}_{i_{\kappa,\omega}}(n) = \begin{cases} \frac{1}{\kappa} \sum_{j=1}^{\kappa} \widehat{M}_{i_\omega}(n-\kappa+j), & \kappa \leq n \\ \frac{1}{n} \sum_{j=1}^{n} \widehat{M}_{i_\omega}(j), & \kappa > n \end{cases} \quad (4)$$

κ depicts the window size in frames that is used to take the *past* of $\widehat{M}_{i_\omega}(n)$ starting at frame n into account. Past values of $\widehat{M}_{i_\omega}$ reflect how feature f_i changed on average over the specified time window in frames. The actual value of the features is compared to the mean minus the corresponding standard deviation for feature f_i. If the actual value of the features is below this threshold, the frame is selected for increasing or decreasing the playout rate.

For calculating the lower threshold $l_i(n)$ and upper threshold $u_i(n)$ for feature f_i we take the empirical standard deviation (which is an unbiased estimator for the variance) $s_{i_{\kappa,\omega}}$ of the normalized M_{i_ω} within a parametrized window κ depicted by Equation 5.

$$s^2_{i_{\kappa,\omega}}(n) = \begin{cases} \frac{\sum_{j=1}^{m} (\widehat{M}_{i_\omega}(n-\kappa+j) - \overline{\widehat{M}}_{i_{\kappa,\omega}}(n))^2}{\kappa-1}, & \kappa \leq n \\ \frac{\sum_{j=1}^{n} (\widehat{M}_{i_\omega}(j) - \overline{\widehat{M}}_{i_{\kappa,\omega}}(n))^2}{n-1}, & \kappa > n \end{cases} \quad (5)$$

The lower $l_i(n)$ and upper $u_i(n)$ thresholds are calculated as depicted by Equation 6 and Equation 7, respectively.

$$l_i(n) = \overline{\widehat{M}}_{i_{\kappa,\omega}}(n) - \delta * \sqrt{s_{i_{\kappa,\omega}}(n)^2} \quad (6)$$

$$u_i(n) = \overline{\widehat{M}}_{i_{\kappa,\omega}}(n) + \delta * \sqrt{s_{i_{\kappa,\omega}}(n)^2} \quad (7)$$

For identifying the content sections for the stimuli of the conducted subjective quality assessment we used following values for the parameters of the proposed algorithm: $\omega = 100$, $\kappa = 125$ and $\delta = 1$. For further information on the algorithm we refer the interested reader to [5].

IV. Subjective Quality Assessment Methodology

This section describes the methodology used for the subjective quality assessment using crowdsourcing.

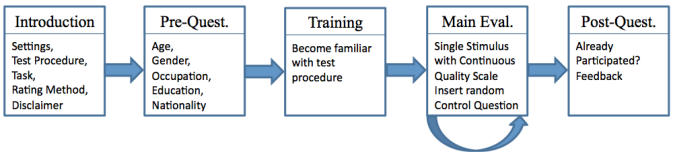

Fig. 1. Evaluation Methodology.

A. Stimuli and Stimulus Presentation

For our subjective quality assessment using crowdsourcing we select excerpts from the Big Buck Bunny and Sintel sequences with the absolute start time and end time of the actual sequence given in brackets (mm:ss) followed by the total length: *i*) Big Buck Bunny (01:10-02:00, 50s); *ii*) Sintel1 (01:30-02:54, 84s); *iii*) Sintel2 (02:54-03:52, 58s).

Please note that Big Buck Bunny was only presented during the training phase of the experiment. All sequences have a resolution of 720p, 25 fps, and a bitrate of about 2.5 Mbit/s. Our aim is to subjectively assess the impact of the selected QoECAMP algorithm and randomly selecting content sections on the QoE. Therefore, we select two excerpts of the Sintel sequence such that the first (Sintel1) contains a fair amount of natural speech (i.e., dialogs) and the second (Sintel2) contains nearly no natural speech.

