Crowdsourcing Multimedia QoE Evaluation: A Trusted Framework

Chen-Chi Wu, Kuan-Ta Chen, *Member, IEEE,* Yu-Chun Chang, *Student Member, IEEE,* and Chin-Laung Lei

Abstract

Crowdsourcing has emerged in recent years as a potential strategy to enlist the general public to solve a wide variety of tasks. With the advent of ubiquitous Internet access, it is now feasible to ask an Internet crowd to conduct QoE (Quality of Experience) experiments on their personal computers in their own residences rather than in a laboratory. The considerable size of the Internet crowd allows researchers to crowdsource their experiments to a more diverse set of participant pool at a relatively low economic cost. However, as participants carry out experiments without supervision, the uncertainty of the quality of their experiment results is a challenging problem.

In this paper, we propose a crowdsourceable framework to quantify the QoE of multimedia content. To overcome the aforementioned quality problem, we employ a paired comparison method in our framework. The advantages of our framework are: 1) trustworthiness due to the support for cheat detection; 2) a simpler rating procedure than that of the commonly-used but more difficult mean opinion score (MOS), which places less burden on participants; 3) economic feasibility since reliable QoE measures can be acquired with less effort compared with MOS; and 4) generalizability across a variety of multimedia content. We demonstrate the effectiveness and efficiency of the proposed framework by a comparison with MOS. Moreover, the results of four case studies support our assertion that the framework can provide reliable QoE evaluation at a lower cost.

Index Terms

Crowdsourcing, Mean Opinion Score, Paired Comparison, Probabilistic Choice Model, Quality of Experience, Subjective Test.

I. Introduction

In recent years, multimedia services have become immensely popular and have been widely accessed via not only desktop and laptop computers but also mobile devices. Because of the heterogeneity of hardware capabilities and network environments, providing users with a satisfying experience no matter what platform they are using is an ultimate goal of service providers. However, measuring the quality of multimedia content efficiently and reliably has long been a real challenge. Here by “quality” we refer to Quality of Experience (QoE) [39], which reflects the degree of a user’s subjective satisfaction. It should not be confused, however, 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.

There have been numerous methods proposed to evaluate the QoE of multimedia content employing *objective* or *subjective* methods. Objective methods, such as PESQ (Perceptual Evaluation of Speech Quality) [37] and PEVQ (Perceptual Evaluation of Video Quality) [36], are convenient because they do not have the expenses associated with participants; however, their QoE assessment is not as factually accurate as that of subjective methods, such as the widely-used MOS rating method [34] for multimedia QoE assessments. In this method, each participant is asked to grade the quality of multimedia content on a five-point scale, which ranges from 1 (Bad) to 5 (Excellent),

A preliminary version of this paper [13] appeared in the Proceedings of the 17th ACM International Conference on Multimedia (ACM Multimedia 2009).

Manuscript received February 19, 2012; revised June 17, 2012; accepted November 02, 2012.

Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei are with the Department of Electrical Engineering, National Taiwan University, Taipei 106, Taiwan (e-mail: bipa@fractal.ee.ntu.edu.tw; congo@fractal.ee.ntu.edu.tw; lei@cc.ee.ntu.edu.tw)

Kuan-Ta Chen is with the Institute of Information Science and the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan (e-mail: ktchen@iis.sinica.edu.tw).

with the MOS score as the arithmetic mean of all the participants' scores. Although the MOS rating method has a long history of pervasive use, it has three fundamental problems:

1) **Rating scale mapping**. As the concept of the five scales, i.e., Bad, Poor, Fair, Good, and Excellent, cannot be concretely defined and explained, participants may confront the problem of mapping their perception on the scale.
2) **Rating scale heterogeneity** [61]. Participants may have dissimilar interpretations of the scales according to their idiosyncratic preferences and strategies. Therefore, they may give different ratings even if they have had similar experiences with the same stimulus.
3) **Lack of a cheat detection mechanism**. Participants may not pay full attention to experiments or determine ratings cautiously. To the best of our knowledge, there is no established mechanism for verifying whether a participant gives false ratings either intentionally or carelessly. Consequently, it is hard to detect untrustworthy inputs, which would degrade the accuracy of assessment.

Apart from the above intrinsic problems, another issue that warrants consideration is the high economic cost, which is also the impediment to all subjective methods. It is expensive to obtain the sufficiently large number of participants that are required in order to obtain reliable QoE assessments. In addition, as QoE experiments are traditionally conducted in laboratory settings, it is not efficient due to the difficulty of collecting so many experiment results in a short time frame.

In this paper, we use the concept of crowdsourcing in order to take advantage of the power of the masses to achieve efficient and reliable QoE assessments. The term "crowdsourcing" is a neologistic portmanteau of "crowd" and "outsourcing" for describing the act of outsourcing tasks to an undetermined crowd of people rather than employees or contractors. The advent of ubiquitous Internet access has opened the door to Internet crowds who can be asked to conduct experiments on their personal computers, thus freeing QoE experiments from the limits of closed laboratories. Because of the considerable size of the Internet crowd, we believe that *crowdsourcing allows researchers to conduct experiments with a more diverse set of participants at a lower economic cost than is possible under laboratory conditions.*

Nonetheless, an obvious challenge of crowdsourcing QoE evaluations is the fact that *not every Internet user is trustworthy*. Due to the lack of supervision when subjects perform experiments, they may provide erroneous responses perfunctorily, carelessly, or dishonestly. Moreover, if tasks are not intuitive and simple enough, participants may also give problematic answers because they do not fully realize how or what they should do. This is a problem because erroneous feedback increases the uncertainty of the evaluation results and leads to biased conclusions. And although one may argue that we could compensate for untrustworthy inputs by conducting more experiments than necessary, this would be a valid course of action only if untrustworthy users comprise a small proportion of an experiment's participants. Further, since participants are paid wages for each experiment, there is more incentive for dishonest users to participate in as many experiments as possible by giving random answers quickly without obeying the instructions[1]. It is therefore important to find a way to prevent or detect untrustworthy inputs to ensure reliable and high-quality evaluation results. In summary, while the MOS method is widely used, we consider it unsuitable in crowdsourcing not only because its five-point scale is not simple enough to be mastered well by an inexperienced crowd of users, but also because its inability to detect untrustworthy results.

To address the above problems, we propose a trusted crowdsourceable framework, based on *paired comparison* [17], for multimedia QoE evaluations. In a paired comparison experiment, a participant is simply asked to compare two stimuli simultaneously, and decide *which one* has the better quality based on his perception. In this task, the dichotomous decision of perceptual quality is clearly much simpler than the five-point rating in the MOS method. There are five key features of the paired comparison method:

1) It is generalizable across various types of multimedia content without any modification.
2) It is simple for participants since they are only asked to make intuitive *comparative judgements* instead of mapping their perception on a categorical or numerical scale. The scale heterogeneity problem of MOS ratings [61] is thus eliminated.
3) Its comparative judgement results can be analyzed by probabilistic choice models [17] to obtain QoE measures on an *interval scale* [69], which allows us to compile an arithmetically computable index for QoE management purposes [10, 33].

[1] It is foreseeable that certain automation schemes, e.g., bots [11], would be used to perform such crowdsourcing tasks repeatedly if it is considered a profitable "business" by malicious attackers.

4) It is trustworthy with the support for cheat detection mechanisms as outlined in Section IV. Our framework relies on *the transitivity property*, which means that if a participant prefers A to B and B to C, then he should also prefer A to C. By employing this property, we can effectively check the consistency of judgements and remove untrustworthy inputs.
5) It is efficient and economic since a reliable QoE assessment can be obtained with less effort compared with the MOS approach (cf. Section VI-E).

To demonstrate the efficacy and generalizability of our framework, we conducted four case studies that targeted audio and visual multimedia QoE evaluations. To compare the cost and performance of laboratory and crowdsourcing strategies, experiments were conducted by part-time employees under supervision and by anonymous Internet users. The results show that, overall, the quality of raw inputs obtained from the crowdsourced experiments was slightly lower than that derived from laboratory experiments. Even so, because of our approach's ability to detect untrustworthy inputs, we can obtain comparable evaluation results at a lower economic cost and with wider participant diversity.

There are three important contributions of our work:

1) We propose a trusted crowdsourcing framework, which comprises paired comparison, probabilistic choice modeling, consistency checking, and cheat detection mechanisms, 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 method; and 3) it exhibits high intra- and inter-subject reliability and provides precise QoE estimates in less time (cf. Section VI).
2) Our crowdsourceable framework not only enables detection of untrustworthy inputs, but also makes "differentiated rewards" possible. That is, the amount of wage for an experiment can be based on the quality (i.e., consistency) of a participant's inputs. This design provides participants an incentive to provide judgements honestly, carefully, and seriously.
3) To demonstrate the efficacy of our framework, we conduct four case studies involving audio and visual multimedia content. The results of the laboratory and crowdsourced experiments indicate that we can obtain comparable evaluation results at a lower economic cost and with wider participant diversity.

The remainder of this paper is organized as follows. Section II contains a review of related works. We describe the proposed framework in Section III, and introduce the cheat detection mechanism in Section IV. In Section V, we demonstrate the effectiveness of the proposed framework by applying it to the evaluation of the QoE of audio and visual content. Section VI provides a side-by-side comparison between the paired comparison and MOS approaches, and we discuss issues related to paired comparison and crowdsourcing in Section VII. Finally, in Section VIII, we present our conclusions.

II. Related Work

A. QoE Assessment Methods

QoE of multimedia content can be measured by subjective methods or objective methods. Subjective methods require participants to indicate their opinions in an evaluation process [12, 14, 34, 38]. The absolute category rating (ACR) approach [38] and the degradation category rating (DCR) approach [38] are two representative examples of subjective assessment methodologies. While ACR requires subjects to grade the quality of individual multimedia content, DCR requests that subjects grade the *quality difference* between a pair of multimedia content. Both approaches take averages of subjects' ratings to quantify the QoE of each content. The mean opinion score (MOS) approach [34], which has been widely used in quality assessment studies [30, 59, 60, 68], is an instance of ACR with a five-level scale (i.e., Bad, Poor, Fair, Good, and Excellent). As for paired comparison [38], which can be considered a variation of DCR, it has in recent years received more attention due to its capacity for quality assessment [7, 16, 32, 45, 48, 56, 65, 71].

Objective methods can be divided into two categories: signal-based methods and parameter-based methods. PESQ (Perceptual Evaluation of Speech Quality) [37], an example of signal-based methods for assessing QoE of speech content, takes the original signal and a degraded signal as inputs, and evaluates the quality of the degraded signal based on noise and audible distortion. E-Model [35] is an example of parameter-based methods that is used for evaluating QoE of VoIP conversations by taking impairment factors such as network delay and quantizing distortions as inputs. A hybrid model [66] that integrates PESQ and E-Model has been proposed to more accurately predict VoIP QoE by leveraging the advantages of both methods.

