Quadrant of Euphoria: A Crowdsourcing Platform for QoE Assessment

Kuan-Ta Chen[1], Chi-Jui Chang[1], Chen-Chi Wu[2], Yu-Chun Chang[2], and Chin-Laung Lei[2]

[1]Institute of Information Science, Academia Sinica

[2]Department of Electrical Engineering, National Taiwan University

Abstract

Existing QoE (Quality of Experience) assessment methods, subjective or objective, suffer from either or both problems of inaccurate experiment tools and expensive personnel cost. The panacea for them, as we have come to realize, lies in the joint application of *paired comparison* and *crowdsourcing*, the latter being a Web 2.0 practice of organizations asking ordinary, unspecific Internet users to carry out internal tasks. We present in this article Quadrant of Euphoria, a user-friendly, web-based platform facilitating QoE assessments in network and multimedia studies, with features low cost, participant diversity, meaningful and interpretable QoE scores, subject consistency assurance, and burdenless experiment process.

Index Terms

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

The everlasting endeavor by network and multimedia researchers to satisfy the end-users' growing needs has spawned many serious disciplines, of which the study and assessment of Quality of Experience [1] has proved one of the most challenging. Quality of Experience, or QoE for short, rivets on the *true feelings* of end-users from *their* perspective when they watch podcasts, listen to digitized music, and leaf through online photo albums, to name a few. It should not, however, be confused with QoS (Quality of Service), which refers to an objective system performance metric, such as the bandwidth, latency, or packet loss rate of a communication network.

Methods assessing QoE are conventionally classified as either subjective or objective. Subjective methods, in particular Absolute Category Rating [2], directly ask human-beings to rate their experience with some received media, also known as stimuli, on a categorical scale, the most adopted of which being the Mean Opinion Score (MOS). A single human rating in a MOS test is expressed as one of the ratings from 1 to 5. The numbers are also given names: Bad for 1, Poor for 2, followed Fair, Good, and Excellent. The MOS for a certain stimulus is then the arithmetic mean of individual ratings. The obvious drawback of subjective methods is the personnel cost and time, especially if evaluation has to be repeatedly conducted in iterative and incremental system development. In response, objective methods estimate QoE by analyzing the delivered content *automatedly*, for example by looking for unnatural noise appearing in a compressed audio clip. Unfortunately, no matter how sophisticated objective methods are, intrinsically they cannot capture all dimensions of QoE. PESQ, for instance, gives inaccurate predictions when used in conjunction with factors like sidetone, listening levels, loudness loss, talker echo, and effect of delays in VoIP conversation. External factors, such as the production quality of headsets (in acoustic QoE assessments) or the distance between viewer and display (in visual QoE assessments), are not considered by objective methods because they are hard to measure and quantify. Conventional subjective and objective approaches to assessing QoE remain more complements than replacements of each other. Subjective experiments are still called for to help develop mathematical models and authenticate results obtained from objective analyses despite their lavishness.

Nonetheless, the subjective methodology is not without pitfalls. It may be burdensome for one to map his own sensation onto the MOS scale, and the fact that it is concomitantly numeral and nominal does not help, either. As a matter of fact, the literature has identified at least two other problems with MOS:

1) **Scale heterogeneity** [3]. The options on the MOS scale are not something readily defined and explained. Consequently, each subject may interpret the scale according to his idiosyncratic preference and strategy. Some may tend to give higher ratings while others give below average ones even if they share similar experience toward the same stimulus.

2) **Scale ordinality** [4]. The MOS scale is actually not interval but ordinal, or put otherwise, a ranked order of five arbitrary numbers. The cognitive distance between Bad (1) and Poor (2) is usually not the same as that between Good (4) and Excellent (5). It is thus questionable to calculate the final score by taking the arithmetic mean, which is only defined on interval scales.