For Sintel1 the total duration of the content sections where the playout rate is adjusted is 10.08 seconds (duration on average of a single section: 0.72 seconds, with a standard deviation of 0.76 seconds, maximum duration among all section is 2.6 seconds) which is 12% of the total length. The playout rate adjustments for Sintel2 have a duration of 7.84 seconds (duration on average of a single section: 1 second, with a standard deviation of 1.06 sec, maximum duration among all section is 2.6 seconds) representing 13.52% of the total length.

For the playout rate adjustment we choose the following values for μ: 0.5, 0.75, 1.5, and 2 times the nominal playout rate, i.e., $\mu = 1$. We selected these playout rates in order to assess whether a change of 25% of the nominal playout has no significant impact on the QoE as stated in [8], [2] and [9]. Furthermore, we are interested in how the QoE is influenced when the playout rate is even higher or lower than the claimed 25%. We present each video with each of the algorithms (Random and QoECAMP) where the content sections selected by the algorithm are played with each of the given playout rates. Therefore, each algorithm is presented nine times including the reference for $\mu = 1$ for each sequence. Thus, in total we have 18 test conditions.

B. Crowdsourcing Platform and Participants

As already mentioned we use Microworkers as crowdsourcing platform as it allows hiring workers outside the USA. The advantages of using a crowdsourcing platform is the instant access to a huge number of participants and the low effort of actually conducting the experiment. We found that the compensation for a task which requires approximately 20 minutes is on average about 0.7 cent (Euro) at the Microworkers platform.

C. Evaluation Methodology

For conducting the experiment we used workers from Europe and the USA which results in a higher reliability of the

responses [16]. The Microworkers platform allows to restrict the origins of workers to specific countries. The assessment was structured as depicted in Figure 1.

Introduction. At the beginning, a short introduction is presented to the participants which explained in detail what the participants have to do and what they will have to assess. In particular, the participants' task is to evaluate the perceived quality of the viewing/hearing experience while watching the sequences. Furthermore, we explain the whole assessment such that no questions are left open. This includes a detailed explanation of what will happen during the experiment, how the rating scale and rating possibility will look like, and the different phases of the experiment. Additionally, we asked the participants to turn off mobile devices, darken the room, and set up their audio devices to a pleasant configuration. Furthermore, the introduction included a disclaimer that persons who are visually impaired or have impairments regarding hearing should not take part in the subjective quality assessment.

Pre-Questionnaire. After the introduction, a pre-questionnaire is shown to gather demographical information about the participants, i.e., age, gender, country of residence, nationality, occupational field, and education. This will provide us with demographical data that can be used to identify influence factors for groups of participants clustered according to one of the demographic variables.

Training. The training phase using the Big Buck Bunny sequence is presented to allow the participants to adjust their audio volume and to become familiar with the stimulus presentation. Furthermore, it allows participants to become familiar with the rating scale. The Big Buck Bunny sequence is presented in three different configurations. The first configuration comprises the training sequence with the nominal playout rate of $\mu = 1$ and, thus, without any temporal impairments. For the other two configurations we modified the playout rate to $\mu = 2$ (i.e., twice the nominal playout rate) and $\mu = 0.5$ (i.e., half the nominal playout rate) for selected content sections. We selected the Big Buck Bunny sequence as training sequence because it does not convey any natural speech.

Main Evaluation. The main evaluation adopts a single stimulus with hidden reference as recommended by the ITU [17], [18]. The idea behind the selection of a single stimulus was that the participants should not know the reference condition (i.e., $\mu = 1$). They should only rate the actual sequence with or without temporal impairments like in a home TV viewing/hearing experience. The hidden reference should allow us to clarify whether there is a significant difference between the reference and the temporal impaired sequences. After each test condition the rating possibility is presented to the participants using a slider. We selected a continuous rating scale with an interval of $[0, 100]$ with 0 indicating a very low QoE and 100 representing a very high QoE. Furthermore, each rating phase was limited to eight seconds. Additionally, we use a control question (i.e., "What was present in the last video sequence?") with three possible answers using an option box to check whether the participants are paying attention. The control question is inserted randomly following one of the 18 test conditions.