Whereas objective methods are indeed convenient to use, subjective methods nonetheless provide factual assessments of users' experiences. No matter how sophisticated objective assessment methods may be, intrinsically they cannot capture every QoE dimension that may affect users' experiences. For example, PESQ yields inaccurate predictions when used in conjunction with factors like listening levels, echo, and sidetone [51]. Meanwhile, E-Model does not consider the variability of network delays and loss rates, and the interaction of factors, such as the interplay of network delay and listening quality [18]. Therefore, to obtain factual QoE evaluation results, subjective methods are still required, even though the cost is higher.

B. Paired Comparison

Paired comparison takes advantage of simple comparative judgements to prioritize a set of stimuli, where subjects' preferences for the stimuli can be quantified via probabilistic choice modeling [17, 72]. This technique is used in various domains, notably decision making and psychometric testing. The Analytic Hierarchy Process (AHP) [62], a well-known application of paired comparison, uses the preference priorities extracted from paired comparison results to construct a hierarchical framework that can help individuals and enterprises make complex decisions. Paired comparison has also been employed in the ranking of universities [19], the rating of celebrities [44], and various subjective sensation measurements, such as pain [52], sound quality [15], and the taste of food [55].

C. Crowdsourcing

Crowdsourcing is a distributed model that assigns tasks traditionally undertaken by employees or contractors to an undefined crowd [5, 20, 31], achieving its 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 unspecified Internet crowd rather than a specified group of people. Performing psychological experiments on the Internet, for example, is one academic application of the crowdsourcing strategy. In [3], the authors thoroughly discussed the pros and cons of Internet psychological experiments based on a number of case studies and proposed some solutions to address the data validity issues of such experiments.

A number of crowdsourcing platforms have emerged in recent years. For example, InnoCentive[2] enables organizations to utilize the intellect of the global scientific community to find innovative solutions to challenging research and development problems. Amazon's Mechanical Turk[3] (MTurk) is probably the most popular crowdsourcing platform, which provides a marketplace for a variety of tasks, and anyone who wishes to seek help from the Internet crowd can post their task requests on the website. On the platform, tasks are known as human intelligence tasks (HITs) and 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 such as image annotation [64, 70], document relevance [1], and document evaluation [43]. Because of MTurk's popularity, we also crowdsourced our QoE evaluation experiments on the platform and found that the results were satisfactory with the proposed framework. We will discuss our experiment results in Section V.

D. QoE Assessment Studies based on Paired Comparison and/or Crowdsourcing

In the past few years, paired comparison and crowdsourcing have been receiving increased attention. There are a number of attempts which employ either or both of these strategies in multimedia QoE assessment studies. We now summarize the latest research categorized by their adopted strategies.

1) Paired Comparison: Paired comparison alone can be seen an alternative to the commonly-used MOS approach due to its simple and intuitive judgements. In [48], Lee et al. investigated how the different layers defined in Scalable Video Coding (SVC) affected the viewers' perceptual quality about video clips. Their results indicate that paired comparison is a robust methodology for externalizing users' opinions even when the QoE of video clips were simultaneously affected by multi-dimensional factors. In [32], Huang et al. investigated participants' quality perceptions when they were interacting with remote parties via a tele-immersive remote conferencing platform, in

[2] `http://www.innocentive.com`
[3] `http://www.mturk.com`

which paired comparison was shown to be able to capture users' opinions about the interactivity experience. In a different context, Lan et al. [45] employed paired comparison to evaluate whether their infrastructure-less VoIP communication protocol performs better than others in terms of user perceived voice quality.

2) Crowdsourcing: The crowdsourcing strategy alone can reduce the cost of hiring subjects and increase the efficiency (i.e., shortening turn-around time) of QoE assessment experiments. In [30], the authors asked an Internet crowd to report their experience in watching YouTube videos using an MOS rating scale. They adopted the "gold standard" approach to detect untrustworthy inputs, with a few "trap" questions whose standard answers are commonly known, and which were randomly spread over the questionnaire for voluntary participants. If a participant failed to provide correct answers to any of the trap questions, his inputs were considered untrustworthy and removed from the dataset. However, although the gold standard approach is considered effective, the "trap" questions still need to be manually designed for each experiment, thus rendering the approach ineffective when automation attacks [2] are used.

Ribeiro et al., on the other hand, adopted an outlier-detection approach in their crowdsourced QoE assessment studies for audio clips [60] and images [59]. A subject's inputs were removed if his own MOS estimates (i.e., the MOS scores estimated from his ratings) had a Pearson's correlation coefficient with the global MOS estimates lower than 0.25. However, whereas this strategy ensures that the global QoE estimates represent the opinions of the majority, it can only be applied after a significant amount of user inputs have been collected and when the ratio of malicious inputs is small.

3) Paired Comparison and Crowdsourcing: In light of the above issues, conducting QoE assessments using paired comparison in a crowdsourcing context is a reasonable strategy, as crowd participants tend to prefer simple tasks such as dichotomous judgements required by paired comparison. We can see that the advantage of such integrated use of paired comparison and crowdsourcing has been noticed and used in a number of QoE assessment studies such as 1) evaluations of users' preferences over text-to-speech (TTS) systems [7], 2) studies of users' perceptions of image quality [56, 65], 3) evaluations of QoS adaptation schemes for 3D video processing [71], and 4) assessments of viewers' preferences of American Sign Language (ASL) videos [16]. Among these studies, Buchholz and Latorre [7] adopted the gold standard strategy and Sprow et al. [65] used a control group (i.e., non-crowdsourced experiments performed in a laboratory) to ensure the validity of the user inputs from the Internet crowd, while the other studies [16, 56, 71] did not include quality control efforts for crowdsourced tasks.

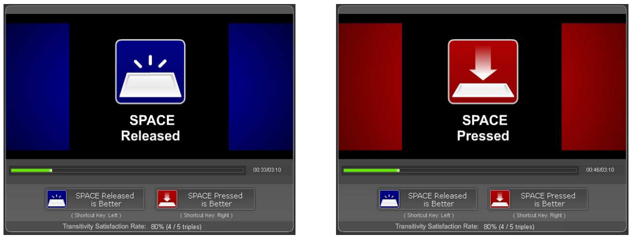
We see the inconsistent uses of data assurance mechanisms as an indication that 1) quality control of crowdsourcing tasks is needed as unreliable inputs will affect the credibility of QoE assessment results, and 2) there is a strong demand for a general, systematic input validation framework for crowdsourced QoE assessments, which is exactly the goal of this present work.

E. Reward and Punishment Mechanisms

In crowdsourced user studies, it is important to provide proper incentives so that participants are motivated to give high quality representative answers. In [29], Horton and Chilton investigated the relationship of wages and task difficulty, and proposed a model to estimate participants' reservation wage in crowdsourcing tasks. A number of reward and punishment mechanisms [40, 46, 57] have also been proposed to encourage users to provide quality contributions in the context of peer production systems. Lee et al. [46], for example, introduced a voting-based reward mechanism for online Q&A forums, where not only the contributor of the winning answer but also the users who voted for the answer were rewarded. Game theory has also been applied to investigate the rationality of incentives in human computation games [27, 28]. While reward and punishment mechanisms are not the focus in this study, our proposed framework provides a trust index for each experiment taken by an Internet participant. The "boss" (i.e., the person who crowdsources tasks) is therefore free to decide how to facilitate the respective distribution of rewards and punishments to task workers who perform exceptionally well and poorly.

III. QoE Assessment Framework

In this section, we present a crowdsourceable framework to quantify the QoE of multimedia content. We describe the experiment designs for evaluating audio and visual multimedia content, and explain how to estimate the QoE score for the evaluated multimedia content by applying a probabilistic choice model to the paired comparison inputs.

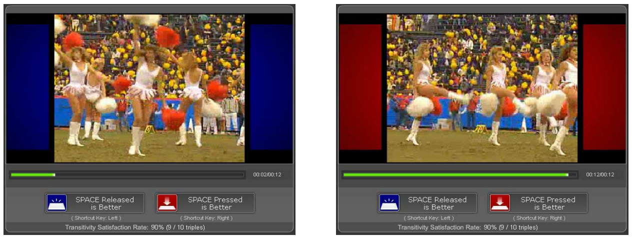


Transitivity Satisfaction Rate: 80% (4 / 5 triples)

Transitivity Satisfaction Rate: 80% (4 / 5 triples)

(a) Released state (b) Pressed state

Fig. 1. The user interface for the audio QoE evaluation experiment.



Transitivity Satisfaction Rate: 90% (9 / 10 triples)

Transitivity Satisfaction Rate: 90% (9 / 10 triples)

(a) Released state (b) Pressed state

Fig. 2. The user interface for the visual QoE evaluation experiment.

A. Experiment Design

Suppose we have n algorithms for processing audio content. The algorithms can be designed for format-conversational purposes, e.g., audio encoding, or for handling impairments due to errors in storage or transmission, such as error correction and loss concealment. Using audio content as an example, we now present our experiment design for evaluating the effect of different audio processing algorithms on the QoE of a given audio recording.

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 are accurately synchronized. That is, with the exception of the level of their presentation quality, every second of the audio samples in each of the n test clips are semantically equivalent.

Next, these n test clips are fed into an evaluation platform. For an n-clip experiment, the total of $m = \binom{n}{2}$ possible pairs of test clips requires m rounds of comparison. In each of the m rounds, the system *randomly* picks a pair 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 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 and held down; otherwise, the clip associated with the **Released** state will be heard. Whenever the participant switches the playout state, say, from the **Pressed** state to the **Released** state at t second, the clip associated with the **Released** state will take over seamlessly and start playing from t second and afterward. This design makes participants feel that the quality level of the source clip is controllable, and allows them to carefully listen to the difference in the quality of the two states by switching back and forth at any time.

Figure 1 shows an exemplar user interface of the evaluation platform. The large upper pane provides state indicators in two colors (blue vs. red) and glyphs (key released vs. key pressed) to indicate the current state (**Released** vs. **Pressed**). Participants are allowed an unrestricted time in 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 quick decision. Sometimes the differences in quality are quite subtle, so participants may require a longer time to make a decision[4]. Once the participants are ready to make a decision, they can press the LEFT arrow key to indicate that the quality is better in the **Released** state or the RIGHT arrow 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 after all m paired comparisons have been made.

[4] Although we did not specify a time limit, each round normally took between 5 and 25 seconds in our experiments.

Performing experiments on the system is quite simple in that participants only need to use three keys, namely the SPACE key, the LEFT arrow key, and the RIGHT arrow key.

