A Crowdsourcable QoE Evaluation Framework for Multimedia Content

Kuan-Ta Chen\textsuperscript{1}, Chen-Chi Wu\textsuperscript{3}, Yu-Chun Chang\textsuperscript{3}, and Chin-Laung Lei\textsuperscript{3}

\textsuperscript{1}Institute of Information Science, Academia Sinica
\textsuperscript{2}Research Center for Information Technology Innovation, Academia Sinica
\textsuperscript{3}Department of Electrical Engineering, National Taiwan University

ABSTRACT

Until recently, QoE (Quality of Experience) experiments had to be conducted in academic laboratories; however, with the advent of ubiquitous Internet access, it is now possible to ask an Internet crowd to conduct experiments on their personal computers. Since such a crowd can be quite large, crowdsourcing enables researchers to conduct experiments with a more diverse set of participants at a lower economic cost than would be possible under laboratory conditions. However, because participants carry out experiments without supervision, they may give erroneous feedback perfunctorily, carelessly, or dishonestly, even if they receive a reward for each experiment.

In this paper, we propose a crowdsourcable framework to quantify the QoE of multimedia content. The advantages of our framework over traditional MOS ratings are: 1) it enables crowdsourcing because it supports systematic verification of participants’ inputs; 2) the rating procedure is simpler than that of MOS, so there is less burden on participants; and 3) it derives interval-scale scores that enable subsequent quantitative analysis and QoE provisioning. We conducted four case studies, which demonstrated that, with our framework, researchers can outsource their QoE evaluation experiments to an Internet crowd without risking the quality of the results; and at the same time, obtain a higher level of participant diversity at a lower monetary cost.

Categories and Subject Descriptors

H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Evaluation/methodology; C.4 [Performance of Systems]: Design studies; H.1.2 [Models and Principles]: User/Machine Systems—Human factors

General Terms

Experimentation, Human Factors, Performance

Keywords

Bradley-Terry-Luce Model, Crowdsourcing, Mean Opinion Score (MOS), Paired Comparison, Probabilistic Choice Model, Quality of Experience (QoE)

1. INTRODUCTION

To provide better service quality for users, multimedia and network researchers endeavor to improve, respectively, the presentation quality of multimedia content and network infrastructure. As the ultimate goal is to provide a satisfying end-user experience, there is a strong need for a technique that can measure the quality of multimedia content efficiently and reliably. By “quality” we mean Quality of Experience (QoE) [22], which indicates the degree of a user’s subjective satisfaction. It is often confused with the more commonly used Quality of Service (QoS) concept, which refers to an objective system performance metric, such as the bandwidth, delay, and loss rate of a communication network.

One of the most well-known experiment frameworks for assessing the QoE of multimedia content is called the MOS (Mean Opinion Score) rating test [19]. In an MOS test, subjects are asked to give a rating from Bad (the worst) to Excellent (the best) to grade the quality of a stimulus, and the overall rating of the stimulus is obtained by averaging the scores from repeated tests. MOS scoring has been widely adopted because it is simple and intuitive; however, it is liable to cause the following problems:

1. The rating standard is somewhat obscure to experiment participants. As the concepts of the five scales, i.e., Bad, Poor, Fair, Good, and Excellent, cannot be concretely defined and explained, subjects may be confused about which scale they should give in each test.

2. Participants may have dissimilar interpretations of the scales, e.g., Poor and Good. Thus, they may provide different ratings even if they have had similar experiences with the same stimulus. According to a user’s own idiosyncratic preferences and strategy, he/she may give higher or lower scales compared to those of other participants. This is the so-called scale heterogeneity problem [31].

3. It is assumed that MOS scores are on an interval scale, but it is only an ordinal scale [36]. In fact, when people use the MOS scale, their cognitive distance between Bad (1) and Poor (2) is usually different from that between Good (4) and Excellent (5) [36]. Thus, the accuracy of MOS scores may be questionable, since they are obtained by taking arithmetic means, which does not apply to ordinal-scale measurements.
4. In MOS tests, participants are asked to grade the MOS scores for stimuli; however, we do not know whether they pay full attention to the scoring procedures, or whether they just give ratings in a perfunctory manner. To the best of our knowledge, there is no established methodology for verifying whether a participant gives false ratings either intentionally or carelessly. Consequently, it is hard to detect problematic inputs, and the measurement accuracy may be degraded due to the behavior of untrustworthy participants.

In this paper, we propose the concept of crowdsourcing experiments to the general public to achieve efficient and reliable QoE evaluations. Crowdsourcing is a neologism that means utilizing the general public’s wisdom rather than the expertise of employees or contractors. Until recently, QoE experiments were conducted in academic laboratories; however, with the advent of ubiquitous Internet access, it is now possible to ask an Internet crowd to conduct experiments on their personal computers. Our rationale is that, since the size of such a crowd can be considerable, crowdsourcing allows researchers to conduct experiments with a more diverse set of participants at a lower economic cost than is possible under laboratory conditions.