Post-Questionnaire. Finally, at the end of the experiment the participants are asked to fill out a post-questionnaire. The post-questionnaire provides participants the opportunity to give feedback using a free text field regarding whether they participated already in a similar experiment. After the post-questionnaire a unique token is shown which is a mandatory proof that a micro-worker had successfully participated in our subjective quality assessment.

D. Filtering of Participants

We introduce a three-level scheme for filtering participants from the result set. As already mentioned in Section II we do not only rely on the ratings obtained by the rating possibilities. Therefore, we use the additional data that is gathered by our Web-based assessment platform [19]. That is the duration of the stimuli presentation and the duration of each rating process for each participant. We first describe the methods and afterwards we give the numbers of participants that have been screened by using the following methods.

The **first level** comprises the control question and we reject participants who did not provide a correct answer to the control question. A wrong answer on the control question may indicate that the participant did not pay attention to the sequences. Furthermore, it may indicate that a participant did not understand the question and therefore it is likely that the participant may not have understood the introduction and, consequently, the actual task. Thus, we reject participants that provided a wrong answer for the control question.

The **second level** is about screening participants who had a significant difference in playout time in comparison to the playout time for the nominal playout rate $\mu = 1$ for each stimulus presentation. Therefore, we used the F-test to test whether there exists a significant difference between the variances of the nominal playout times and the playout times of each participant. We rejected those participants for which the F-test stated a significant difference for a significance level of $\alpha = 0.05$.

The **third level** filters those participants with abnormal rating behavior. Therefore, we take a closer look at the ratings of each participant and found that some participants moved the slider only a few times out of n rating possibilities (in our case $n = 18$). In order to detect whether a participant just leaves the slider at the initial position or moves it to one of the extreme values of the presented rating scale (0 or 100), we model the selection of an extreme value by the use of a binomial distribution ($X \sim B(n, p)$) with $p = \frac{1}{2}$ the probability of selecting an extreme value. We rejected a participant if $\alpha \geq 1 - \mathbb{P}(X \leq k)$ with $\alpha = 0.05$.

In total 119 micro-worker participated in the subjective quality assessment. The filtering using the control question reveals that four participants did not provide a correct answer. The screening according to the playout time results in nine participants. Four of these nine participants did try to skip at least one stimulus presentation. Five out of the nine participants paused the playout of at least one stimulus presentation. For the ratings we rejected ten participants that did not provide viable ratings, i.e., either not moving the slider at all or selecting an extreme value very often. In total we screened 23 participants out of 119 by applying our filtering scheme. We further used the Median Absolute Deviation (MAD) to detect outliers according to the threshold of two times the standard

deviation for the QoE ratings [20]. The MAD did not reveal any outliers. Thus, we did the analysis of the results based on the responses we received from 96 participants which are presented in the following section.

V. Statistical Analysis of the Responses

A. Pre- and Post-Questionnaire

For the pre-questionnaire the participants provided us with the following data. The majority of the participants is between 20 and 25 years old with 85% of the participants 35 or younger and from the 96 participants are 20 female and 76 male. 11% stated that they are experts and working in the field of computer and mathematics. Furthermore, 26% of the participants stated that they are students.

For the post-questionnaire we got the following feedback. Approximately 80% of the participants stated that they have not participated in a similar experiment. The remaining 20% stated that they already participated in a similar subjective quality assessment hosted at Microworkers. Furthermore, approximately 4% stated that the subjective quality assessment was too long or that the number of sequences should be reduced.

B. Main Evaluation

The aim of the subjective quality assessment was to assess whether it makes a difference when selecting content sections by using a more sophisticated approach in contrast to randomly selecting the content sections for which the playout rate should be increased or decreased. Therefore, our focus is on identifying whether there exist significant differences between the two algorithms for the selected playout rates. We analyzed the responses according to significant differences between their means by using a Student's t-test. Prior to the Student's t-test we ensured that the variance between two tested samples are equal by conducting an F-test. For testing whether there is no normal distribution present for the ratings of each test condition we used the Lilliefors-test. If the analysis of the variances rejected the hypothesis that two variances are equal we used the Welch's t-test instead of the Student's t-test which assumes a normal distribution of the samples. According to the Lilliefors-test the hypothesis that the samples are not drawn from a normal distribution was rejected. Therefore, we use parametric statistical tests to assess the significance of our results.