Our experiment design for evaluating the QoE of visual multimedia content, such as video clips, is similar to that used for audio content. We also generate n video clips from a source clip with n processing algorithms and conduct $m = \binom{n}{2}$ rounds of paired comparison for each experiment. In each round, participants need to decide (vote) which state (**Released** or **Pressed**) yields a better visual quality. An exemplar user interface for visual QoE evaluation is shown in Figure 2. Similar to the audio interface, users can watch the video clips repeatedly until they can determine the quality difference between the two states, and then vote by pressing the LEFT arrow key or RIGHT arrow key accordingly.

Since the participants are allowed to switch back and forth between two stimuli at any time, one may voice the concern that such a user interface may cause the so-called temporal masking effect, i.e., the visibility (or audibility) of one stimulus being temporally reduced by the presence of another stimulus. According to [25, 53], a visual temporal masking effect could be present only in the first 30–100 ms after a scene change, while the auditory temporal masking effect could be present only in the first 100–200 ms [8]. Such a short duration of time of the temporal masking effect indicates that the effect has minimal, if any, impact on quality assessment because users tend to spend much longer time, say, 2 to 10 seconds according to our user study, on a stimulus after each switching. Another potential concern is that participants may overlook quality degradations of a stimulus as they may switch to another stimulus right before artifacts occur. Even though this could happen at times, we consider that it would not significantly impact the results of quality assessment because the chance to overlook the quality degradations of one stimulus is equal to that of another stimulus. It is thus unlikely that participants overlook all the quality degradations of a "lucky" stimulus while not overlooking any of the quality degradations of the other "unlucky" stimulus. If the participants watch or listen to each stimulus for a sufficient time before making judgements, the quality degradation overlook effect would be eliminated.

Note that we do not fix the user interface component in the proposed framework, i.e., how stimuli should be presented to participants and how participants should report their judgements. We believe that the user interface presented in Figure 1 and Figure 2 is suitable for audio and video QoE assessments, and for this reason adopt it in our case studies. However, researchers are free to develop any user interface suitable to their research, e.g., watching two video clips side-by-side synchronously, while still being able to employ the proposed framework for QoE assessments without any modification.

B. Overall Consistency Checks

After collecting inputs performed by a number of participants, we can assess the overall consistency of judgements among different experiments and participants by checking the stochastic transitivity properties [67] or computing Kendall's u-coefficient [42]. The stochastic transitivity approach involves checking three variants of the stochastic transitivity (ST) property, namely the weak (WST), moderate (MST), and strong (SST) stochastic transitivity [67]. Let $\hat{P}_{ij}$ be the empirical probability that the quality level T_i is considered better than the quality level T_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, where $i \neq j \neq k$ and $1 \leq i, j, k \leq n$. 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 thus help decide whether probabilistic choice modeling is suitable for analyzing the choice frequencies.

Kendall's u-coefficient [42] is defined as follows:

$$u = \frac{2 \sum_{i \neq j} \binom{a_{ij}}{2}}{\binom{m}{2}\binom{n}{2}} - 1.$$

If m 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 (cf. Section III-C), 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 agreement equals $-1/(m-1)$ if m is even, and $-1/m$ if m is odd.

C. QoE Score Estimation

If the consistency of the collected inputs is confirmed, we can proceed to infer the QoE scores for the quality levels being evaluated. The n quality levels in experiments are denoted as $T_1, ..., T_n$ and the number of comparisons for the pair (T_i, T_j) is denoted 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. Note that $a_{ij} + a_{ji} = n_{ij}$.

TABLE I
A MATRIX OF CHOICE FREQUENCIES FOR FOUR QUALITY LEVELS.

	T_1	T_2	T_3	T_4
T_1	$-$	a_{12}	a_{13}	a_{14}
T_2	a_{21}	$-$	a_{23}	a_{24}
T_3	a_{31}	a_{32}	$-$	a_{34}
T_4	a_{41}	a_{42}	a_{43}	$-$

By applying a probabilistic choice model [17] 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, 50], 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:

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

where $\pi(T_i)$ represents the relative preference probability for T_i, which can be obtained by maximum likelihood estimation, and $u(T_i) = \log \pi(T_i)$ is the estimated QoE score of the quality level T_i. Note that $\pi(T_i) \geq 0, i = 1, ..., n$ and $\sum_{i=1}^{n} \pi(T_i) = 1$. We treat $u(T_i)$ rather than $\pi(T_i)$ as the QoE score because it has the properties required by interval-scale metrics, whereas $\pi(T_i)$ does not. For example, if the probability of choosing T_i over T_j is equal to that of choosing T_j over T_k, that is, $P_{ij} = P_{jk}$, then the difference in QoE scores between T_i and T_j is equal to that between T_j and T_k, i.e., $u(T_i) - u(T_j) = u(T_j) - u(T_k)$.

To evaluate the BTL model's goodness of fit with the choice frequencies, we compare the likelihood L of the given model and the likelihood L_U of the unrestricted model, which fits the frequencies perfectly. The test statistic $-2\log(L/L_U)$ is approximately χ^2-distributed with $n-1$ degrees of freedom.

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 can normalize all the QoE scores between 0 and 1. By so doing, the quality level with the highest QoE always 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 this 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.

IV. CHEAT DETECTION MECHANISM

Since our framework aims to facilitate the crowdsourcing strategy for multimedia QoE assessment, an inevitable issue involves the possibility that participants may provide erroneous inputs which would cause inaccuracy and bias in the estimated QoE scores. As such, a cheat detection mechanism is required to ensure reliable QoE assessment. In this section, we introduce a metric for quantifying the input consistency of individual participants, describe the derivation of a threshold for detecting untrustworthy inputs, and conclude this section with a discussion on rewards and punishments.

A. *Transitivity Satisfaction Rate*

In an experiment with n stimuli, we collect $m = \binom{n}{2}$ paired comparison inputs. The participant's preferences in the m rounds of comparisons are supposed to have *transitive relation* in that if he prefers A to B as well as B to C, then he would also prefer A to C. Based on this transitivity property, we define a metric called the Transitivity Satisfaction Rate (TSR) to quantify the individual consistency of a participant's judgements in an experiment. The TSR is computed as the number of triplets satisfying the transitivity property divided by the number of triplets that the transitivity rule may apply to; thus, the value of the TSR must be between 0 and 1. The algorithm for computing the TSR is shown in Algorithm 1. The TSR will be 1 if and only if a participant's judgements are consistent throughout all the rounds in an experiment.

Algorithm 1 TSR Calculation

```
$m$ is an $n$ by $n$ matrix, where $m[i,j] = 1$ indicates that $i$ is considered better than $j$; otherwise $m[i,j] = 0$.

$n\_test \leftarrow 0$
$n\_pass \leftarrow 0$
for all $i, j, k, 1 \leq i, j, k \leq n, i \neq j \neq k$ do
    if $m[i,j] = 1$ and $m[j,k] = 1$ then
        $n\_test \leftarrow n\_test + 1$
        if $m[i,k] = 1$ then
            $n\_pass \leftarrow n\_pass + 1$
        end if
    end if
end for
$TSR \leftarrow n\_pass/n\_test$
```

B. *Trust Thresholding*

Having defined the TSR metric for quantifying the consistency of a participant's inputs in an experiment, we proceed to derive a TSR threshold to determine whether a participant's inputs in a particular experiment are trustworthy or not. We shall call the TSR value that can best discriminate trustworthy and untrustworthy experiments as the "trust threshold." In the following, we describe how we derive the trust threshold using an empirical approach.

1) Why Use Real Traces?: To derive the trust threshold, we need to obtain a set of trustworthy inputs and untrustworthy inputs and search for the TSR value which can best discriminate the two types of user inputs. However, obtaining such inputs is a challenging task because it is difficult, if not impossible, to define what trustworthy and untrustworthy inputs would look like. Trustworthy inputs are not perfectly correct inputs, as people make mistakes even though they may pay full attention to the QoE experiments. Also, not all the inputs from a "trustworthy" participant are necessarily "trustworthy;" this is because even if a participant is carefully making judgements, his decisions can still be erroneous due to distractions or limits in recognition capability. For the above reasons, we consider that real traces would best fit to our purpose as they contain real trustworthy and untrustworthy inputs, so the inferred trust threshold would be the most realistic.

2) Principles for Detecting Suspicious User Inputs: We use the user inputs from our case studies (presented in Section V) to serve as the real trace. Our next step is to distinguish trustworthy and untrustworthy inputs in the trace. For this purpose, we identify suspicious inputs by devising two heuristic rules: the distinct pairs heuristic and the decision time heuristic.

2.A) **The distinct pairs heuristic.** The authors of [26] indicated that in a paired comparison, the greater the difference two stimuli have, the less likely they will be judged incorrectly. By this intuition, we believe that distinct pairs, i.e., *stimuli pairs with a significant quality difference*, tend to receive consistent judgements across participants if the participants are competent and careful. On the other hand, dishonest or indiscreet participants who choose without careful examination or make random choices would not give judgements that are consistent with others'. Therefore, *if the comparative judgements on distinct pairs in an experiment are not consistent with those from other experiments, we deem the inputs from the experiment suspicious*. In our case, we use the top five distinct pairs (out

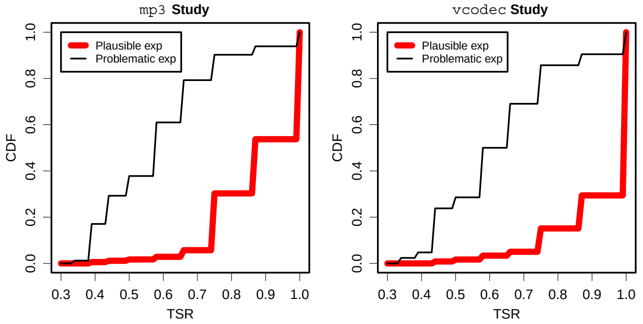


Fig. 3. The cumulative distribution functions of TSRs from plausible and problematic experiments (classified using the distinct pairs heuristic).

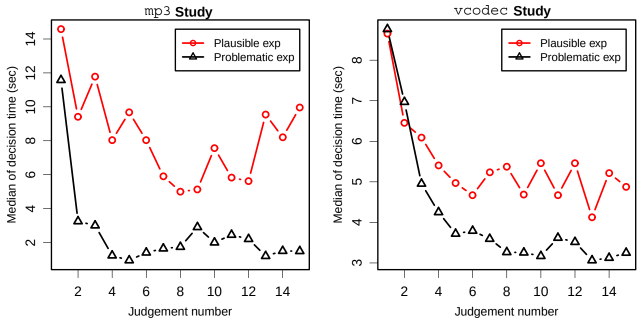


Fig. 4. The median decision time of the ith judgement in the plausible and problematic experiments, respectively.

of 15 possible pairs) and consider an experiment to be problematic if *any of* the distinct pairs are inconsistent with the consensus, i.e., the QoE scores estimated based on all user inputs using the BTL model (Section III-C).