To give an overview of the article, we identify that there are two motivations or problems to be solved: that *subjective QoE tests, in particular MOS, have identifiable intrinsic problems*, and that *they incur high personnel and time costs*. We propose replacing category rating with paired comparison to tackle the first ("A Kickoff with Paired Comparison"), and employing crowdsourcing to address the second ("A QoE Assessment Platform"). Crowdsourcing, however, brings about another problem ("The Crowd Is Not All Trustworthy"), inspiring the invention of Transitivity Satisfaction Rate to ensure participant consistency ("Ensuring Subject Consistency"). The combination of all these endeavors is an unprecedented QoE assessment platform, the titular Quadrant of Euphoria, whose user interface and operation are described in "Platform Design". Finally, we evaluate the platform and and discuss the use of crowdsourcing in general in "Case Studies and Evaluation".

A Kickoff with Paired Comparison

To tackle the problems aforementioned, we propose to assess QoE with *paired comparison*, so that the test subject needs only to choose one better stimulus at a time from two based on his perception. The dichotomous decision is clearly less onerous and less confusing to participants of QoE experiments than a scaled rating. The features of paired comparison include:

- Applicability to all kinds of network and multimedia systems;
- Elimination of scaled rating and all accompanying problems;
- Quantified assessments of QoE, i.e., QoE scores, on an interval scale through the use of available probabilistic choice models [5];

and finally,

- The transitive property of preference, which is instrumental in checking the consistency of unsupervised subjects (described later).

Suppose we have n algorithms for, say, audio compression to rate. One by one we apply them to a "testbed" audio clip, creating n synchronized stimuli to be paired with each other. There are $\binom{n}{2}$ possible pairs, each with two stimuli semantically equivalent at every second except for their presentation quality. In our design, each pair corresponds to a decision or judgment to be made by the subject. The $\binom{n}{2}$ judgments in turn constitute one of the subject's many runs of a QoE experiment. (We do encourage participants to perform an experiment multiple times.) The results of all runs of an experiment can be collectively summarized in a frequency matrix resembling

	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}	–

where $T_1, T_2, ..., T_n$ are the n prepared stimuli ($n = 4$ in the above matrix) and a_{ij} denotes the number of runs (not participants) where T_i is preferred to T_j. The total number of runs is of course $a_{ij} + a_{ji}$.

The Bradley-Terry-Luce (BTL) model [5] states that the probability of choosing T_i over T_j, $\frac{a_{ij}}{a_{ij}+a_{ji}}$, is a function associated with the "true" ratings of the two stimuli:

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

from which we obtain $u(T_i) = \log \pi(T_i)$ by maximum likelihood estimation. The numerals $u(T_i), i = 1, 2, ..., n$ are comparable with each other on an interval scale and are thus chosen as the raw estimates of QoE score for $T_1, T_2, ..., T_n$, respectively. $u(T_i)$ is negative since $\pi(T_i)$ is a positive real number smaller than 1. To facilitate interpretation, we further shift and normalize the raw estimates to within $[0, 1]$, where the stimulus with the best QoE scores 1 and the one with the worst scores 0.

A QoE Assessment Platform

We beef up our methodological reform of QoE experiments with the proposition of an *assessment platform* powered by paired comparison. We name it Quadrant of Euphoria as a backronym of QoE. Such a platform is not complete without addressing the cost issue aforesaid. Recent technology advances have made available ubiquitous Internet access and rich Internet applications, giving rise to a generation of more participative and self-aware end-users, the

Internet crowd. We argue that they are the ideal subjects of QoE experiments for their headcount, diversity, *and* (relative) nonchalance to monetary rewards. They often respond to problem-solving requests solely for kudos, intellectual satisfaction, or a sense of being helpful. It is thus perfectly reasonable to *crowdsource* QoE assessing experiments instead of hiring part-timers to carry them out in the laboratory, since, after all, the crowd is to whom researchers labor to provide ever-improving service quality.

Crowdsourcing (yet another Web 2.0 neologism) advocates mass collaboration and the wisdom of the commons. The academia has long embraced it even before the name was coined. NASA-sponsored Clickworkers[1], for instance, began to utilize online volunteers to classify Martian craters as early as November 2000. Crowdsourcing also has potential in the public sector, most notably in the area of urban planning [6]. Commercially speaking, crowdsourcing is a further step from outsourcing in that the task performers are no longer specific, identifiable persons. Footwear manufacturers like Portland-based RYZ are known to have adopted *community-based designs*, thereby cutting costs and cultivating a kinship between themselves and customers.