Figure 2 depicts the Mean Opinion Score (MOS) and the 95% Confidence Interval (CI) for the first sequence Sintel1 for each test condition with the QoECAMP and Random algorithm. As already mentioned, Sintel1 contains a fair amount of natural speech (dialogs) with high audio volume. The x-axis shows the different playout rate adjustments and the hidden reference is depicted by $\mu = 1$, i.e., participants voted only once and, thus, the MOS for both QoECAMP and Random are equal.

At a first glance it can be observed that for playout rates close to the reference $\mu = 1$ the MOS does not change that much for both algorithms. This finding is supported by the results of a Student's t-test between the means of QoECAMP for $\mu = 0.75$ and $\mu = 1.5$ and the reference condition $\mu = 1$.

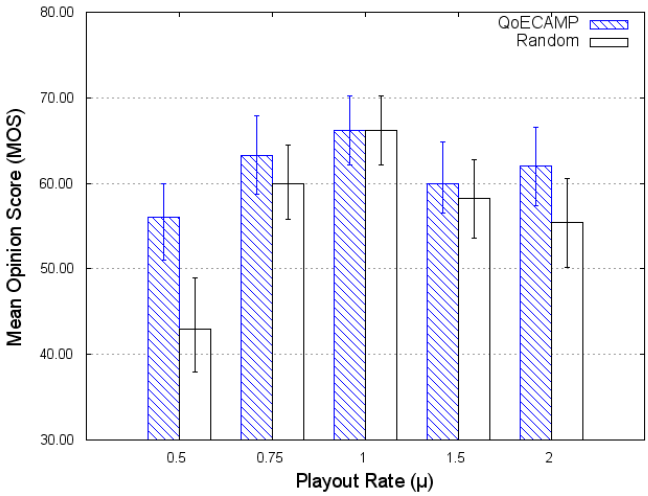

Fig. 2. MOS and 95% CI for the Sintel1 sequence.

For Random there is in fact a significant difference between the means for $\mu = 0.75$ and $\mu = 1.5$ and the reference with: $\mu = 0.75$, $p = 0.048$ and $t = 1.99$; $\mu = 1.5$, $p = 0.01$ and $t = 2.5387$. Taking a look at the test conditions where the playout rate was decreased to $\mu = 0.5$ and increased to $\mu = 2$ it can be observed that QoECAMP starts to outperform the Random algorithm. For $\mu = 0.5$ the difference of the means of both algorithms compared to the reference is statistically significant ($p = 4.3 * 10^{-10}$, $t = 6.587$ for Random and $p = 0.0011$ and $t = 3.321$ for QoECAMP). According to a Student's t-test the difference of the means between both algorithms for $\mu = 0.5$ is statistically significant too ($p = 0.0007$ and $t = 3.4$). These results state that QoECAMP performs significantly better in extreme situations where the playout rate is very low or very high.

For $\mu = 2$ the same behavior can be observed. The QoECAMP algorithm is able to maintain a QoE of above 70 MOS points. The Random algorithm scores below 65 MOS points. Thus, a Student's t-test revealed a significant difference for the mean of Random and the reference ($p = 0.0016$, $t = 3.198$). There is no significant difference between QoECAMP and the reference for $\mu = 2$. Another interesting finding is that decreasing the playout rate results into higher QoE degradations than increasing the playout rate by the reciprocal factor of the decrease.