2.B) **The decision time heuristic.** One problem of the distinct pairs heuristic is that it cannot handle problematic inputs which comprise correct judgements for pairs easier to distinguish (i.e., distinct pairs) and incorrect judgements for the remaining pairs. Such inputs may be given by incompetent participants who cannot recognize subtle differences in content quality or dishonest participants who only make easy judgements faithfully but cheat on difficult judgements by giving a random answer. In view of this reason, we devise the decision time heuristic. The intuition of the heuristic is that *the decision time of comparative judgements would be shorter in problematic experiments*, as dishonest and careless participants tend to be impatient for making careful judgements that they can otherwise make. In our case, we determine an experiment problematic if the decision time of its last two judgements (out of 15) is shorter than a certain threshold.

3) Suspicious User Inputs in Real Trace: The user inputs from our case studies came from 388 participants in $1,094$ experiments. Since the results from the four case studies are similar, we only present the results of the MP3 bitrate case study (Section V-A1) and the video codec study (Section V-B1) here, though all the case studies are included in the derivation for the trust threshold. For the sake of brevity, we denote the former study as the `mp3` study (262 experiments taken by 124 participants) and the latter as the `vcodec` study (300 experiments taken by 141 participants). In both studies, 6 stimuli are compared with each other, thus yielding a total of 15 paired comparisons in each experiment.

3.A) **Applying the distinct pairs heuristic.** By using the distinct pairs heuristic, we detected that 82 out of 262 experiments in the `mp3` study and 71 out of 300 experiments in the `vcodec` study are problematic. We plot the cumulative distribution function (CDF) of the TSRs of both plausible and problematic experiments in Figure 3, which shows that the TSRs of the two categories are clearly different. The Wilcoxon rank-sum test suggests that the TSRs of plausible and problematic experiments are significantly different (p-value smaller than 0.001), which indicates the effectiveness of the heuristic in detecting suspicious inputs.

3.B) **Applying the decision time heuristic.** We begin with an observation to check if any pattern in the judgement decision time exists. Figure 4 shows the median of decision time spent in the ith judgement in the `mp3` and `vcodec` studies respectively. We can make two observations according to the graph. The first is that the decision time of comparative judgements tends to decline over time, which echoes the earlier finding reported in [54]. This phenomenon should be due to the participants becoming more adept at quality comparison, and thus requiring less

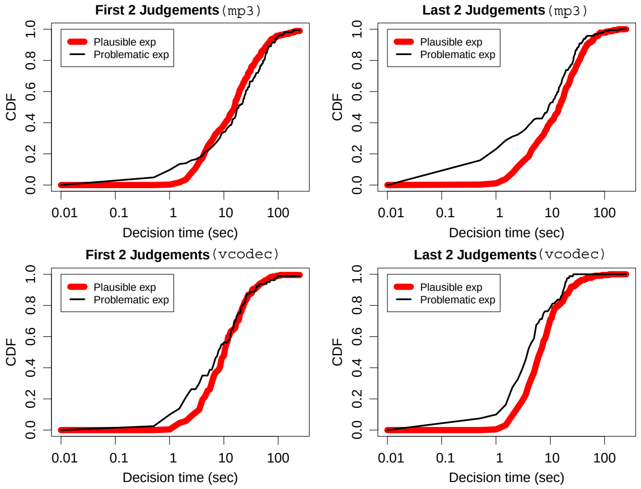


Fig. 5. The cumulative distribution functions of decision time of the first two and the last two judgements in our experiments.

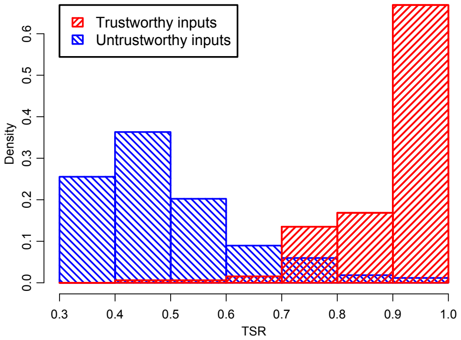


Fig. 6. The histograms of TSRs from the trustworthy and untrustworthy inputs (classified using both distinct pairs and decision time heuristics).

time to tell the subtle difference in each pair of stimuli. The second observation is that the decision time of the last few rounds in the problematic experiments is generally shorter than that in the plausible experiments. While the time for the former tends to be around 2–3 seconds, for the latter it is on average 5–10 seconds.

To quantify the difference of plausible and problematic experiments in terms of judgement decision time, we plot the cumulative distribution functions of decision time for the first two and the last two rounds in Figure 5. Using the Wilcoxon rank-sum test, the decision time of plausible experiments and that of problematic experiments in the first two rounds are statistically equivalent under the significance level 0.01 for both case studies. On the other hand, the decision time of both types of experiments in the last two rounds are significantly different with a p-value smaller than 0.001 in both studies. This confirms our intuition that the judgement time in the late rounds is an indicator whether the participants are serious about an experiment or not. As a consequence of this confirmatory finding, we adopt a binary classification approach to detect suspicious inputs based on the decision time of the last two comparisons in an experiment. The classifier is simple as it searches for a decision time threshold that yields the largest TPR$-$FPR, where TPR stands for the true positive rate and FPR stands for the false positive rate. The decision time threshold we obtained is 4.6 seconds and 5.2 seconds for the mp3 and vcodec studies, respectively.

Note that even though we use the decision time heuristic to detect suspicious inputs for the purpose of trust threshold derivation, we still do not consider these heuristics to be reliable enough to be used *directly* in cheat detection. The reason for this is that we are aware that dishonest participants can deliberately lengthen the decision time without paying attention to the content quality if they know of this rule. In contrast, the TSR, i.e., the judgement consistency, is robust to countermeasures (cf. Section IV-C). Therefore, the heuristic is only used in deriving the trust threshold for TSR.

4) Trust Threshold Derivation: We partition each trace into trustworthy and untrustworthy experiments based on both of the previously-mentioned heuristics, such that an experiment is considered untrustworthy if it is detected by either heuristic rule. Figure 6 shows the histogram of TSRs for the inputs from trustworthy and untrustworthy

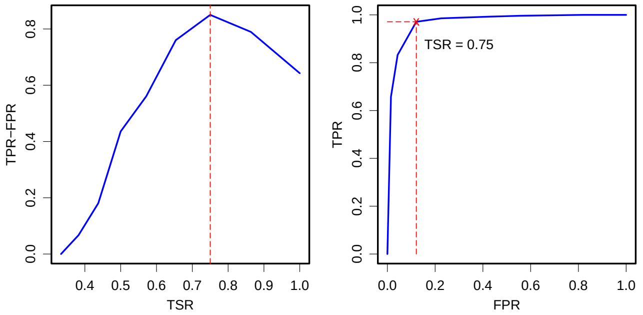


Fig. 7. Searching for the trust threshold that best discriminate trustworthy and untrustworthy inputs.

experiments in the four case studies. We can observe that the histogram is clearly bimodal: the TSRs of trustworthy inputs are mostly greater than 0.7, while those of untrustworthy inputs concentrate around 0.4–0.5.

We search for the best discrimination threshold by maximizing TPR$-$FPR over all the possible TSR thresholds. In Figure 7, the TPR$-$FPR over TSR and the ROC curve are plotted. From the left graph we can see that the most discriminatory TSR threshold is 0.75, which yields a TPR of 97% and a FPR of 14%. We therefore set the trust threshold to 0.75. In other words, we will consider an experiment is trustworthy if its TSR is higher than 0.75; otherwise, it is considered untrustworthy and will be discarded in QoE score estimation.

C. *Reward and Punishment*

For a crowdsourced experiment, we suggest announcing the logic of the TSR and punishment rules for participants who constantly produce TSR scores lower than the trust threshold. More specifically, a participant will only be paid a reward if the TSR score of his experiments is higher than 0.75. We enforced such policy in our case studies, and did not receive any complaints about the rule. We maintain that certain rules for rewards and/or punishments are required to ensure that resources are not wasted on untrustworthy inputs. Such rules will also maintain the quality of the QoE assessment, as untrustworthy inputs are excluded at the outset.

We believe that there is no systematic way for participants to cheat our system by submitting “trick” answers to achieve high TSR scores. First, the presentation order of each pair and the order within each pair (i.e., which clip corresponds to **Pressed** state or **Released** state) are totally random in each experiment, and 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 between the states and make judgements that are as consistent as possible. Second, although a malicious participant can still achieve a high TSR by making consistently “wrong” judgements (i.e., by always claiming that the state with the lower quality is the better one), such extreme cases can be detected easily by applying the distinct pairs heuristic rule.

V. Framework Evaluation

In this section, we present four case studies based on our framework for audio and visual QoE evaluations. The experiments were conducted by participants from the three sources of laboratory, MTurk, and community.

- **Laboratory**: We recruited part-time workers at an hourly rate of US\$8. They were asked to participate in the experiments in our laboratory. Each participant was asked to take a 5-minute break every 30 minutes.
- **MTurk**: We posted each experiment as a HIT (Human Intelligence Task) on the Mechanical Turk platform. If the outcome of an experiment was qualified, i.e., it yielded a TSR higher than 0.75, we paid the participant 0.15 US dollars.
- **Community**: We posted an open call on the website of an Internet community `telnet://ptt.cc` with 1.5 million members to seek participants for our experiments. For each qualified experiment, we paid the participant an amount of virtual currency that was equivalent to one US cent.

In the following, we first present the experiment setup of four case studies and their evaluation results, which were inferred from the combined inputs from the three sources. We then compare the performance of the laboratory and crowdsourced experiments from the perspectives of the data quality, participant diversity, and monetary cost.

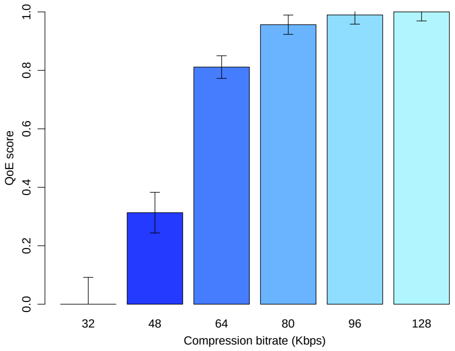


Fig. 8. QoE scores of MP3-compressed songs at different bitrates.