A major challenge of crowdsourcing QoE evaluations is that not every Internet user is trustworthy. Since users perform experiments without supervision, they may give erroneous feedback perfunctorily, carelessly, or dishonestly, even if they receive a reward for each experiment. For example, if we were to ask an Internet crowd to rate several video clips compressed by different codecs and we received dozens of ratings for each clip, it would be difficult, if not impossible, to determine which ratings were believable. Erroneous ratings may increase the variance of the evaluation results and lead to biased conclusions. One may argue that we could compensate for problematic inputs by conducting more experiments than necessary, but it would only be valid if dishonest and careless users comprise a small proportion of an experiment’s participants. Moreover, since dishonest users may choose random answers without following the experiment regulations, they can earn rewards more easily than other users; thus, they may be motivated to participate in as many experiments as possible to maximize their rewards. Therefore, we must find a way to detect problematic inputs in order to obtain reliable and high-quality evaluation results.

To resolve the above problem, we propose a crowdsourcable framework, based on paired comparison, for multimedia QoE evaluations. The framework not only enables us to verify the consistency of users’ inputs systematically, but also addresses the disadvantages of the MOS rating test mentioned earlier. In a paired-comparison test, a participant is simply asked to compare two stimuli simultaneously, and vote (decide) which one has the better quality based on his/her perception. Clearly, making a decision is simpler than in the MOS test, as the five-scale rating is reduced to a dichotomous choice. The features of paired comparison are as follows:

1. Like MOS, paired comparison is generalizable across a variety of multimedia applications. Thus, it can be applied to various genres of multimedia content without any modification.

2. The burden on participants is low, since they do not have to map their sensation magnitude on a categorical or numerical scale. They are only required to make simple comparative judgments, and thereby avoid the scale heterogeneity problem of MOS ratings [31].

3. We can apply probabilistic choice models [10] to analyze paired-comparison results and obtain QoE scores on an interval scale [36]. The scale enables us to quantify the discrepancy between the QoE of different evaluated targets, and also allows us to compile an arithmetically computable index for QoE management purposes [7,17].

4. The key property of paired comparison is that the experiment results can be verified. The verification basically relies on the transitivity property; that is, if A is preferred over B, and B is preferred over C, then A should also be preferred over C by the same participant. By employing this property, we can detect inconsistent judgments and remove problematic data before performing further analysis and modeling.

We demonstrate the effectiveness and generalizability of the proposed crowdsourcable experiment framework in four case studies: the first evaluates the quality of audio clips encoded by MP3 with different bit rates; the second evaluates the quality of VoIP speech with different packet loss rates; the third compares the quality of several video codecs that have similar bandwidth usage; and the fourth compares two loss concealment schemes for video playout. The studies cover a variety of multimedia applications and various factors that may affect the content’s quality. For each study, we conducted both laboratory and crowdsourced experiments. The former were performed by part-time employees under supervision; and the latter were carried out by anonymous Internet users who were interested in making some money. The results show that, overall, the quality of data obtained from the crowdsourced experiments was slightly lower than that derived from laboratory experiments. Even so, because of our approach’s ability to detect inconsistent inputs, we can still obtain comparable evaluation results at a lower economic cost and with wider participant diversity. In addition, supervision of the experiment does not require a physical space or involve labor costs.

Our contribution in this work is three-fold:

1. We propose a crowdsourcable framework, which comprises paired comparison, consistency checking, probabilistic choice modeling, and Web-based implementations, to quantify the QoE of multimedia content. The advantages of our framework over traditional MOS ratings are that 1) it facilitates crowdsourcing because it supports systematic verification of the participants’ inputs; 2) the rating procedure is simpler than that of the MOS test, so the burden on participants is lower; and 3) it derives interval-scale scores that enable further quantitative analysis and QoE management [17].

2. Our crowdsourcable framework not only enables detection of problematic inputs, but also makes “differentiated rewards” possible. That is, the reward for performing an experiment can be based on the quality

---

1Before each experiment, we informed the participants that no reward would be given if their inputs were not self-consistent; all the participants seemed to accept the rule.
2. RELATED WORK

2.1 Quality of Experience Assessment

Methods for assessing the QoE of multimedia content can be classified as either subjective or objective in nature. Subjective methods ask for human participants’ opinions in the evaluation process [8, 18]; while objective methods evaluate the QoE of a multimedia clip by analyzing its content [20, 21], e.g., checking if unnatural noise occurs in a compressed video segment. Note that the two approaches are in fact complementary rather than mutually exclusive. Subjective methods provide factual assessments of users’ experiences; on the other hand, objective methods are more convenient to use, but they require the results of subjective experiments for model development and verification.

No matter how sophisticated objective assessment methods may be, intrinsically they cannot capture all the QoE dimensions that may affect users’ experiences. For example, PESQ yields inaccurate predictions when used in conjunction with factors like listening levels, loudness loss, echo, sidetone, and the effect of delays on conversations [21]. Meanwhile, external factors, such as the quality of the headsets used in acoustic QoE evaluations, and the distance between the viewer and the display in optical QoE evaluations, are not considered by objective methods because they are hard to measure and quantify. Therefore, subjective experiments provide the most factual QoE evaluations of multimedia content, although the cost is usually high.

2.2 Paired Comparison

Paired comparison takes advantage of simple comparative judgments to prioritize a set of stimuli, and subjects’ preferences for the stimuli can be quantified via probabilistic choice modeling [10]. The technique is used in various domains, notably decision making and psychometric testing. The Analytic Hierarchy Process (AHP) [32], a well-known application of paired comparison, uses the preference priorities extracted from paired comparison results to construct a hierarchical framework that can help people make complex decisions. Paired comparison is also used in the ranking of universities [11], the rating of celebrities [26], and various subjective sensation measurements, such as pain [28], sound quality [9], and the taste of food [29].