Conceptually, we picture Quadrant of Euphoria as a fulfilment of network and multimedia researchers' need to conduct QoE experiments without *any* interface programming, thus being able to focus on their fields of expertise while benefiting from the advantages of paired comparison. The idea of a crowdsourcing platform is drawn from websites like InnoCentive [7] and the Amazon Mechanical Turk (MTurk)[2]. InnoCentive is a service through which organizations seek the assistance of a global scientific community for innovative solutions to challenging R&D problems, and give professional recognition and cash rewards to problem-solvers in return. MTurk is a popular crowdsourcing marketplace where anyone calling for help from the Internet crowd can post their tasks. The tasks to be performed can be of any kind, ranging from surveys, experiments to answering mundane questions.

The Crowd Is Not All Trustworthy

We focus now on the one formidable challenge to crowdsource QoE experiments: Not every Internet user is good-natured. Since subjects perform experiments without supervision, they

[1] National Aeronautics and Space Administration, http://clickworkers.arc.nasa.gov/

[2] Mechanical Turk, http://www.mturk.com/

may give erroneous feedback and still receive payment therefor. Such results are products of a careless, perfunctory attitude, or more perturbingly, of dishonesty and malign conduct. Whatever the reason is, erroneous ratings do increase the variances of QoE scores and lead to biased conclusions.

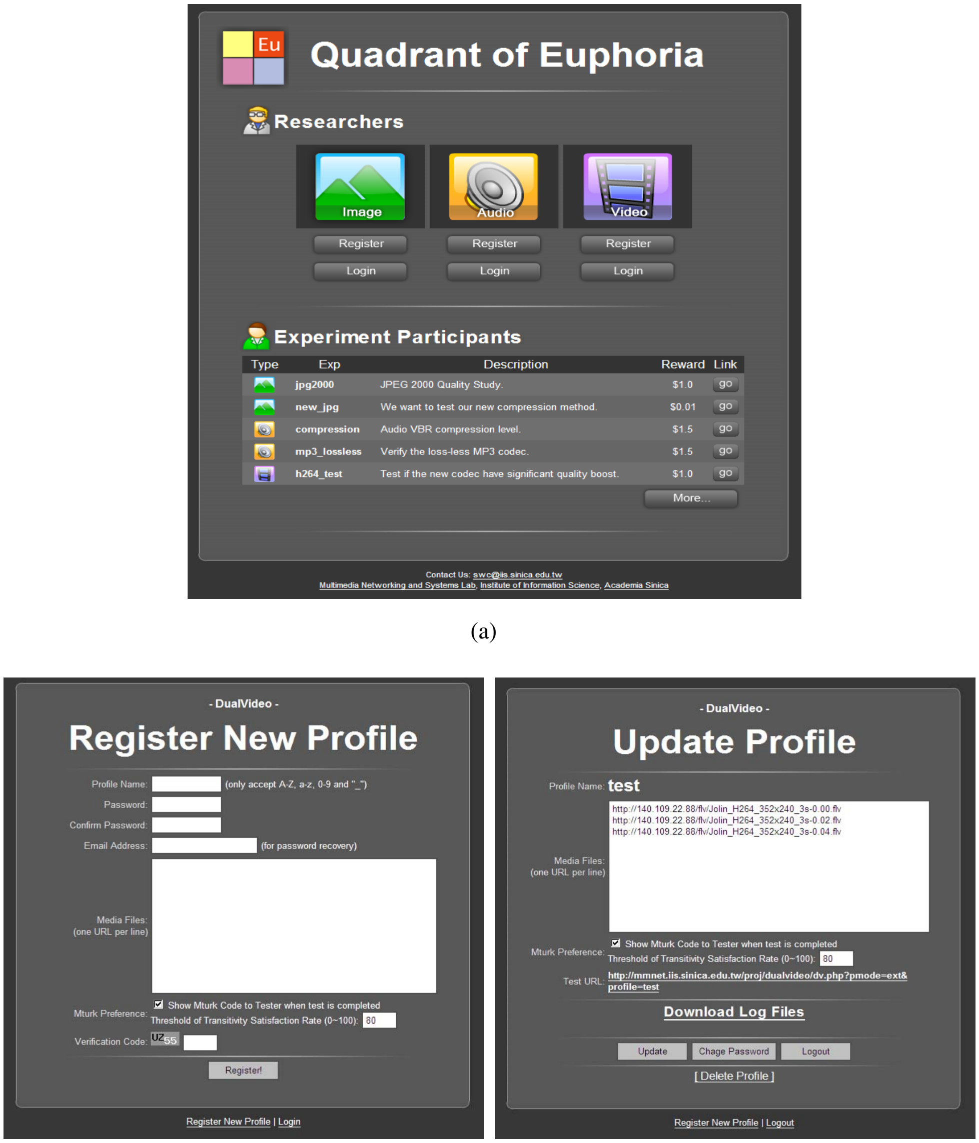
Given a handful of results turned in by anonymous subjects, it is difficult, if not impossible, for an experimenter to sift the wheat from the chaff. One may argue that we can compensate problematic inputs by amassing more experiment runs than necessary, but the approach is valid only if ill-natured users occupy a small portion of the crowd. However, since they may choose random answers by ignoring instructions and effectively earn more than honest participants, they are motivated to run (and sabotage) an experiment as many times as they can. We are in dire need of a countermeasure to finally establish a theoretically consummate platform. Problematic inputs must be pruned. Reward and punishment rules can also ensue to encourage a high caliber of participation and thwart potentially uninterested subjects.