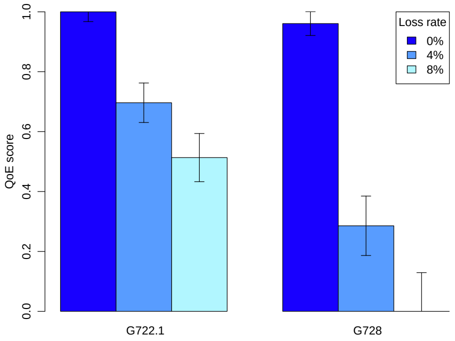


Fig. 9. QoE scores of VoIP speeches encoded by different codecs at different packet loss rates.

A. Audio QoE Evaluation

1) Effect of MP3 Compression Level: In this case study, we investigated the QoE of MP3-compressed audio clips with different compression levels. We selected an English song named "Garden of Graves" as the source clip. To obtain test clips, we converted the song into MP3 constant-bit-rate format with six bitrate levels, namely, 32, 48, 64, 80, 96, and 128 Kbps. Consequently, we had six test clips with different quality levels and required $\binom{6}{2} = 15$ paired comparisons in each experiment.

There were a total of 124 participants who performed 262 experiments. For each paired comparison, the participant used the interface shown in Figure 1 to indicate which quality level yielded a better listening experience. Based on our cheat detection mechanism (cf. Section IV), 128 experiments achieved TSRs lower than the trust threshold 0.75, and were therefore excluded from the QoE estimation. Using the BTL model described in Section III-C, we estimated the QoE scores of the six compression levels and plotted them in Figure 8, where the vertical bar denotes the 95% confidence interval of the score. It can be observed that a higher bitrate constantly leads to a higher QoE score. Also, the law of diminishing marginal utility was in evidence insofar as the increased rate of audio quality declines when the compression bitrate is high. This phenomenon is often seen in the relationship between a system's quality and users' perception because there must be an upper limit, beyond which increasing the system quality will not enhance the users' experience any further.

2) Effect of Packet Loss on VoIP QoE: This case study investigates the effect of the packet loss rate on VoIP speech quality. The source clip was a three-minute speech recording made by concatenating uncompressed speech segments from the Open Speech Repository[5]. We compressed the clip with two speech codecs, G722.1 and G728, respectively, into voice packets. Then, we simulated packet loss events in a Gilbert-Elliott channel [22, 24], where the loss rates were set at 0%, 4%, and 8%. Because of the combination of two speech codecs and three loss rates, six test clips were generated.

A total of 66 participants performed 135 experiments, among which 35 experiments were excluded based on the trust threshold. Figure 9 shows the estimated QoE scores for the six test clips. From the graph, it is evident that a higher packet loss rate leads to a lower QoE score. While G722.1 and G728 achieve similar QoE scores when the loss rate is zero, their robustness to packet loss is significantly different. Specifically, the QoE of G722.1 at the

[5] http://www.voiptroubleshooter.com/open_speech

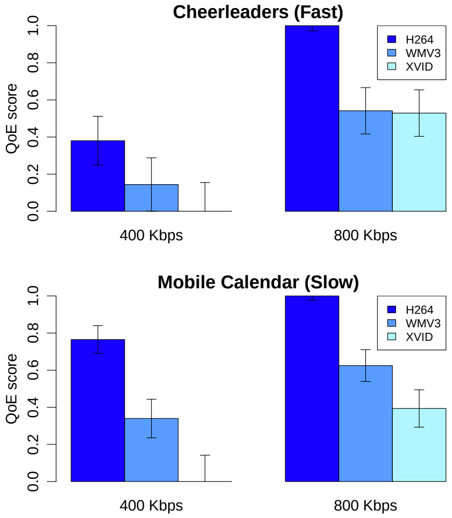


Fig. 10. QoE scores of video clips compressed by different codecs at different bitrates.

8% loss rate is much better than that of G728 at the 4% loss rate. The result conforms to our expectation because G722.1 operates at 32 Kbps, while G728 operates at 16 Kbps. As a result, G722.1 can use a higher encoding bitrate and a higher sampling rate than G728, and thus is more capable of maintaining speech information under a lossy situation.

B. *Visual QoE Evaluation*

1) Comparison of Video Codecs: In video compression, the codec and the compression level both play important roles in users' perceptions of video quality. Here, we assess the impact of codecs and compression levels on the QoE of video clips. From the database provided by the Video Quality Experts Group[6], we selected two 12-second raw video clips, the fast-motion "Cheerleaders" that contains a cheerleader dance performance, and the slow-motion "Mobile Calendar" that consists of a wall calendar, a toy train, and a ball, all moving slowly. We compressed both source clips with three codecs, H.264, WMV3, and XVID, at the two bitrates of 400 Kbps and 800 Kbps. In other words, for each source clip, we obtained six test clips with different codec-and-bitrate combinations.

A total of 141 participants, both part-time employees and Internet volunteers, performed 300 experiments, of which 27% of the experiments are excluded based on the cheat detection mechanism. Figure 10 shows the QoE scores of each test clip for the "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. For the "Cheerleaders" video, we find that H.264 performs better than WMV3 and XVID. While WMV3 is better than XVID at 400 Kbps, the quality of both these codecs is comparable at 800 Kbps. Interestingly, on the "Mobile Calendar" video, WMV3 performs significantly better than XVID at the same bitrate. This indicates that WMV3 is generally better than XVID. Also for the "Mobile Calendar" video, H.264 at 400 Kbps performs even better than WMV3 and XVID at 800 Kbps. This surprising result indicates that by using H.264, we can compress slow-motion videos at a low bitrate and have better perceptual quality than that of using WMV3 and XVID at a high bitrate. The study demonstrates that H.264 significantly outperforms the other two codecs at the same compression level, while XVID yields the worst overall ratings.

2) Comparison of Loss Concealment Schemes: In the design of a high-quality IPTV system, one of the most challenging issues is how to deal with video packet loss caused by network loss or excessive variations in network delay. A large number of loss concealment schemes have been proposed, such as the intuitive frame copy method and the more sophisticated error resilient coding approach. In this case study, we evaluated the two loss concealment schemes of the frame copy (FC) scheme and the frame copy with frame skip (FCFS) scheme [68], 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. In contrast, FCFS is a hybrid scheme that integrates

[6] `http://www.its.bldrdoc.gov/vqeg`

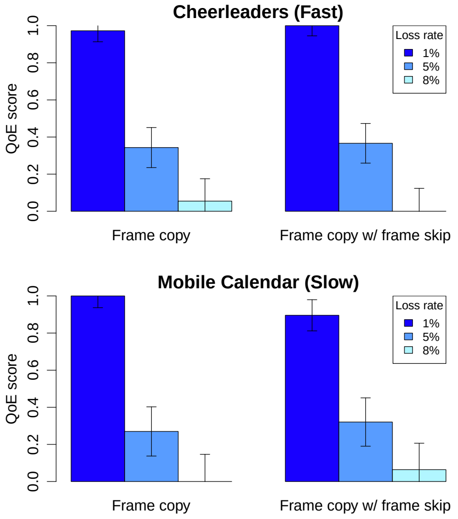


Fig. 11. QoE scores of video clips decoded with different loss concealment schemes at different packet loss rates.

the frame copy and the frame skip technique that 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%, the frame will be dropped; otherwise the errors in the frame will be concealed by the frame copy method. We continued to use the "Cheerleaders" and "Mobile Calendar" video clips in this study. We compressed both clips by using JM H.264/AVC reference software[7], and simulated packet loss rates at 1%, 5%, and 8%, to obtain degraded test clips. During the decoding process, we applied FC and FCFS for loss concealment. Since there are three packet loss rates and two loss concealment schemes, we obtained six test clips.

Overall, a total of 173 participants performed 397 experiments, of which 54% are considered untrustworthy. A look at Figure 11, showing the QoE score for each test clip of "Cheerleaders" and "Mobile Calendar," indicates that FCFS performs slightly better than FC on "Cheerleaders" when the loss rate is moderate ($\leq 5\%$). A possible explanation is that FCFS skips seriously corrupted frames so that the participants perceive better spatial quality. However, when the loss rate is high (8%), a large number of frames are seriously corrupted and thus dropped by FCFS, thus making the QoE of FCFS 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 ($\geq 5\%$). We believe that this is because the dropping of 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 given that 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 that the effect of loss concealment largely depends on the joint characteristics of the target video clips and network conditions.

C. Cost and Performance Analysis

As our case studies were conducted using both the laboratory and crowdsourcing strategies, we are able to inspect how much cost the crowdsourcing strategy saved and evaluate its performance. In this subsection, we present an analysis comparing both approaches in terms of economic cost, outcome quality, and participant diversity.

Economic cost: In total we spent US$$191.8$ on $1,094$ experiments, which were performed by 388 participants and involved $16,410$ rounds of paired comparison. The cost and performance of all the participant sources in the case studies are summarized in Table II. The laboratory experiments accounted for 89% of the total monetary cost. Since the number of experiments performed by participants from each source was different, the economic cost of each source is compared in terms of wage per round, as shown in Table II. 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 4.6 cents; and for the crowdsourced MTurk and community experiments it was 1 and 0.07 cents respectively, which yields respective ratios with the laboratory experiments of $4.6 : 1$ and $66 : 1$.

[7] http://iphome.hhi.de/suehring/tml

TABLE II
A COMPARISON OF LABORATORY AND CROWDSOURCED EXPERIMENTS IN TERMS OF THEIR COST AND PERFORMANCE.

Case Study	Participant Source	Total Cost	# Rounds	# Person	Qualified Rate	Cost / Round (cent)	Avg. TSR	WST Violation	MST Violation	SST Violation	Kendall
MP3 Bitrate	Laboratory	\$46.47	1,200	13	59%	6.59	0.95	0	0	0.00	0.62
	MTurk	\$4.20	1,035	26	41%	1.00	0.95	0	0	0.00	0.61
	Community	\$0.62	1,695	85	52%	0.07	0.96	0	0	0.25	0.60
VoIP Quality	Laboratory	\$30.45	990	10	67%	4.61	0.98	0	0	0.05	0.78
	MTurk	\$2.85	390	17	73%	1.00	0.98	0	0	0.10	0.78
	Community	\$0.39	645	39	86%	0.07	0.98	0	0	0.15	0.80
Video Codec	Laboratory	\$28.19	1,860	10	80%	1.90	0.98	0	0	0.15	0.57
	MTurk	\$4.80	750	28	64%	1.00	0.98	0	0	0.35	0.57
	Community	\$0.93	1,890	103	71%	0.07	0.97	0	0	0.30	0.57
Loss Concealment	Laboratory	\$66.59	1,800	13	69%	5.35	0.96	0	0	0.30	0.60
	MTurk	\$5.70	1,620	38	35%	1.00	0.97	0	0	0.20	0.60
	Community	\$0.62	2,535	122	35%	0.07	0.96	0	0	0.30	0.58
Overall		\$191.80	16,410	388	58%	2.02	0.97	0	0	0.18	0.64

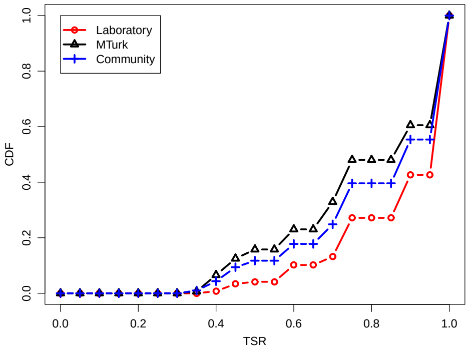


Fig. 12. The cumulative distribution functions of TSRs from the laboratory, MTurk, and community experiments.