Figure 3 illustrates the MOS and the 95% CI for the second sequence Sintel2 comprising nearly no natural speech and low audio volume. Interestingly, when increasing the playout rate ($\mu > 1$), the MOS remains almost the same for QoECAMP while it decreases for the Random case. Furthermore, the QoECAMP algorithm scores a higher MOS than the Random algorithm for all playout rate adjustments ($\mu \neq 0$). Again, it can be observed that playout rates close to the reference of $\mu = 1$ do not show a statistically significant difference, except for Random with $\mu = 0.75$ ($p = 0.0017$, $t = 3.18$). For $\mu = 0.5$ and $\mu = 2$ Figure 3 paints nearly the same picture as Figure 2. For the playout rate $\mu = 0.5$ the means of both algorithms are statistically significant different in comparison to the reference ($p = 8.9 * 10^{-7}$, $t = 5.1$ for QoECAMP; $p = 1.1 * 10^{-11}$, $t = 7.24$ for Random). For $\mu = 0.5$ there is a significant difference of the means of both algorithms with

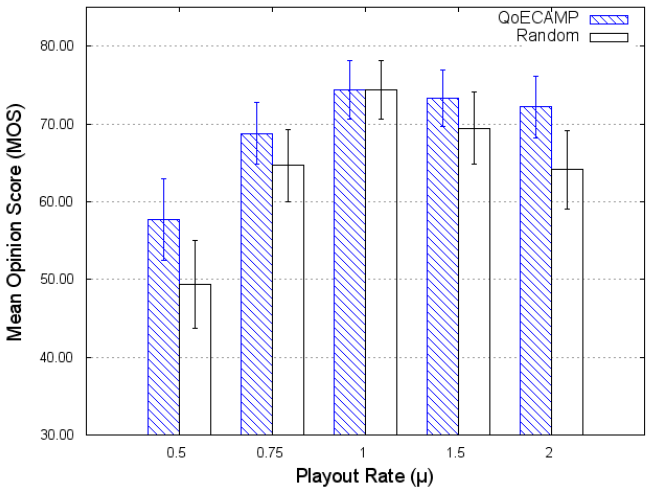

Fig. 3. MOS and 95% CI for the Sintel2 sequence.

$p = 0.03373$ and $t = 2.1388$. The results state that QoECAMP performs better than Random from a QoE point of view.

If we increase the playout rate the same behavior can be observed as with $\mu = 0.5$. QoECAMP is able to maintain a high QoE in comparison to Random. A Student t-test supports this finding by stating a significant difference between the means of Random for $\mu = 2$ and the reference ($p = 0.0016$, $t = 3.198$) but not for QoECAMP. Furthermore, a Student t-test revealed a statistically significant difference for both means of both algorithms for $\mu = 2$ ($p = 0.01476$, and $t = 2$). As before, we observe that decreasing the playout rate leads to a higher decrease in QoE than increasing the playout rate.

VI. Discussion and Conclusions

The results presented in Section V clearly show that the impact of increasing the playout rate on the QoE is lower than the impact of decreasing the playout rate. On the one hand, the findings contradict the results of informal tests mentioned in [8], [2] and [9], where it is stated that playout variations of 25% up to 50% of the nominal playout rate may not be perceptible by users. The results state that this does not hold, especially, when the decision of increasing or decreasing the playout rate is based on a random variable. The results of the study show that a more sophisticated selection of content sections for increasing or decreasing the playout rate allows decreasing or increasing the playout rate without significantly degrading the QoE. Furthermore, the results state that the selection of the content sections gets even more important the higher or lower the playout rate is. This is a very important finding regarding the synchronization of the media playback between clients in IDMS because it states that increasing the playout rate by 25% does not have a significant impact on the QoE. This allows us to overcome asynchronism without bothering whether we cause a significant impact on the QoE assuming that the buffer fill state is high enough.

Interestingly, increasing the playout rate does not have that huge impact on the QoE. Especially, for playout rates of about 25% of the nominal rate it seems that the selection of content sections does not matter. Thus, increasing the playout rate should always be preferred if possible. Another very important fact that can be observed when comparing Figure 2 and Figure 3 is that with more information present in the audio domain the QoE decreases. This indicates that audio plays a very important role when selecting the content sections for playout rate variations. Therefore, we declare this finding as subject to future work. In particular, it raises the question on the impact of the distortion in the audio domain on the resulting QoE.

Acknowledgments

This work was supported in part by the European Commission in the context of the SocialSensor (FP7-ICT-287975), QUALINET (COST IC 1003) projects and partly performed in the Lakeside Labs research cluster at AAU.

