A Crowdsourceable QoE Evaluation Framework for Multimedia Content

Kuan-Ta Chen[12], Chen-Chi Wu[3], Yu-Chun Chang[3], and Chin-Laung Lei[3]

[1]Institute of Information Science, Academia Sinica
[2]Research Center for Information Technology Innovation, Academia Sinica
[3]Department of Electrical Engineering, National Taiwan University

ABSTRACT

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

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

Categories and Subject Descriptors

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

General Terms

Experimentation, Human Factors, Performance

Keywords

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

MM'09, October 19–24, 2009, Beijing, China.

1. INTRODUCTION

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

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

1. The rating standard is somewhat obscure to experiment participants. As the concepts of the five scales, i.e., Bad, Poor, Fair, Good, and Excellent, cannot be concretely defined and explained, subjects may be confused about which scale they should give in each test.
2. Participants may have dissimilar interpretations of the scales, e.g., Poor and Good. Thus, they may provide different ratings even if they have had similar experiences with the same stimulus. According to a user's own idiosyncratic preferences and strategy, he/she may give higher or lower scales compared to those of other participants. This is the so-called scale heterogeneity problem [31].
3. It is assumed that MOS scores are on an interval scale, but it is only an ordinal scale [36]. In fact, when people use the MOS scale, their cognitive distance between Bad (1) and Poor (2) is usually different from that between Good (4) and Excellent (5) [36]. Thus, the accuracy of MOS scores may be questionable, since they are obtained by taking arithmetic means, which does not apply to ordinal-scale measurements.

4. In MOS tests, participants are asked to grade the MOS scores for stimuli; however, we do not know whether they pay full attention to the scoring procedures, or whether they just give ratings in a perfunctory manner. To the best of our knowledge, there is no established methodology for verifying whether a participant gives false ratings either intentionally or carelessly. Consequently, it is hard to detect problematic inputs, and the measurement accuracy may be degraded due to the behavior of untrustworthy participants.

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

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

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

1. Like MOS, paired comparison is generalizable across a variety of multimedia applications. Thus, it can be applied to various genres of multimedia content without any modification.
2. The burden on participants is low, since they do not have to map their sensation magnitude on a categorical or numerical scale. They are only required to make simple *comparative judgments*, and thereby avoid the scale heterogeneity problem of MOS ratings [31].
3. We can apply probabilistic choice models [10] to analyze paired-comparison results and obtain QoE scores on an *interval scale* [36]. The scale enables us to quantify the discrepancy between the QoE of different evaluated targets, and also allows us to compile an arithmetically computable index for QoE management purposes [7, 17].
4. The key property of paired comparison is that *the experiment results can be verified*. The verification basically relies on *the transitivity property*; that is, if A is preferred over B, and B is preferred over C, then A should also be preferred over C by the same participant. By employing this property, we can detect inconsistent judgments and remove problematic data before performing further analysis and modeling.

We demonstrate the effectiveness and generalizability of the proposed crowdsourceable experiment framework in four case studies: the first evaluates the quality of audio clips encoded by MP3 with different bit rates; the second evaluates the quality of VoIP speech with different packet loss rates; the third compares the quality of several video codecs that have similar bandwidth usage; and the fourth compares two loss concealment schemes for video playout. The studies cover a variety of multimedia applications and various factors that may affect the content's quality. For each study, we conducted both laboratory and crowdsourced experiments. The former were performed by part-time employees under supervision; and the latter were carried out by anonymous Internet users who were interested in making some money. The results show that, overall, the quality of data obtained from the crowdsourced experiments was slightly lower than that derived from laboratory experiments. Even so, because of our approach's ability to detect inconsistent inputs[1], we can still obtain comparable evaluation results at a lower economic cost and with wider participant diversity. In addition, supervision of the experiment does not require a physical space or involve labor costs.