Outcome quality: Figure 12 plots the cumulative distribution functions of TSRs for the experiments from three participant sources. We observed that nearly 70% of laboratory experiments were qualified, having a TSR higher than the trust threshold 0.75 (cf. Section IV), while only 46% and 54% of MTurk and community experiments were qualified. Although the proportion of suspicious inputs is larger with the crowdsourcing strategy, we can remove such inputs to ensure the accuracy of estimated QoE scores. In addition, it is a common practice that no wages are paid if a worker's output quality does not meet a certain standard in a crowdsourcing task, so we can assume that no wage needed to be paid for the experiments with TSRs lower than the trust threshold. In other words, our framework makes crowdsourced QoE assessment studies immune to dishonest participants because the latter will neither affect the quality of QoE assessment nor the economic cost.

We define the Qualified Rate as the ratio of experiments that yield a TSR higher than the trust threshold. As listed in Table II, the laboratory experiments achieved the highest qualified rates in all cases, except for the VoIP case study. Moreover, in the study of loss concealment schemes, the rates of laboratory experiments were as high as 69%, compared to approximately 35% from both crowdsourcing experiments. The low rate of 35% suggests that it is hard to differentiate the quality of video clips with different loss concealment schemes. It is our view that the superiority of laboratory experiments in this case is mainly due to the relatively higher proficiency of the participants. On average, the laboratory participants and the crowdsourcing participants respectively performed 138 and 26 comparisons in the study. Hence, the former had more opportunities to gain experience in distinguishing the subtle differences in the quality of video clips. After removing unqualified experiments, we computed the average TSRs of the experiments from the three sources and found that they were all above 0.95.

We also checked the overall consistency (cf. Section III-B) of paired comparison results from three participant sources. The stochastic transitivity checks reveal that no WST and MST violations occurred in any of the datasets. SST violations were observed in all case studies, but the numbers of violations were moderate. It is worth noting that the laboratory experiments had the fewest SST violations. This is reasonable because stochastic transitivity checks assess the consistency of judgements among different participants and the laboratory experiments involved the fewest participants. In all the case studies, the Kendall's u-coefficients were higher than 0.5, which indicates that the judgements provided by all three sources were reasonably consistent.

Participant diversity: Like many other user studies, the diversity of participants is crucial to QoE assessment studies. Since the purpose of these types of studies is to understand people's perception of certain multimedia content, a more diverse set of experiment participants enables us to collect a broader range of opinions. From this perspective, crowdsourcing is a clearly more appropriate and efficient strategy for QoE assessment because it substantially diversifies the participant pool. Quantitatively, the crowdsourced experiments accounted for only 11% of the total cost, but they accounted for 372 out of the 388 participants (96%) in our case studies.

VI. Paired Comparison and MOS: A Side-by-Side Comparison

Thus far, we have presented a framework that utilizes paired comparison in a crowdsourcing context for quantifying the QoE of multimedia content. Our case studies indicate that the framework yields reliable QoE estimates at a lower cost compared with the traditional in-laboratory MOS experiments. Nevertheless, as some researchers adopt the MOS methodology, and others adopt paired comparison for the same purpose—QoE estimation, one may wonder: *are the QoE estimates from the two methodologies compatible and mutually consistent*? It would be meaningless to compare QoE estimates from different methodologies if the methodologies do not generate comparable results.

In this section, motivated by these comparability concerns, we present a comparative analysis of the two widely-used QoE assessment methodologies, the mean opinion score (MOS) and the paired comparison (PC). Although we have demonstrated the qualitative advantages of PC over MOS, most notably in its capability to validate a single participant's inputs (Section IV), the compatibility of their results and a comparison of their quantitative characteristics, such as reliability and efficiency, have not yet been discussed. In the following, we first investigate the rating consistency between PC and MOS, and then compare the two methodologies in terms of three important properties, namely, their intra-subject reliability, inter-subject reliability, and how fast the QoE estimates converge.

A. Experiment Design

To facilitate a fair comparison between PC and MOS, we re-executed the `vcodec` study (cf. Section V-B1) solely in the laboratory by hiring another group of 13 subjects. The 13 subjects participated in 45 experiments in total using both the PC and MOS methodologies.

The setup of the PC experiment was exactly identical with that described in Section III-A. For the MOS experiment, we started with a *preview phase*, in which participants were asked to preview the test clips one by one and that would be rated later. The purpose of this phase is to provide participants an overall impression of the quality of the test clips. In the *rating phase*, followed immediately after the preview phase, participants were asked to watch each test clip and rate its quality level using the MOS scale, which consists of the five levels of Bad (1), Poor (2), Fair (3), Good (4), and Excellent (5). The presentation order of the test clips was independently randomized in both the preview and rating phases. The order of the PC and MOS experiments for each participant was also randomized with a five-minute break to prevent any systematic bias.

B. Consistency between PC and MOS

We now investigate whether the estimated QoE scores from the PC experiments are consistent with the MOS scores obtained by averaging the opinion scores (1–5) from the MOS experiments. For the sake of brevity, we shall denote the QoE scores estimated using the BTL model based on comparative judgements as "PC scores." Since the scales of both scores are not identical, when comparing two sets of scores, we normalize them by aligning the minimum and maximum of PC scores to those of MOS scores. In Figure 13, we plot the PC and MOS scores for each test clip, where the vertical bars in the graph denote the 95% confidence interval of the estimated scores. We

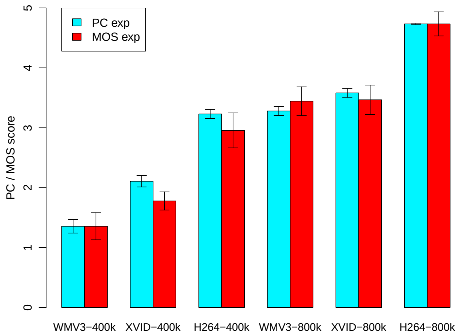


Fig. 13. A comparison of QoE scores estimated using PC and MOS methodologies.

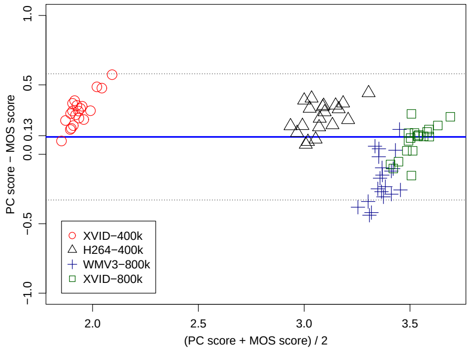


Fig. 14. Tukey mean-difference plot for inspecting systematic differences (if any) between PC and MOS scores.

can see from the graph that, the PC scores and MOS scores are generally consistent, with their 95% confidence intervals mutually overlapping. The Pearson's correlation coefficient between the PC and MOS scores is 0.99, which indicates a significantly strong consistency between the two methodologies.

We further examine whether any systematic difference between the PC and MOS scores exists using the Tukey mean-difference plot (or called the Bland-Altman plot) [4], as shown in Figure 14. We inspect only the four (out of six) quality levels, because given that the minimum and maximum scores from PC and MOS are aligned, their differences are always zero. To reduce the effect of anomalous inputs (if any), we adopt a re-sampling approach and re-estimate the PC and MOS scores based on a random 80% subset of the raw inputs for 20 times. For this reason, there are a total of 80 points (20 points for each of the 4 quality levels) in Figure 14. The blue line on the graph stands for the mean difference of the scores, while the black lines define the 95% confidence interval of the mean difference. The graph indicates that though differences exist between the estimated PC and MOS scores, the differences are mostly smaller than 0.5. It is also the case that there is no systematic bias as the magnitude and polarity of the differences are not dependent on the scores, which further supports the outcome consistency between PC and MOS.

C. Intra-Subject Reliability

Ideally, a QoE assessment methodology should be able to solicit users' preferences *reliably* and *consistently* if the preference of certain stimuli is repeatedly inquired. For this reason, we evaluate which methodology, PC or MOS, leads to a higher intra-subject reliability.

We used the Intra-class Correlation Coefficient (ICC) [63] to quantify the agreement of repeated measures from a single participant. The ICC takes n repeated measures as the input and outputs a coefficient ranging from 0 to 1, where 0 indicates a complete lack of agreement and 1 indicates a perfect agreement between the repeated measures. We use the ranks of stimuli as the input measures. The ranks of the stimuli in the PC experiments are obtained by sorting the number of preferred votes for each stimulus, while those in MOS experiments are obtained by sorting the opinion scores.

Among the 13 participants we recruited, 7 of them participated in both PC and MOS experiments at least three times. The ICCs of the 7 participants are shown in Figure 15, from which we can see that, except for Bill, all the

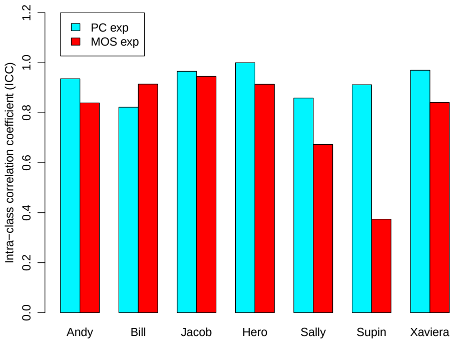


Fig. 15. A comparison of intra-subject reliability of PC and MOS methodologies.

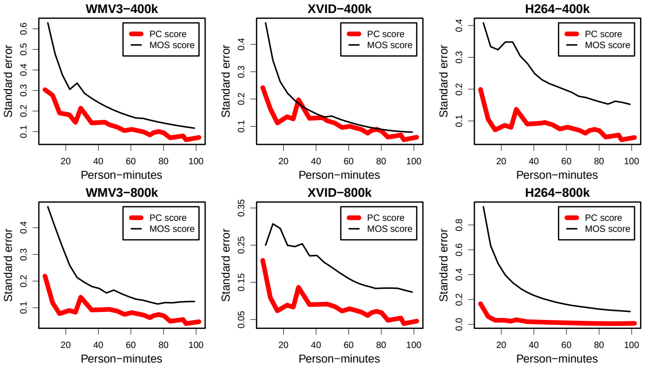


Fig. 16. A comparison of the convergence rate of the estimated QoE scores using PC and MOS methodologies.

participants achieved higher judgement consistency in the PC than in the MOS experiments. While the ICCs of all the 7 participants are higher than 0.8 in PC experiments, the ICCs of two participants are lower than 0.7 and the ICC of one participant, Supin, is even lower than 0.4 in MOS experiments. Generally speaking, PC solicited higher intra-subject reliability than MOS, which we affirm is due to the simpler and more intuitive *comparative preference judgements* in paired comparisons.