2.3 Crowdsourcing

Crowdsourcing is a distributed model that assigns tasks traditionally undertaken by employees or contractors to an undefined crowd [5, 16]. It achieves the goal of mass collaboration via Web 2.0 technologies. The main difference between crowdsourcing and ordinary outsourcing is that a task is carried out by an unspecific Internet crowd rather than a specific group of people.

Online surveys may be the most popular application of the crowdsourcing strategy for user studies. Previous works [4, 12, 33, 37] have shown that online surveys possess a number of advantages over traditional face-to-face surveys. First, they are more efficient in terms of time and monetary cost, since it is relatively easy to collect responses from a large number of people within a short time frame online. Second, they do not have the “interviewer effect,” where the interviewer may influence how respondents answer the questions. In addition, online surveys fit in with subjects’ lifestyles; that is, subjects can respond at their convenience, so they may be more willing to complete questionnaires. However, online surveys do have disadvantages. One major issue is that researchers cannot use sampling techniques to select candidate respondents, as they do for face-to-face surveys [4, 12, 33, 37], because there is no database that covers all Internet users and their demographics.

Crowdsourcing services, such as Amazon Mechanical Turk (MTurk) [25], extend the interactivity of crowdsourcing tasks by more comprehensive user interfaces and micro-payment mechanisms. MTurk is a popular crowdsourcing service that provides a marketplace for a variety of tasks, and anyone who wishes to seek help from the Internet crowd can post their tasks on the website. Tasks can involve any kind of effort, such as participating in surveys, performing experiments, or answering certain specialized questions. Researchers have adopted MTurk to conduct user studies on image annotation [34], document relevance [2], and document evaluation [25]. Because of MTurk’s popularity, we crowdsourced our QoE evaluation experiments on the website and found that the results were satisfactory. We discuss the quality of the crowdsourced experiment results in Section 6.

2.4 Reward and Punishment Mechanisms

In crowdsourced user studies, it is important to provide proper incentives so that participants are motivated to give high quality answers. Reward and punishment mechanisms have been discussed extensively in works on peer production systems and reputation systems [30]. In peer production systems, such as Wikipedia, Yahoo! Answers, and Games with A Purpose (GWAP), users collaborate in the hope of achieving a global outcome. A number of studies have used game theoretic analysis to examine the rationality of incentives in human computation games [14, 15] and online Q&A forums [23]. In reputation systems, a user’s reputation depends on the accumulated ratings given by other users on
his/her performance. Several mechanisms [3, 13, 24] have been proposed to integrate an incentive structure into reputation systems in order to encourage good behavior and stop bad behavior.

3. THE PROPOSED FRAMEWORK

In this section, we present our framework for evaluating the QoE of multimedia content. We describe the experiment designs for evaluating audio and visual multimedia content; discuss how to assess the consistency of participants’ judgments in order to remove problematic inputs; and explain how to develop a statistical model based on the paired comparison results and how to estimate the QoE scores for the evaluated multimedia content.

3.1 Experiment Designs

Suppose we have \( n \) algorithms for processing a series of audio samples. The algorithms can be used for representation purposes, e.g., audio encoding, or for handling impairments due to errors in storage or transmission, such as error correction or loss concealment. We now present our experiment design for evaluating the effect of different audio processing algorithms on the QoE of audio recordings.

First, we need to select an audio clip, which we call the “source clip,” as the evaluation target. We apply the \( n \) audio processing algorithms to the source clip and generate \( n \) different versions of the clip, which we call the “test clips.” Since all the test clips are processed, e.g., encoded, from the same source clip, their content will be synchronized exactly. That is, except for their presentation quality, every second of the audio samples in each of the \( n \) test clips will be semantically equivalent.

Second, we create an Adobe Flash-based system for users to evaluate the \( n \) test clips. Performing experiments under our system is quite simple, as a participant only needs to use three keys, namely the SPACE key, the LEFT key, and the RIGHT key. For an \( n \)-clip experiment, \( m = \binom{n}{2} \) paired comparisons (rounds) are required. In each of the \( m \) rounds, the system randomly picks a pair of test clips that has not appeared yet, and randomly assigns one clip in the pair to the Pressed state and the other to the Released state. Once a round starts, the participant will hear one of the test clips playing continuously, depending on whether or not the SPACE key is pressed. The test clip associated with the Pressed state will be heard if the SPACE key is pressed; otherwise, the clip associated with the Released state will be heard.

Even though the clip being played is switched back and forth whenever the participant presses or releases the SPACE key, it seems that the quality level of the source clip is controllable. This is because all test clips have identical semantic content and timing structures. The design allows participants to press and release the SPACE key in an exploratory manner, and carefully listen to the difference in the quality of the two states before deciding which state yields a more pleasant experience.

Fig. 1 shows our system’s user interface. The large upper pane provides state indicators in two colors (red vs. blue) and glyphs (key pressed vs. key released) to indicate the current state (Pressed vs. Released). We do not restrict the time allowed for each round, and the test clips are played repeatedly. If the quality of two test clips differs significantly, participants should be able to tell the difference easily and make a decision within a few seconds. Sometimes the differences in quality are quite subtle, so participants may require an much longer time to make a decision. In both cases, once the participant is ready to make a decision, he/she can press the LEFT key to indicate that the quality is better in the Released state or the RIGHT key to indicate that the quality is better in the Pressed state. The system proceeds to the next round automatically after the participant has voted, and informs the participant that the experiment is finished once \( m \) paired comparisons have been made. Fig. 2 illustrates the flow of an acoustic paired comparison experiment. Readers can experience our experiment platform online at \url{http://mmnet.iis.sinica.edu.tw/link/da}.