Our contribution in this work is three-fold:

1. We propose a crowdsourceable framework, which comprises paired comparison, consistency checking, probabilistic choice modeling, and Web-based implementations, to quantify the QoE of multimedia content. The advantages of our framework over traditional MOS ratings are that 1) it facilitates crowdsourcing because it supports systematic verification of the participants' inputs; 2) the rating procedure is simpler than that of the MOS test, so the burden on participants is lower; and 3) it derives interval-scale scores that enable further quantitative analysis and QoE management [17].
2. Our crowdsourceable framework not only enables detection of problematic inputs, but also makes "differentiated rewards" possible. That is, the reward for performing an experiment can be based on the quality

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

(i.e., consistency) of a participant's inputs. This design encourages participants to ensure that their judgments in experiments are consistent.

3. To demonstrate the efficacy of our framework, we conduct four case studies involving audio and visual multimedia content. The results of laboratory and crowdsourced experiments indicate that we can obtain comparable evaluation results at lower economic cost and with wider participant diversity.

Note that our framework is effective even if the experiments are not conducted by an anonymous Internet crowd. Its simple rating procedure and capability to provide differentiated rewards can reduce the burden on participants and encourage them to give quality inputs. Consequently, the framework can also be used to maintain the quality of evaluation results in traditional laboratory and focus-group studies.

The remainder of this paper is organized as follows. Section 2 contains a review of related works. We describe the proposed framework in Section 3, and, explain how to apply it to evaluate the QoE of audio and visual content in Section 4 and Section 5 respectively. We then discuss the effectiveness of the experiment crowdsourcing strategy in Section 6. Finally, in Section 7, we present our conclusions.

2. RELATED WORK

2.1 Quality of Experience Assessment

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

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

2.2 Paired Comparison

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

2.3 Crowdsourcing

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

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

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

2.4 Reward and Punishment Mechanisms

In crowdsourced user studies, it is important to provide proper incentives so that participants are motivated to give high quality answers. Reward and punishment mechanisms have been discussed extensively in works on peer production systems and reputation systems [30]. In peer production systems, such as Wikipedia, Yahoo! Answers, and Games with A Purpose (GWAP), users collaborate in the hope of achieving a global outcome. A number of studies have used game theoretic analysis to examine the rationality of incentives in human computation games [14, 15] and online Q&A forums [23]. In reputation systems, a user's reputation depends on the accumulated ratings given by other users on

(a) Released state (b) Pressed state

Figure 1: The user interface for the acoustic QoE evaluation experiment.

his/her performance. Several mechanisms [3, 13, 24] have been proposed to integrate an incentive structure into reputation systems in order to encourage good behavior and stop bad behavior.

3. THE PROPOSED FRAMEWORK

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

3.1 Experiment Designs

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

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

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

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

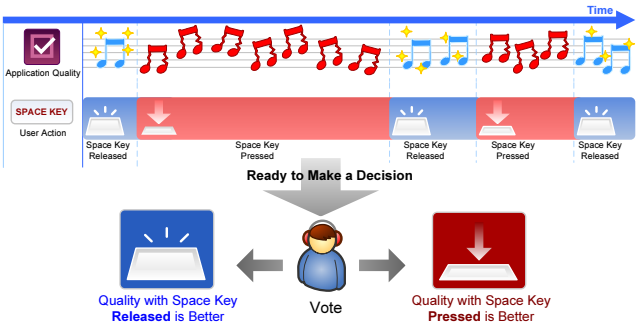
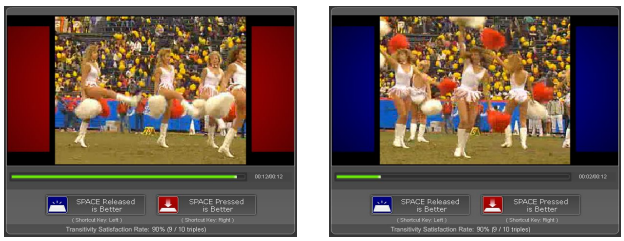


Figure 2: The concept flow of an experiment participant in an acoustic paired comparison.



(a) Released state (b) Pressed state

Figure 3: The user interface for the optical QoE evaluation experiment.