D. Inter-Subject Reliability

Similar to the intra-subject reliability consideration, we believe that a good QoE assessment methodology should also be able to solicit *similar or even identical opinions from different participants*, unless the quality levels of stimuli are contentious.

We use the ICC as well as the Kendall's W (as known as Kendall's coefficient of concordance) [41], to quantify the agreement between the opinions from different participants. The inputs to both ICC and Kendall's W are the ranks of stimuli according to each participant, and both measures output a statistic ranging from 0 to 1 with 1 denoting complete mutual agreement. For each participant, we first aggregate all his inputs from all the experiments he participated in. Then, the ranks of stimuli in the PC experiments are obtained by sorting the aggregated number of preferred votes for each stimulus, while those in the MOS experiments are obtained by sorting the averaged opinion scores. Further, as we had 13 participants in this experiment, the length of the input vector to ICC and Kendall's W is 13.

Based on our dataset, the inter-subject ICCs of PC and MOS are 0.85 and 0.72, respectively, while the Kendall's W of PC and MOS are 0.85 and 0.73, respectively. Both statistics suggest that PC tends to yield more consistent judgements across participants than MOS. We consider the difficulty in rating scale mapping and the rating scale heterogeneity issue (cf. Section I) of MOS to be the major causes of its relatively lower inter-subject reliability. Our results indicate that paired comparison constitutes a reliable QoE assessment methodology due to its simple and unambiguous comparative judgements.

E. Convergence Rate

Another important property of subjective QoE assessment methodologies is whether a methodology can provide confident QoE estimates using *as little human effort as possible*. Thus, we proceed to evaluate the reliability of estimated QoE scores using both PC and MOS assuming a certain amount of human effort is invested. We used the standard error of the estimated QoE scores to represent the score reliability, and used the total person-minutes spent in experiments to measure the human effort. A comparison chart of the score reliability versus human effort is depicted in Figure 16. The graph shows that with similarly spent human effort, PC generally provides more reliable QoE estimates than MOS does. We see this being due to the higher intra-subject reliability and inter-subject reliability of PC, as detailed in the previous subsections.

In sum, by analyzing a side-by-side comparison of the PC and MOS QoE assessment methodologies, we have shown that PC yields similar results with MOS. Moreover, PC can be seen as performing even better in terms of the intra-subject reliability, inter-subject reliability, and outcome convergence rate owing to its simpler and less-challenging judgement subtasks that participants need to do.

VII. Discussion

In this section, we discuss a variety of issues related to the proposed framework. We start with a discussion of the scope of QoE supported by the framework, and then elaborate on several issues related to paired comparison and crowdsourcing.

A. QoP or QoE?

As defined in [23], the quality of perception (QoP) reflects "a user's detectability of a change in quality or the acceptability of a quality level," while the quality of experience (QoE) is primarily concerned with "the overall acceptability of an application or service as perceived subjectively by the end-user." Strictly speaking, the case studies presented in Section V focus more on the QoP aspect (e.g., user perceptions about certain quality levels of the MP3 compression algorithm) of the multimedia content. However, our framework can also be applied to evaluate a multimedia system's QoE with proper adjustments to the raters' user interface to facilitate the paired comparison of two stimuli. For example, Huang et al. [32] evaluated participants' experiences in a tele-immersive interactivity using paired comparison; Lan et al. [45] used paired comparison to evaluate the quality of interactive VoIP conversations; and Chang et al. [9] evaluated the impact of network delay, packet loss rate, and delay jitter on players' online gaming experiences. Whatever situations that paired comparison has been applied to, the proposed trusted framework is also applicable to for the detection of untrustworthy user inputs and for the ensuring of the quality of QoE estimates.

B. Issues with Paired Comparison

Having shown that paired comparison has several advantages over traditional methodologies, we acknowledge that paired comparison comes at a price. The main disadvantage of paired comparison is that the *judgements* required to evaluate the quality of n stimuli are in the order of O(n^2), in contrast to the order of O(n) with the MOS methodology. Fortunately, a number of studies have been devoted to resolve this complexity issue of paired comparisons. In [21], for example, Eichhorn et al. proposed a scheme in which each subject only needs to respond to a unique random subset of pairs instead of all the pairs. They demonstrated that their scheme can provide reliable QoE estimates based on users' judgements on merely 29% of all possible pairs. In addition, Xu et al. [72] further explored how to efficiently select pairs that require users' preference judgements the most. They proposed a random partial paired comparison approach based on random graph theory and Hodge theory, and showed that the complexity O(n^2) of "traditional" paired comparisons can be reduced to O($n^{1.5}$) without significantly sacrificing the accuracy of estimated quality.

Note that though the number of judgements of paired comparisons is larger than that of MOS, the effort required by each judgement is not identical. While PC requires a dichotomous decision in each judgement, MOS requires a multi-choice decision, which inevitably incurs longer thinking time and higher mental overhead. That is one of the reasons why PC can outperform MOS with the same amount of person-minutes invested (cf. Figure 16). Given all these reasons, we believe that the numerous benefits of paired comparison, especially when n is not large and combined with crowdsourcing, can compensate for the relatively more judgements the methodology requires.

Tie Handling. From time to time there could be situations where the two stimuli in a paired comparison exhibit similar quality levels, and subjects may not know how to judge which stimulus is better. To cope with such tie situations, Rao and Kupper [58] proposed to generalize paired comparison to include the handling of tie situations so that a participant can report the quality levels of two stimuli as “equivalent” if he considers that both stimuli have similar quality levels. An example of such extended use of paired comparison is [47], where Lee et al. evaluated the QoE of 3D images using the ternary-response paired comparison methodology.

C. Issues with Crowdsourcing

While our case studies indicate that crowdsourcing is a proper way for large-scale and lower-cost QoE evaluation, we note that the strategy has several limitations that may affect its applicability in certain scenarios.

Environment control. In crowdsourced experiments, participants may view media content under various conditions, such as lighting, screen size, and the quality of headsets. In contrast, laboratory experiments are normally conducted in a controlled environment that equalizes experiment conditions. However, the crowdsourcing strategy can be considered as either advantageous or disadvantageous, depending on the viewpoint. On the one hand, it is an advantage because users’ perceptions can now be assessed in real-life scenarios, in which, for example, users’ headsets may be not as good as those in laboratories, and ambient sound may be unavoidable. While it is difficult to simulate and define a “typical” user environment in a laboratory, crowdsourcing allows us to assess how people actually experience in their daily life. On the other hand, it can be a disadvantage if the purpose of experiments is to measure the quality of multimedia content in a specific scenario.

Experiment devices. Since most people connect to the Internet with personal computers, the crowdsourcing strategy is most suitable for evaluating the media content on such platforms. That is, it could be a problem if evaluations are conducted on other input/output devices, such as HDTV or e-book devices. Fortunately, since these non-PC devices are gradually becoming Internet-capable, the problem may be resolved in the near future.

Demographic factors. The demographic make-up of participants is essential for certain types of QoE experiments. However, as we are unable to confidently identify each crowd worker, it is difficult for us to relate the assessment 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, we believe that it is sufficiently general for evaluating the QoE of multimedia content and systems for a variety of applications. It is especially helpful for assessing the effect of techniques related to coding, processing, and transmission of media content, as we have shown in the four case studies. Moreover, with the rapid advent of technologies for rich user interfaces, such as HTML 5, we expect the framework to be more convenient for assessing user experiences associated with interactive content like computer games [49].

VIII. Conclusion

In this paper, we have proposed a trusted crowdsourceable framework for assessing the QoE of multimedia content and systems. The rating procedure is simple for participants since they only need to make comparative judgements throughout experiments. The support for detecting participants’ problematic inputs is particularly essential since not all Internet crowd are trustworthy. Moreover, the cheat detection mechanism makes “differentiated rewards” possible so that participants can be paid according to the quality of their experiment inputs. By using our framework, researchers can outsource QoE evaluation experiments to a diverse labor pool without compromising the quality of the QoE assessment.

We have demonstrated the efficacy of our framework with four case studies that involve audio and visual QoE evaluations, and have shown that the monetary cost is relatively lower for the crowdsourcing strategy than in laboratory experiments. A comparison with the commonly-used MOS methodology reveals that our framework provides comparable QoE ratings while yielding high intra- and inter-subject consistency. In summary, we expect that the proposed crowdsourceable framework for QoE evaluations will be helpful to researchers in multimedia signal processing, content analysis, and system development.

Acknowledgements

This work was supported in part by the National Science Council under the grants NSC101-2221-E-001-012-MY3 and NSC101-2221-E-002-190-MY3.

References

[1] O. Alonso, D. E. Rose, and B. Stewart, "Crowdsourcing for relevance evaluation," *ACM SIGIR Forum*, vol. 42, no. 2, pp. 9–15, 2008.

[2] A. Basso and S. Sicco, "Preventing massive automated access to web resources," *Computers and Security*, vol. 28, no. 3, pp. 174–188, 2009.

[3] M. H. Birnbaum, *Psychological Experiments on the Internet*. Academic Press, 2000.

[4] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, no. 8476, pp. 307–310, 1986.

[5] D. C. Brabham, "Crowdsourcing as a model for problem solving: An introduction and cases," *Convergence: The International Journal of Research into New Media Technologies*, vol. 14, no. 1, pp. 75–90, 2008.

[6] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[7] S. Buchholz and J. Latorre, "Crowdsourcing preference tests, and how to detect cheating," in *Proceedings of Interspeech 2011*, August 2011, pp. 3053–3056.

[8] B. Carnero and A. Drygajlo, "Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms," *IEEE Transactions on Signal Processing*, vol. 47, no. 6, pp. 1622–1635, 1999.

[9] Y.-C. Chang, K.-T. Chen, C.-C. Wu, C.-J. Ho, and C.-L. Lei, "Online game QoE evaluation using paired comparisons," in *Proceedings of IEEE CQR 2010*, June 2010, pp. 1–6.

[10] K.-T. Chen, C.-Y. Huang, P. Huang, and C.-L. Lei, "Quantifying Skype user satisfaction," in *Proceedings of ACM SIGCOMM 2006*, Pisa, Italy, Sep 2006.