The experiment design for evaluating the QoE of optical multimedia content, such as video clips, is very similar to that used for acoustic content. We also generate \( n \) video clips from a source clip with \( n \) processing algorithms and conduct \( m = \binom{n}{2} \) paired comparisons for each experiment. In each round, participants need to decide (vote) which state (Pressed or Released) yields a better visual quality.

The user interface of our system for optical QoE evalu-
In each experiment, we collect \( m = \binom{n}{2} \) paired comparison results for \( n \) quality levels. Since we encourage crowdsourcing of experiments in which participants may input random decisions carelessly or intentionally, we need a way to detect untrustworthy inputs. There are two reasons for this. First, problematic inputs must be removed as they may cause inaccuracies in the estimation of QoE scores. Second, we can set up punishment and reward rules, such as not paying rewards to participants who provide problematic judgments or giving more rewards to participants who consistently provide reliable inputs.

We define the Transitivity Satisfaction Rate (TSR) to quantify the consistency of a participant’s judgments over \( m \) rounds in an experiment. The computation of TSR relies on the transitivity property: that is, if \( A \) is preferred over \( B \), and \( B \) is preferred over \( C \), then \( A \) should also be preferred over \( C \) by the same participant. Based on this rule, we compute the TSR as the number of triples that “satisfy” the transitivity rule divided by the number of triples that the transitivity rule “may” apply to; thus, the value of the TSR is always between 0 and 1. The algorithm for computing the TSR is detailed in Algorithm 1. If a participant’s judgments are consistent throughout all the rounds of an experiment, the TSR will be 1; otherwise it will be less than 1. In our experience, if a participant pays full attention to an experiment, the TSR is usually higher than 0.8.

**Algorithm 1 TSR Calculation**

\[
\begin{align*}
&\text{m is an } n \times n \text{ matrix, where } m[i,j] = 1 \text{ indicates that } i \text{ is considered better than } j; \text{ otherwise } m[i,j] = 0. \\
&1: \text{n}_\text{test} \leftarrow 0 \\
&2: \text{n}_\text{pass} \leftarrow 0 \\
&3: \text{for all } i,j,k, 1 \leq i,j,k \leq n, i \neq j \neq k \text{ do} \\
&4: \quad \text{if } m[i,j] = 1 \text{ and } m[j,k] = 1 \text{ then} \\
&5: \quad \quad \text{n}_\text{test} \leftarrow \text{n}_\text{test} + 1 \\
&6: \quad \text{if } m[i,k] = 1 \text{ then} \\
&7: \quad \quad \text{n}_\text{pass} \leftarrow \text{n}_\text{pass} + 1 \\
&8: \quad \text{end if} \\
&9: \text{end if} \\
&10: \text{end for} \\
&11: \text{TSR} \leftarrow \text{n}_\text{pass}/\text{n}_\text{test}
\end{align*}
\]

For a crowdsourced experiment, we suggest presenting the computation rule of the TSR before each experiment and setting up certain punishment rules for participants who constantly produce low TSR scores. For example, in our experiments, we only pay a reward if the TSR score is higher than 0.8. Thus far, we have not received any complaints about the reward rule, which ensures that resources are not wasted on untrustworthy experiment results. It also maintains the quality of the results, as problematic data is excluded at the outset.

We believe that there is no systematic way for participants to cheat our system by inputting “smart” answers. This is because the presentation order of each pair and the order within each pair (i.e., which clip corresponds to each state) are totally random in each experiment; moreover, the information about the ordering is not available outside the system. Therefore, the only way for participants to achieve a high TSR is to pay attention to the difference in the quality of states and make judgments that are as consistent as possible. Of course a participant can still achieve a high TSR by making consistently “wrong” judgments, i.e., by always claiming that the state with the lower quality is a better one; however, such extreme cases can be detected easily by comparing their choices with those of other participants.

### 3.3 Overall Consistency Checking

After collecting the paired comparison results from a number of experiments performed by one or many participants, we can assess the overall consistency of judgments across different experiments and participants by checking the stochastic transitivity properties or computing Kendall’s \( n \)-coefficient. The stochastic transitivity approach involves checking three variants of the stochastic transitivity property, namely the weak (WST), moderate (MST), and strong (SST) stochastic transitivity. Let \( \hat{P}_{ij} \) be the empirical probability that the quality level \( i \) is considered better than the quality level \( j \), the three transitivity variants imply that if \( \hat{P}_{ij} \geq 0.5 \) and \( \hat{P}_{jk} \geq 0.5 \), then

\[
\hat{P}_{ik} \geq \begin{cases} 
0.5 & \text{(WST)}, \\
\min\{\hat{P}_{ij}, \hat{P}_{jk}\} & \text{(MST)}, \\
\max\{\hat{P}_{ij}, \hat{P}_{jk}\} & \text{(SST)},
\end{cases}
\]

for all quality levels \( T_i, T_j, \) and \( T_k \). WST is the least restrictive of the three properties. Systematic violations of WST indicate that the paired comparison results from different experiments cannot be integrated into a global preference ordering. Less severe violations of MST or SST can help decide whether probabilistic choice modeling is suitable for analyzing the choice frequencies.