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

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

The user interface of our system for optical QoE evalua-

[2]In our experiments, each round normally took between 5 and 25 seconds.

tion is shown in Fig. 3. The major difference between it and the acoustic interface (Fig. 1) is that the state indicators are more sophisticated. Since participants need to pay full attention to the video on the screen, we provide an audio effect so that participants will hear a short, sharp tone whenever the SPACE key is pressed or released. The pitch is higher when the SPACE key is released and lower when the SPACE key is pressed; thus, participants can focus on the quality difference of the presented video and rely on the audio effect to determine the current state. In addition, the background color of the video playout pane indicates the current state (red and blue correspond to **Pressed** and **Released** respectively), so that a participant can perceive the current state via his/her peripheral vision. Similar to the acoustic interface, users can watch the video clip repeatedly until they can determine the quality difference between the two states, and then vote by pressing the LEFT key or RIGHT key accordingly. Readers can also experience this experiment design at `http://mmnet.iis.sinica.edu.tw/link/dv`.

3.2 Individual Consistency Checking

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

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

Algorithm 1 TSR Calculation

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 ← 0
n_pass ← 0
for all i, j, k, 1 ≤ i, j, k ≤ n, i ≠ j ≠ k do
    if m[i,j] = 1 and m[j,k] = 1 then
        n_test ← n_test + 1
        if m[i,k] = 1 then
            n_pass ← n_pass + 1
        end if
    end if
end for
TSR ← n_pass/n_test
```

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

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

3.3 Overall Consistency Checking

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

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

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

Kendall's u-coefficient is defined as follows:

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

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

3.4 Inference of QoE Scores

If the consistency of the collected experiment results is confirmed, we can proceed to the modeling step and infer

the quantitative QoE scores for the quality levels being evaluated.

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

	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}	–

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

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

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

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

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

Model Interpretation

The computed $u(T_i)$ for the quality level T_i from the fitted BTL model conforms to the relationship in Eq. 1. It must be negative since $u(T_i) = \log \pi(T_i)$ and $\pi(T_i)$ is a positive real number smaller than 1. To extract interpretable QoE scores, we normalize all the QoE scores between 0 and 1. By so doing, the quality level with the highest QoE 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 it allows us to compare the QoE scores of different quality levels, assuming that the perfect scheme achieves a QoE score of 1 and the worst scheme achieves a score of 0.

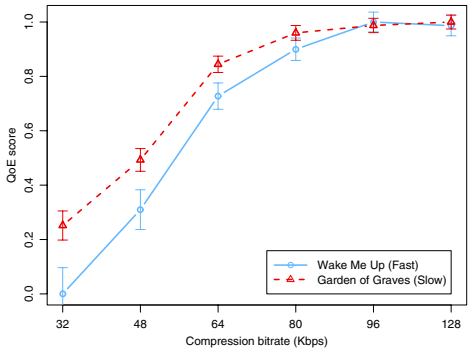


Figure 4: QoE scores of MP3-compressed songs at different bit rates.

4. CASE STUDIES: ACOUSTIC QOE EVALUATION

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

4.1 MP3 Compression Level

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

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

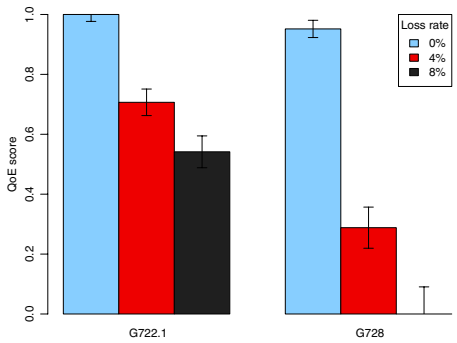