[11] K.-T. Chen, J.-W. Jiang, P. Huang, H.-H. Chu, C.-L. Lei, and W.-C. Chen, "Identifying MMORPG bots: A traffic analysis approach," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, 2009.

[12] K.-T. Chen, C. C. Tu, and W.-C. Xiao, "OneClick: A framework for measuring network quality of experience," in *Proceedings of IEEE INFOCOM 2009*, April 2009.

[13] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *Proceedings of ACM Multimedia 2009*, 2009.

[14] ——, "Quantifying QoS requirements of network services: A cheat-proof framework," in *Proceedings of the second annual ACM Conference on Multimedia systems*, February 2011, pp. 81–92.

[15] S. Choisel and F. Wickelmaier, "Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference," *The Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 388–400, 2007.

[16] F. M. Ciaramello and S. S. Hemami, "Quality versus intelligibility: Studying human preferences for American sign language video," in *Proceedings of SPIE Vol. 7865, Human Vision and Electronic Imaging XVI*, January 2011.

[17] H. A. David, *The Method of Paired Comparisons*, 1988.

[18] L. Ding and R. Goubran, "Speech quality prediction in VoIP using the extended E-Model," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM 2003)*, December 2003, pp. 3974–3978.

[19] R. Dittrich, R. Hatzinger, and W. Katzenbeisser, "Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings," *Journal of the Royal Statistical Society (Series C): Applied Statistics*, vol. 47, no. 4, pp. 511–525, 1998.

[20] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the World-Wide Web," *Communications of the ACM*, vol. 54, no. 4, pp. 86–96, April 2011.

[21] A. Eichhorn, P. Ni, and R. Eg, "Randomised pair comparison: An economic and robust method for audiovisual quality assessment," in *Proceedings of International Workshop on Network and Operating Systems Support for Digital Audio and Video*, June 2010, pp. 63–68.

[22] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell System Technical Journal*, vol. 42, no. 5, pp. 1977–1997, 1963.

[23] ETSI TR 102 643, "Quality of experience (QoE) requirements for real-time communication services," 2010.

[24] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell System Technical Journal*, vol. 39, no. 5, pp. 1253–1265, 1960.

[25] B. Girod, "The information theoretical significance of spatial and temporal masking in video signals," in *Proceedings of SPIE Vol. 1077, Human Vision, Visual Processing, and Digital Displays*, 1989, pp. 178–187.

[26] R. J. Hill, "A note on inconsistency in paired comparison judgments," *American Sociological Review*, vol. 18, no. 5, pp. 564–566, 1953.

[27] C.-J. Ho, T.-H. Chang, and J. Y.-J. Hsu, "Photoslap: A multi-player online game for semantic annotation," in *Proceedings of the 22nd Conference on Artificial Intelligence*, July 2007, pp. 1359–1364.

[28] C.-J. Ho and K.-T. Chen, "On formal models for social verification," in *Proceedings of Human Computation Workshop 2009 (affiliated to ACM KDD 2009)*, June 2009, pp. 62–69.

[29] J. J. Horton and L. B. Chilton, "The labor economics of paid crowdsourcing," in *Proceedings of ACM Conference on Electronic Commerce*, June 2010, pp. 209–218.

[30] T. Hossfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via crowdsourcing," in *Proceedings of the 2011 IEEE International Symposium on Multimedia*, 2011, pp. 494–499.

[31] J. Howe, "The rise of crowdsourcing," *Wired Magazine*, vol. 14, no. 6, pp. 176–183, 2006.

[32] Z. Huang, A. Arefin, P. Agarwal, K. Nahrstedt, and W. Wu, "Towards the understanding of human perceptual quality in tele-immersive shared activity," in *Proceedings of the ACM Multimedia Systems Conference*, February 2012, pp. 29–34.

[33] Y. Ito and S. Tasaka, "Quantitative assessment of user-level QoS and its mapping," *IEEE Transactions on Multimedia*, vol. 7, no. 3, pp. 572–584, 2005.

[34] ITU-R Recommendation P.800, "Methods for subjective determination of transmission quality," 1996.

[35] ITU-T Recommendation G.107, "The E-model, a computational model for use in transmission planning," 2005.

[36] ITU-T Recommendation J.247, "Objective perceptual multimedia video quality measurement in the presence of a full reference," 2008.

[37] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[38] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," 2008.

[39] R. Jain, "Quality of experience," *IEEE Multimedia*, vol. 11, no. 1, pp. 96–97, 2004.

[40] S. Jain, Y. Chen, and D. C. Parkes, "Designing incentives for online question and answers forums," in *Proceedings of ACM Conference on Electronic Commerce*, July 2009, pp. 129–138.

[41] M. G. Kendall and B. B. Smith, "The problem of m rankings," *The Annals of Mathematical Statistics*, vol. 10, no. 3, pp. 275–287, 1939.

[42] ——, "On the method of paired comparisons," *Biometrika*, vol. 31, no. 3/4, pp. 324–345, 1940.

[43] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with Mechanical Turk," in *Proceedings of ACM CHI 2008*, April 2008, pp. 453–456.

[44] C. L. Knott and M. S. James, "An alternate approach to developing a total celebrity endorser rating model using the analytic hierarchy process," *International Transactions in Operational Research*, vol. 11, no. 1, pp. 87–95, 2004.

[45] K.-C. Lan and T.-H. Wu, "Evaluating the perceived quality of infrastructure-less VoIP," in *Prceedings of IEEE Workshop on Streaming and Media Communications*, July 2011.

[46] C. T. Lee, E. M. Rodrigues, G. Kazai, N. Milic-Frayling, and A. Ignjatovic, "Model for voter scoring and best answer selection in community Q&A services," in *Proceedings of IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, September 2009, pp. 116–123.

[47] J.-S. Lee, L. Goldmann, and T. Ebrahimi, "A new analysis method for paired comparison and its application to 3D quality assessment," in *Proceedings of ACM Multimedia 2011*, 2011, pp. 1281–1284.

[48] J.-S. Lee, F. D. Simone, and T. Ebrahimi, "Subjective quality evaluation via paired comparison: Application to scalable video coding," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 882–893, 2011.

[49] Y.-T. Lee, K.-T. Chen, H.-I. Su, and C.-L. Lei, "Are all games equally cloud-gaming-friendly? An electromyographic approach," in *Proceedings of IEEE/ACM NetGames 2012*, October 2012.

[50] R. D. Luce, *Individual Choice Behavior: A Theoretical Analysis*, 1959.

[51] T. Manjunath, "Limitations of perceptual evaluation of speech quality on VoIP systems," in *Proceedings of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, May 2009, pp. 1–6.

[52] J. N. S. Matthews and K. P. Morris, "An application of Bradley-Terry-type models to the measurement of pain," *Journal of the Royal Statistical Society (Series C): Applied Statistics*, vol. 44, no. 2, pp. 243–255, 1995.

[53] R. R. Pastrana-Vidal, J. C. Gicquel, C. Colomes, and H. Cherifi, "Temporal masking effect on dropped frames at video scene cuts," in *Proceedings of SPIE Vol. 5292, Human Vision and Electronic Imaging IX*, 2004, pp. 194–201.

[54] G. L. Peterson and T. C. Brown, "Economic valuation by the method of paired comparison, with emphasis on evaluation of the transitivity axiom," *Land Economics*, vol. 74, no. 2, pp. 240–261, 1998.

[55] N. L. Powers and R. M. Pangborn, "Paired comparison and time-intensity measurements of the sensory properties of beverages and gelatins containing sucrose or synthetic sweeteners," *Journal of Food Science*, vol. 43, no. 1, pp. 41–46, 1978.

[56] G. Qiu, Y. Mei, and J. Duan, "Evaluating HDR photos using Web 2.0 technology," in *Proceedings of SPIE Vol. 7867, Image Quality and System Performance VIII*, 2011.

[57] D. G. Rand, A. Dreber, T. Ellingsen, D. Fudenberg, and M. A. Nowak, "Positive interactions promote public cooperation," *Science*, vol. 325, no. 5945, pp. 1272–1275, 2009.

[58] P. V. Rao and L. L. Kupper, "Ties in paired-comparison experiments: A generalization of the Bradley-Terry model," *Journal of the American Statistical Association*, vol. 62, no. 317, pp. 194–204, 1967.

[59] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *Proceedings of IEEE International Conference on Image Processing*, September 2011, pp. 3097–3100.

[60] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, "crowdMOS: An approach for crowdsourcing mean opinion score studies," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2011, pp.

2416–2419.

[61] P. E. Rossi, Z. Gilula, and G. M. Allenby, “Overcoming scale usage heterogeneity: A Bayesian hierarchical approach,” *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 20–31, 2001.

[62] T. L. Saaty, “A scaling method for priorities in hierarchical structures,” *Journal of Mathematical Psychology*, vol. 15, no. 3, pp. 234–281, 1977.

[63] P. E. Shrout and J. L. Fleiss, “Intraclass correlations: Uses in assessing rater reliability,” *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, 1979.

[64] A. Sorokin and D. Forsyth, “Utility data annotation with Amazon Mechanical Turk,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2008, pp. 1–8.

[65] I. Sprow, Z. Baranczuk, T. Stamm, and P. Zolliker, “Web-based psychometric evaluation of image quality,” in *Proceedings of SPIE Vol. 7242, Image Quality and System Performance VI*, 2009.

[66] L. Sun and E. C. Ifeachor, “Voice quality prediction models and their application in VoIP networks,” *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 809–820, 2006.

[67] P. Suppes, D. H. Krantz, R. D. Luce, and A. Tversky, *Foundations of measurement, Volume II: Geometrical, Threshold, and Probabilistic Representations*, December 2006.

[68] S. Tasaka, H. Yoshimi, A. Hirashima, and T. Nunome, “The effectiveness of a QoE-based video output scheme for audio-video IP transmission,” in *Proceeding of ACM Multimedia 2008*, 2008, pp. 259–268.

[69] A. Watson and M. A. Sasse, “Measuring perceived quality of speech and video in multimedia conferencing applications,” in *Proceedings of ACM Multimedia 1998*, 1998, pp. 55–60.

[70] P. Welinder and P. Perona, “Online crowdsourcing: Rating annotators and obtaining cost-effective labels,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2010, pp. 25–32.

[71] W. Wu, A. Arefin, G. Kurillo, P. Agarwal, K. Nahrstedt, and R. Bajcsy, “A psychophysical approach for real-time 3D video processing,” in *Proceedings of ACM Multimedia 2011*, November 2011, pp. 683–686.

[72] Q. Xu, Q. Huang, T. Jiang, B. Yan, W. Lin, and Y. Yao, “HodgeRank on random graphs for subjective video quality assessment,” *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 844–857, June 2012.