Kendall’s \( u \)-coefficient is defined as follows:

\[
u = \frac{2 \sum_{i \neq j} \binom{n}{2}}{\binom{n}{2}} - 1.
\]

If participants are in complete agreement, there will be \( \binom{n}{2} \) elements containing the number \( m \) and \( \binom{n}{2} \) elements with zero in the matrix of choice frequencies, so \( u = 1 \). As the number of agreements decreases, \( u \) also decreases. The minimum agreement occurs when each element is \( m/2 \) if \( m \) is even, and \( (m \pm 1)/2 \) if \( m \) is odd; therefore, the minimum equals \(-1/(m-1)\) if \( m \) is even, and \(-1/m\) if \( m \) is odd.

### 3.4 Inference of QoE Scores

If the consistency of the collected experiment results is confirmed, we can proceed to the modeling step and infer
the quantitative QoE scores for the quality levels being evaluated.

Assume that our experiment is composed of \( n \) quality levels, \( T_1, \ldots, T_n \), thus, there are \( \binom{n}{2} \) quality-level pairs. We denote the number of comparisons for the pair \((T_i, T_j)\) as \( n_{ij} \), where \( n_{ij} = n_{ji} \). The results of paired comparisons can be summarized by a matrix of choice frequencies, represented as \( \{a_{ij}\} \), where \( a_{ij} \) denotes the number of choices that participants prefer \( T_i \) over \( T_j \) and \( a_{ij} + a_{ji} = n_{ij} \).

\[
\begin{array}{cccc}
T_1 & T_2 & T_3 & T_4 \\
T_1 & a_{12} & a_{13} & a_{14} & a_{11} \\
T_2 & a_{21} & a_{23} & a_{24} & a_{22} \\
T_3 & a_{31} & a_{32} & a_{34} & a_{33} \\
T_4 & a_{41} & a_{42} & a_{43} & a_{44} \\
\end{array}
\]

Table 1: A matrix of choice frequencies for four quality levels

By applying a probabilistic choice model [10] to the paired comparison results, we can extract an interval-scale score for each quality level. One of the most widely used models for this purpose is the Bradley-Terry-Luce (BTL) model [6,27], which predicts \( P_{ij} \), the probability of choosing \( T_i \) over \( T_j \), as a function associated with the “true” ratings of the two quality levels:

\[
P_{ij} = \frac{\pi(T_i)}{\pi(T_i) + \pi(T_j)} = \frac{e^{u(T_i)-u(T_j)}}{1 + e^{u(T_i)-u(T_j)}},
\]

where \( u(T_i) = \log \pi(T_i) \) is the estimated QoE score of the quality level \( T_i \), which can be obtained by using the maximum likelihood estimation method. We treat \( u(T_i) \) rather than \( \pi(T_i) \) as our QoE score because it comprises interval-scale metrics, but \( \pi(T_i) \) does not.

To evaluate the BTL model’s goodness of fit with the choice frequencies, we compare the likelihood \( L_0 \) of the given model and the likelihood \( L \) of the unrestricted model, which fits the frequencies perfectly. The test statistic \(-2 \log(L_0/L)\) is approximately \( \chi^2 \)-distributed with \( n - 1 \) degrees of freedom. The goodness of fit of the model can also be used to check the overall consistency of the paired comparison results. The model can be expressed in a linear form as \( P_{ij} = H(u(T_i) - u(T_j)) \) and \( H(x) = \frac{e^x}{1 + e^x} \). \( H(x) \) is a monotonically increasing function, which implies that the BTL model possesses the SST property. In other words, systematic violations of SST will preclude the validity of the BTL model. Therefore, we can also ensure the consistency of the paired comparison results by checking whether the BTL model fits the data well.

**Model Interpretation**

The computed \( u(T_i) \) for the quality level \( T_i \), from the fitted BTL model conforms to the relationship in Eq. 1. It must be negative since \( u(T_i) = \log \pi(T_i) \) and \( \pi(T_i) \) is a positive real number smaller than 1. To extract interpretable QoE scores, we normalize all the QoE scores between 0 and 1. By so doing, the quality level with the highest QoE score has a score of 1, and that with the lowest QoE always has a score of 0. Thus it is more reasonable to include a “perfect,” or at least “near-perfect,” quality level in the experiment if this normalization approach is adopted. The rationale is that it allows us to compare the QoE scores of different quality levels, assuming that the perfect scheme achieves a QoE score of 1 and the worst scheme achieves a score of 0.

4. **CASE STUDIES: ACOUSTIC QOE EVALUATION**

In this section, we present two case studies based on our experiment design for acoustic QoE evaluations (cf. Section 3.1). In the first case, we study how the QoE of MP3-encoded songs varies by altering the compression bit rates. In the second case, we investigate how packet loss and speech codec affect VoIP speech quality. While these topics may not be new to the research community, we consider that they are good starting points to demonstrate the efficacy of our framework. All the QoE evaluations were performed in three ways: in our own (physical) laboratory, crowdsourced to MTurk users, and crowdsourced to an Internet community. Here, we focus on the QoE evaluation results, which were inferred from the combined frequency choices of participants from the three sources, and their implications in each case study. We discuss how the experiments were crowdsourced and then compare the performances of the laboratory and crowdsourced experiments in Section 6.