Figure 5: QoE scores of VoIP speeches encoded by different codecs at different packet loss rates.

diminishing marginal utility exists; that is, the increase rate of audio quality declines when the compression bit rate is higher. This type of utility curve is often seen in the relationship between a system's quality and users' perceptions because there must be an upper limit, after which increasing the system quality will not enhance users' experience any further.

We also find that the characteristics of an audio clip may affect the listening quality even when identical compression levels are used. It is known that a slower song contains a larger number of similar pitches than a faster song, so they tend to have higher compressibility; therefore, an MP3 encoder can include more granular information in a fixed-bit-rate output stream for a slower song. This explains why the QoE scores of "Garden of Graves" are higher than those of "Wake Me Up Before You Go Go" at low bit rates. However, the difference in the audio quality of the songs is less obvious at higher bit rates because less information is dropped during the compression process, which means that slower songs no longer benefit from their high compressibility.

4.2 Effect of Packet Loss on VoIP

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

A total of 62 participants performed 103 experiments, which involved 1,545 paired comparisons. Fig. 5 shows the QoE scores for the 6 test clips. From the graph, it is clear that a higher packet loss rate leads to a lower QoE score. 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 8% loss rate is still 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 bit rate and a higher sampling rate than G728, and it is also more robust to packet loss. This case study demonstrates

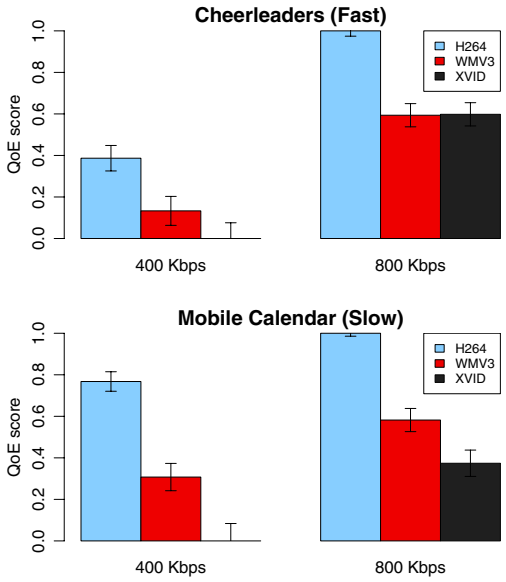


Figure 6: QoE scores of video clips compressed by different codecs at different bit rates.

the effectiveness of our framework in VoIP speech quality assessment studies.

5. CASE STUDIES: OPTICAL QOE EVALUATION

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

5.1 Video Codec

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

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

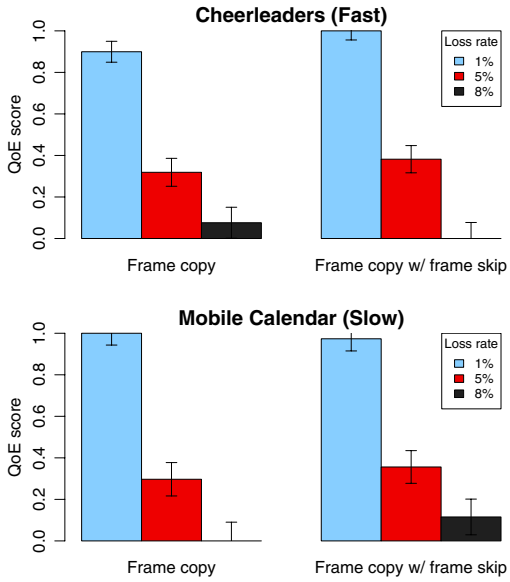


Figure 7: QoE scores of video clips decoded by different loss concealment schemes at different packet loss rates.

800 Kbps. Interestingly, on the "Mobile Calendar" video, WMV3 performs significantly better than XVID at the same bit rate. This indicates that WMV3 is generally better than XVID; the exception is that XVID is comparable to WMV3 on fast-motion videos at high bit rates.

5.2 Loss Concealment Scheme

When designing a high-quality IPTV system, one of the most challenging issues is how to deal with video packet loss due to network loss or excessive variations in network delay. A number of loss concealment schemes have been proposed, e.g., the intuitive frame copy method and the more sophisticated error resilient coding approach. In this case study, we evaluated two loss concealment schemes, namely the frame copy (FC) scheme and the frame copy with frame skip (FCFS) scheme [35], under different degrees of packet loss. The FC scheme conceals errors in a video frame by replacing a corrupted block with the block in the corresponding position in the previous frame. On the other hand, FCFS is a hybrid scheme that integrates the frame copy and the frame skip technique, which simply drops a frame that is corrupted due to packet loss. In our implementation of FCFS, if the percentage of corrupted slices in a frame exceeds 10%, it will skip (i.e., drop) the frame; otherwise it will apply the frame copy method to conceal the errors. The "Cheerleaders" and "Mobile Calendar" video clips (Section 5.1) were also used in this case study. We compressed both clips by JM [1], the H.264/AVC reference software, 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 6 test clips.