References

[1] M. Montagud, F. Boronat, H. Stokking, and R. Brandenburg, "Inter-Destination Multimedia Synchronization: schemes, use cases and standardization," in *Multimedia Systems*, vol. 18. IEEE, 2012, pp. 459–482.

[2] Y.-F. Su, Y.-H. Yang, M.-T. Lu, and H. H. Chen, "Smooth control of adaptive media playout for video streaming," in *Transactions on Multimedia*, vol. 11, no. 7. Piscataway, NJ, USA: IEEE, Nov. 2009, pp. 1331–1339.

[3] M. Kalman, E. Steinbach, and B. Girod, "Adaptive media playout for low-delay video streaming over error-prone channels," in *Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 6. IEEE, 2004, pp. 841–851.

[4] T. Hossfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via Crowdsourcing," in *International Symposium on Multimedia (ISM), IEEE*, 2011, pp. 494–499.

[5] B. Rainer and C. Timmerer, "Adaptive Media Playout for Inter-Destination Media Synchronization," in *Proceedings of the 5th International Workshop on Quality of Multimedia Experience (QoMEX'13)*. Los Alamitos, CA, USA: IEEE, jul 2013, pp. 44–45.

[6] F. Boronat, J. Lloret, and M. García, "Multimedia group and inter-stream synchronization techniques: A comparative study," in *Information Systems*, vol. 34, no. 1. Oxford, UK, UK: Elsevier Science Ltd., Mar. 2009, pp. 108–131.

[7] D. Geerts, I. Vaishnavi, R. Mekuria, O. van Deventer, and P. Cesar, "Are we in sync?: synchronization requirements for watching online video together." in *Proceedings of the SIGCHI*. ACM, 2011, pp. 311–314.

[8] M. Montagud and F. Boronat, "On the Use of Adaptive Media Playout for Inter-Destination Synchronization," in *Communications Letters*, vol. 15, no. 8. IEEE, 2011, pp. 863–865.

[9] Y. Li, A. Markopoulou, J. Apostolopoulos, and N. Bambos, "Content-Aware Playout and Packet Scheduling for Video Streaming Over Wireless Links," *Multimedia, IEEE Transactions on*, vol. 10, no. 5, pp. 885–895, 2008.

[10] Mechanical Turk, "https://www.mturk.com/mturk/."

[11] Microworkers, "http://www.microworkers.com."

[12] K.-T. Chen, C.-J. Chang, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "Quadrant of euphoria: a crowdsourcing platform for QoE assessment," *IEEE Network*, vol. 24, no. 2, pp. 28–35, 2010.

[13] C. Keimel, J. Habigt, and K. Diepold, "Challenges in crowd-based video quality assessment," in *2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2012, pp. 13–18.

[14] C.-C. Wu, K.-T. Chen, Y.-C. Chang, and C.-L. Lei, "Crowdsourcing Multimedia QoE Evaluation: A Trusted Framework," *Multimedia, IEEE Transactions on*, vol. 15, no. 5, pp. 1121–1137, 2013.

[15] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best Practices for QoE Crowdtesting: QoE Assessment with Crowdsourcing," *Multimedia, IEEE Transactions on*, vol. PP, no. 99, 2013.

[16] J. Redi, T. Hossfeld, P. Korshunov, F. Mazza, I. Povoa, and C. Keimel, "Crowdsourcing-based Multimedia Subjective Evaluations: A Case Study on Image Recognizability and Aesthetic Appeal," *CrowdMM*, 2013.

[17] ITU-R Recommendation BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Geneva, Switzerland, Tech. Rep., Jan. 2012.

[18] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union, Geneva, Switzerland, Tech. Rep., Apr. 2008.

[19] B. Rainer, M. Waltl, and C. Timmerer, "A Web based Subjective Evaluation Platform," in *Proceedings of the 5th International Workshop on Quality of Multimedia Experience (QoMEX'13)*. Los Alamitos, CA, USA: IEEE, jul 2013, pp. 24–25.

[20] V. Barnett and T. Lewis, *Outliers in statistical data*, ser. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley & Sons, 1994.