4.1 **MP3 Compression Level**

In audio compression, there is a trade-off between maintaining good sound quality and reducing the size of the audio data. A higher encoding bit rate usually yields better quality output; however, the cost is a larger file, which increases the demand for data storage and network bandwidth in streaming applications. In this case study, we investigate the QoE of MP3-compressed audio clips with different compression levels. We selected two English songs, the fast-paced “Wake Me Up Before You Go Go” and the slow-paced “Garden of Graves,” as the source clips. To obtain test clips, we converted the songs into MP3 CBR format with six bit rate levels, namely, 32, 48, 64, 80, 96, and 128 Kbps. Consequently, for each song, we had 6 stimuli (quality levels) and \( \binom{6}{2} = 15 \) paired comparisons in each experiment.

There were 127 participants, both part-time employees and Internet volunteers (cf. Section 6), who performed 244 experiments on a total 3,660 paired comparisons. For each paired comparison, the participant used the interface shown in Fig. 1 to indicate which quality level yielded a better listening experience. Using the probabilistic choice modeling technique described in Section 3.4, we estimated the QoE scores of the 6 compression levels for each song, and plotted them on the graph shown in Fig. 4, where the vertical bar denotes the 95% confidence band of the score on each point. From the graph, we observe that a higher bit rate constantly leads to a higher QoE score. Also, the law of
4.2 Effect of Packet Loss on VoIP

In this case study, we investigated the effect of packet loss on VoIP speech quality. Our source clip was a three-minute speech recording made by concatenating uncompressed speech segments from the Open Speech Repository. We compressed the clip with two speech codecs, G722.1 and G728, into voice packets by using the Intel Integrated Performance Primitives (IPP) library. Then, we simulated packet loss events, such as those due to network loss or variable network delays, and decoded the remaining voice packets into degraded speech recordings. The simulated packet loss rates were 0%, 4%, and 8%. Because of the combination of two speech codecs and three loss rates, we obtained a total of 6 test clips for the QoE evaluation experiments.

A total of 62 participants performed 103 experiments, which involved 1,545 paired comparisons. Fig. 5 shows the QoE scores for the 6 test clips. From the graph, it is clear that a higher packet loss rate leads to a lower QoE score.

5. CASE STUDIES: OPTICAL QOE EVALUATION

In this section, we present two case studies based on our experiment design for optical QoE evaluations (cf. Section 3.1). The first evaluates the QoE of video clips compressed by different codecs with comparable bit rates; and the second considers the impact of two loss concealment schemes during playout of IPTV videos. Here, we focus on the QoE evaluation results, which were inferred from the combined frequency choices of participants from the three sources, and their implications in each case study. We explain how the experiments were crowdsourced and then compare the performances of the laboratory and crowdsourced experiments in Section 6.

5.1 Video Codec

In this case study, we assess the impact of codecs and compression levels on the QoE of video clips. From the video database of the Video Quality Experts Group (VQEG), we selected two 12-second raw video clips, the fast-motion “Cheerleaders” and the slow-motion “Mobile Calendar.” We compressed both source clips with three codecs, H.264, WMV3, and XVID, at two bit rates, 400 Kbps and 800 Kbps respectively. Therefore, for each source clip, we obtained a total of 6 test clips with different codec-and-bit-rate combinations.

A total of 121 participants, both part-timer employees and Internet volunteers, performed 223 experiments that involved 3,345 paired comparisons. Fig. 6 shows the QoE score of each test clip for both “Cheerleaders” and “Mobile Calendar.” Generally, the quality of 800-Kbps clips is much better than that of 400-Kbps clips because more information is encoded in the output video clips. If we consider the “Cheerleaders” video, we find that H264 performs better than WMV3, which is better than XVID at 400 Kbps; however, the quality of WMV3 and XVID is comparable at
5.2 Loss Concealment Scheme

When designing a high-quality IPTV system, one of the most challenging issues is how to deal with video packet loss due to network loss or excessive variations in network delay. A number of loss concealment schemes have been proposed, e.g., the intuitive frame copy method and the more sophisticated error resilient coding approach. In this case study, we evaluated two loss concealment schemes, namely the frame copy (FC) scheme and the frame copy with frame skip (FCFS) scheme [35], under different degrees of packet loss. The FC scheme conceals errors in a video frame by replacing a corrupted block with the block in the corresponding position in the previous frame. On the other hand, FCFS is a hybrid scheme that integrates the frame copy and the frame skip technique, which simply drops a frame that is corrupted due to packet loss. In our implementation of FCFS, if the percentage of corrupted slices in a frame exceeds 10%, it will skip (i.e., drop) the frame; otherwise it will apply the frame copy method to conceal the errors. The “Cheerleaders” and “Mobile Calendar” video clips (Section 5.1) were also used in this case study. We compressed both clips by the frame copy method to conceal the errors. The “Cheerleaders” and “Mobile Calendar” video clips are a hybrid scheme that integrates the frame copy and the frame skip technique, which simply drops a frame that is corrupted due to packet loss. In our implementation of FCFS, if the percentage of corrupted slices in a frame exceeds 10%, it will skip (i.e., drop) the frame; otherwise it will apply the frame copy method to conceal the errors. The “Cheerleaders” and “Mobile Calendar” video clips (Section 5.1) were also used in this case study. We compressed both clips by the frame copy method to conceal the errors. The “Cheerleaders” and “Mobile Calendar” video clips (Section 5.1) were also used in this case study. We compressed both clips by the frame copy method to conceal the errors. The “Cheerleaders” and “Mobile Calendar” video clips (Section 5.1) were also used in this case study. We compressed both clips by the frame copy method to conceal the errors.