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

6. DISCUSSION

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

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

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

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

We also checked the overall consistency of experiment results from three participant sources. After removing unqualified experiments, we computed the average TSR of the

Table 2: A comparison of laboratory and crowdsourced experiments in terms of their cost and performance

Case Study	Participant Source	Total Cost (dollar)	# Rounds	# Person	Qualified Rate	Cost / Round (cent)	Time / Round (sec)	Avg. TSR	WST Violation	MST Violation	SST Violation	Kendall
MP3 Bit Rate	Laboratory	50.97	1,440	10	67%	3.54	16	0.96	0	0	0.05	0.65
	MTurk	7.50	750	24	47%	1.00	9	0.96	0	0.05	0.20	0.58
	Community	1.03	1,470	93	54%	0.07	25	0.96	0	0	0.30	0.58
VoIP Quality	Laboratory	22.95	675	10	67%	3.40	16	0.98	0	0	0.05	0.78
	MTurk	3.00	300	15	74%	1.00	19	0.98	0	0	0.20	0.78
	Community	0.40	570	37	86%	0.07	24	0.98	0	0	0.10	0.85
Video Codec	Laboratory	23.73	1,500	10	80%	1.58	7	0.98	0	0	0.15	0.57
	MTurk	4.95	495	23	65%	1.00	17	0.98	0	0	0.20	0.68
	Community	0.95	1,350	88	71%	0.07	11	0.97	0	0	0.25	0.57
Loss Concealment	Laboratory	51.93	1,260	11	69%	4.12	19	0.96	0	0	0.30	0.60
	MTurk	5.85	585	21	36%	1.00	25	0.97	0	0	0.40	0.60
	Community	0.63	900	59	35%	0.07	21	0.96	0	0	0.25	0.52
Overall		173.88	11,295	298	59%	1.54	17	0.97	0	0.01	0.20	0.65

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

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

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

Limitations

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

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

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

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

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

7. CONCLUSION AND FUTURE WORK

In this work, we have proposed and evaluated a crowdsourceable framework for assessing the QoE of multimedia content. We have shown that, under our framework, researchers can outsource QoE evaluation experiments to an

Internet crowd without compromising the quality of the results; and, at the same time, achieve wider participant diversity at a lower monetary cost.

In the future, we plan to release our experiment platform for public use. Currently the platform is under alpha release and can be accessed via `http://mmnet.iis.sinica.edu.tw/link/qoe`. Researchers will be able to publish their multimedia content on the platform and utilize an Internet crowd to conduct experiments at a lower cost. We hope that the proposed crowdsourcing framework and platform for QoE evaluations will prove helpful to researchers interested in evaluating the quality of multimedia content.

Acknowledgements

The authors would like to thank Wei Tsang Ooi and the anonymous reviewers for their constructive comments. This work was supported in part by the Taiwan E-learning and Digital Archives Program (TELDAP), sponsored by the National Science Council of Taiwan under grants NSC98-2631-001-011 and NSC98-2631-001-013. It was also supported in part by the National Science Council of Taiwan under grants NSC96-2628-E-001-027-MY3 and NSC98-2221-E-001-017.

8. REFERENCES

[1] H.264/AVC reference software JM 15.1. http://iphome.hhi.de/suehring/tml/.

[2] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[3] A. Blanc, Y.-K. Liu, and A. Vahdat. Designing incentives for peer-to-peer routing. In *Proceedings of IEEE INFOCOM 2005*, pages 374–385, March 2005.

[4] P. Bordia. Face-to-face versus computer-mediated communication: A synthesis of the experimental literature. *Journal of Business Communication*, 34(1):99–118, 1997.