A total of 91 participants performed 183 experiments that involved 2,745 paired comparisons. Fig. 7 shows the QoE score for each test clip of “Cheerleaders” and “Mobile Calendar.” We find that FCFS performs slightly better than FC on “Cheerleaders” when the loss rate is moderate (≤5%). This may be because FCFS skips seriously corrupted frames so that the subjects perceive better temporal quality. However, when the loss rate is high (8%), FCFS drops a large number of frames, so its QoE is inferior to that of FC. Interestingly, the situation is reversed in the case of the “Mobile Calendar” clip. FCFS outperforms FC at moderate to high loss rates (≥5%). We believe this is because dropping frames in a slow-motion video does not lead to significant freezing effects. On the other hand, FC provides better QoE at the 1% loss rate. This is reasonable because, when the damage caused by packet loss is small in a slow-motion video, FC can easily repair most of the corrupted blocks. This case study demonstrates two points: 1) the effectiveness of our framework for evaluating the QoE of video clips; and 2) the effect of loss concealment depends to a large extent on the characteristics of the target video clips.

6. DISCUSSION

In this section, we compare the efficiency and effectiveness of laboratory and crowdsourced experiments performed for the case studies in Section 4 and Section 5. The experiments were conducted in the three ways:

- **Laboratory:** We recruited part-time workers at an hourly rate of US$8. They were asked to perform the experiments repeatedly during their working hours.
- **MTurk:** We posted each experiment as a HIT (Human Intelligence Task) on the Mechanical Turk website. If an experiment was qualified, i.e., it yielded a TSR higher than 0.8, we paid the participant 0.15 US dollars.
- **Community:** We posted an advertisement on the website of an Internet community with 1.5 million members to seek participants for our experiments. For each experiment that qualified, we paid the participant an amount of virtual currency that was equivalent to one US cent.

In total we spent US$173.88 on 753 experiments, which were performed by 298 participants and involved 11,295 paired comparisons. The performances and costs of all the participant sources in the case studies are summarized in Table 2. Next, we discuss the differences between the laboratory and crowdsourced experiments in terms of quality, cost, and participant diversity.

**Quality.** We define the Qualified Rate as the ratio of experiments that yield a TSR higher than 0.8. In the experiments, the rate was generally between 60% and 70%. The laboratory experiments achieved the highest rates in all cases, except for the VoIP case study. Moreover, in the study of loss concealment schemes, the rates for laboratory experiments were as high as 69%, compared to approximately 35% on both crowdsourcing sites. The latter rates indicate that it is difficult to differentiate the quality of video clips with different loss concealment schemes. We believe the superiority of laboratory experiments in this case is due to the difference in the participants' proficiency. On average, the laboratory participants and crowdsourcing participants performed 115 and 18 comparisons respectively. Hence, the former had more opportunities to gain experience in distinguishing the subtle differences in the quality of video clips.

We also checked the overall consistency of experiment results from three participant sources. After removing unqualified experiments, we computed the average TSR of the
Table 2: A comparison of laboratory and crowdsourced experiments in terms of their cost and performance

<table>
<thead>
<tr>
<th>Case Study</th>
<th>Participant Source</th>
<th>Total Cost (dollar)</th>
<th># Rounds</th>
<th># Person</th>
<th>Qualified Rate</th>
<th>Cost / Round (cent)</th>
<th>Time / Round (sec)</th>
<th>Avg. TSR</th>
<th>WST Violation</th>
<th>MST Violation</th>
<th>SST Violation</th>
<th>Kendall</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP3 Bit Rate</td>
<td>Laboratory</td>
<td>50.97</td>
<td>1,440</td>
<td>10</td>
<td>67%</td>
<td>3.54</td>
<td>16</td>
<td>0.96</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>MTurk</td>
<td>1.00</td>
<td>750</td>
<td>24</td>
<td>47%</td>
<td>1.00</td>
<td>9</td>
<td>0.96</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Community</td>
<td>1.00</td>
<td>1,470</td>
<td>95</td>
<td>94%</td>
<td>0.07</td>
<td>25</td>
<td>0.96</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
<td>0.38</td>
</tr>
<tr>
<td>VoIP Quality</td>
<td>Laboratory</td>
<td>22.95</td>
<td>675</td>
<td>10</td>
<td>67%</td>
<td>3.40</td>
<td>16</td>
<td>0.98</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>MTurk</td>
<td>3.00</td>
<td>300</td>
<td>15</td>
<td>75%</td>
<td>1.00</td>
<td>19</td>
<td>0.98</td>
<td>0</td>
<td>0</td>
<td>0.20</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Community</td>
<td>0.40</td>
<td>570</td>
<td>31</td>
<td>86%</td>
<td>0.56</td>
<td>24</td>
<td>0.98</td>
<td>0</td>
<td>0</td>
<td>0.18</td>
<td>0.85</td>
</tr>
<tr>
<td>Video Codec</td>
<td>Laboratory</td>
<td>25.74</td>
<td>1,500</td>
<td>10</td>
<td>80%</td>
<td>1.58</td>
<td>7</td>
<td>0.98</td>
<td>0</td>
<td>0</td>
<td>0.12</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>MTurk</td>
<td>3.05</td>
<td>406</td>
<td>23</td>
<td>65%</td>
<td>1.00</td>
<td>11</td>
<td>0.95</td>
<td>0</td>
<td>0</td>
<td>0.20</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>Community</td>
<td>0.95</td>
<td>1,340</td>
<td>88</td>
<td>71%</td>
<td>0.07</td>
<td>11</td>
<td>0.97</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
<td>0.60</td>
</tr>
<tr>
<td>Loss Concealment</td>
<td>Laboratory</td>
<td>51.93</td>
<td>1,260</td>
<td>11</td>
<td>69%</td>
<td>1.41</td>
<td>19</td>
<td>0.96</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>MTurk</td>
<td>5.85</td>
<td>980</td>
<td>21</td>
<td>36%</td>
<td>1.00</td>
<td>25</td>
<td>0.97</td>
<td>0</td>
<td>0</td>
<td>0.40</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>Community</td>
<td>0.63</td>
<td>900</td>
<td>39</td>
<td>35%</td>
<td>0.07</td>
<td>21</td>
<td>0.96</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
<td>0.60</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>173.88</td>
<td>11,295</td>
<td>298</td>
<td>59%</td>
<td>1.54</td>
<td>17</td>
<td>0.97</td>
<td>0</td>
<td>0.01</td>
<td>0.20</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Experiments from the three sources and found that it was above 0.95. The statistical transitivity checks show that no WST violations occurred in any of the data sets, and only 1 in 20 MST checks failed for the MTurk experiments in one of the case studies. SST violations occurred in all case studies, but the numbers of such violations were moderate. We remark that the laboratory experiments had the fewest SST violations. This is reasonable because statistical transitivity checks assess the consistency of judgments among different participants and the laboratory experiments involved the fewest participants. In all the case studies, the Kendall-\(u\) parameters were higher than 0.5, which indicates that the judgments provided by all three sources were reasonably consistent.

Cost. The laboratory experiments accounted for 86% of the total monetary cost. Since the number of experiments performed by participants in each source was different, we compare the cost of the sources in terms of the cost per round, as shown in Table 2. The cost per round was not a constant price in laboratory experiments because the part-time employees were paid an hourly rate, but the number of experiments they performed varied. On average, the cost per round of laboratory experiments was 3 cents; and for the crowdsourced MTurk and community experiments it was 1 and 0.07 cents respectively, which yields a ratio of 3 : 1 and 43 : 1 respectively.

Participant diversity. The diversity of participants in QoE evaluation experiments is important. Since the purpose of such experiments is to understand people’s perceptions of certain stimuli, such as multimedia content, a more diverse set of experiment participants enables us to collect a broader range of opinions. From this perspective, crowdsourcing is obviously a more appropriate strategy for QoE experiments because it increases participant diversity substantially. Quantitatively, the crowdsourced experiments accounted for only 14% of the total cost, but they accounted for 289 out of the 298 participants (97%) in our case studies.

Limitations

Although we have demonstrated the advantages of crowdsourcing in QoE evaluation experiments, the strategy has several limitations that may affect its applicability in certain types of evaluations.

Environment control. In crowdsourced experiments, the subjects might listen to or view the presented media content under different conditions, such as lighting, screen size, and the quality of headsets; however, in a laboratory, the experiments can be conducted under a common, controlled environment. This could be considered as an advantage or a disadvantage. It is an advantage because users’ perceptions can be assessed in real-life scenarios. In practice, users’ headphones may not be as good as those in laboratories, and ambient sound may be unavoidable. While it is difficult to simulate a “typical” user environment in a laboratory, crowdsourcing allows us to assess people’s real-life experiences. On the other hand, it can be a disadvantage if the purpose of experiments is to accurately measure the quality of multimedia content in a specific scenario.

Media. Since most people connect to the Internet via personal computers, the crowdsourcing strategy is most suitable for evaluating the media content on personal computers. It could be a problem if evaluations are conducted on other input/output devices, such as TV or electronic papers. Fortunately, those non-PC devices are gradually becoming Internet-capable, so this problem may be resolved in the near future.

Demography. As any Internet user can participate in crowdsourced experiments, it is difficult, if not impossible, to relate the experiment results to demographic factors, such as gender and age. For example, we cannot investigate the effect of age on the perceptions of certain colors in video clips, because the ages self-reported by the participants may not be trustworthy.

Even though the crowdsourcing strategy has the above limitations, it can still be used to evaluate the QoE of multimedia content in a variety of applications. We believe that it is especially helpful for assessing the effect of techniques related to coding, processing, and transmission of media content, such as the case studies we conducted in this work. Moreover, with the rapid advent of technologies on rich user interface, such as Flash, the framework can be extended to assess users’ experiences in interactive applications like computer games, which will also be part of our future work.

7. CONCLUSION AND FUTURE WORK

In this work, we have proposed and evaluated a crowdsourcing framework for assessing the QoE of multimedia content. We have shown that, under our framework, researchers can outsource QoE evaluation experiments to an
evaluating the quality of multimedia content. Evaluations will prove helpful to researchers interested in the proposed crowdsourcing framework and platform for QoE studies with an application to university rankings. Journal of the Aca