[5] D. Brabham. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence*, 14(1):75, 2008.

[6] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[7] K.-T. Chen, C.-Y. Huang, P. Huang, and C.-L. Lei. Quantifying Skype user satisfaction. In *Proceedings of ACM SIGCOMM 2006*, Pisa, Itlay, Sep 2006.

[8] 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.

[9] S. Choisel and F. Wickelmaier. Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. *The Journal of the Acoustical Society of America*, 121(1):388–400, 2007.

[10] H. A. David. *The Method of Paired Comparisons*. Oxford University Press, 1988.

[11] 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*, 47(4):511–525, 1998.

[12] B. Duffy, K. Smith, G. Terhanian, and J. Bremer. Comparing data from online and face-to-face surveys. *Internation Journal of Market Research*, 47(6):615–639, 2005.

[13] A. Fernandes, E. Kotsovinos, S. Otring, and B. Dragovic. Pinocchio: Incentives for honest participation in global-scale distributed trust management. In *Proceedings of iTrust2004*, pages 63–77, 2003.

[14] C.-J. Ho, T.-H. Chang, and J. Y.-j. Hsu. Photoslap: A multi-player online game for semantic annotation. In *Twenty-Second Conference on Artificial Intelligence (AAAI-07)*, Vancouver, British Columbia, July 2007.

[15] 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)*, Paris, France, 2009.

[16] J. Howe. The rise of crowdsourcing. *Wired Magazine*, 14(6):176–183, 2006.

[17] Y. Ito and S. Tasaka. Quantitative assessment of user-level QoS and its mapping. *IEEE Transactions on Multimedia*, 7(3):572–584, June 2005.

[18] ITU-R Recommendation BT.500-11. Methodology for the subjective assessment of the quality of television pictures, 2002.

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

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

[21] 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.

[22] R. Jain. Quality of experience. *IEEE Multimedia*, 11(1):96–97, Jan.-March 2004.

[23] S. Jain, Y. Chen, and D. C. Parkes. Designing incentives for online question and answers forums. In *10th ACM Electronic Commerce Conference (EC'09)*, 2009.

[24] R. Jurca and B. Faltings. An incentive compatible reputation mechanism. In *Proceedings of IEEE International Conference on E-Commerce Technology*, pages 285–292, June 2003.

[25] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of ACM CHI'08*, pages 453–456, 2008.

[26] 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*, 11(1):87–95, 2004.

[27] R. D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959.

[28] J. N. S. Matthews and K. P. Morris. An application of bradley-terry-type models to the measurement of pain. *Applied Statistics*, 44:243–255, 1995.

[29] 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*, 43(1):41–46, 1978.

[30] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman. Reputation systems. *Commun. ACM*, 43(12):45–48, 2000.

[31] P. Rossi, Z. Gilula, and G. Allenby. Overcoming scale usage heterogeneity: A bayesian hierarchical approach. *Journal of the American Statistical Association*, 96(453):20–31, 2001.

[32] T. L. Saaty. A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, 15(3):234–281, 1977.

[33] M. V. Selm and N. W. Jankowski. Conducting online surveys. *Quality and Quantity*, 40(3):435–456, 2006.

[34] A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. In *Computer Vision and Pattern Recognition Workshops (CVPRW '08)*, pages 1–8, June 2008.

[35] 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*, pages 259–268, Vancouver, Canada, 2008.

[36] A. Watson and M. A. Sasse. Measuring perceived quality of speech and video in multimedia conferencing applications. In *Proceedings of ACM Multimedia 1998*, pages 55–60. ACM, 1998.

[37] K. B. Wright. Researching Internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of Computer-Mediated Communication*, 3(10), 2005.