Ensuring Subject Consistency

For our purposes, it is asserted that preference is a *transitive* relation; that is, if some participant prefers A to B and B to C, than he will normally prefer A to C. In light of this, we define the Transitivity Satisfaction Rate (TSR) to be the number of judgment triplets (e.g., the three preference relations between A, B, and C) satisfying transitivity divided by the total number of triplets where transitivity may apply. The TSR is the quantification of a participant's consistency throughout a run of an experiment. A rule of thumb is that a fully attended subject can attain a TSR higher than 0.8 without difficulty.

We suggest that experiment administrators posit before the tests specific requirements of a paying result and possibly a warning to rogue participants. In our demonstrative studies, for example, only the results with TSRs of 0.8 and above were rewarded, and never once did we receive complaints.

In addition, we argue that there is no systematic way to cheat our system. The fact that the order in which the pairs appear and the correspondence between key states and stimuli are completely randomized and unavailable to outsiders contributes to this assertion. The only ways a participant can achieve high TSR and get paid are to give sound and honest answers and, though easily dismissed, to consistently make wrong judgments.



Quadrant of Euphoria

Researchers

Experiment Participants

Type	Exp	Description	Reward	Link
🟩	jpg2000	JPEG 2000 Quality Study.	$1.0	go
🟩	new_jpg	We want to test our new compression method.	$0.01	go
🟧	compression	Audio VBR compression level.	$1.5	go
🟧	mp3_lossless	Verify the loss-less MP3 codec.	$1.5	go
🟪	h264_test	Test if the new codec have significant quality boost.	$1.0	go

Contact Us: swc@iis.sinica.edu.tw
Multimedia Networking and Systems Lab, Institute of Information Science, Academia Sinica

(a)

- DualVideo -

Register New Profile

Profile Name: (only accept A-Z, a-z, 0-9 and "_")
Password:
Confirm Password:
Email Address: (for password recovery)
Media Files: (one URL per line)
Mturk Preference: ☑ Show Mturk Code to Tester when test is completed
Threshold of Transitivity Satisfaction Rate (0~100): 80
Verification Code:

Register New Profile | Login

(b)

- DualVideo -

Update Profile

Profile Name: test
Media Files: (one URL per line)
http://140.109.22.88/flv/Jolin_H264_352x240_3s-0.00.flv
http://140.109.22.88/flv/Jolin_H264_352x240_3s-0.02.flv
http://140.109.22.88/flv/Jolin_H264_352x240_3s-0.04.flv
Mturk Preference: ☑ Show Mturk Code to Tester when test is completed
Threshold of Transitivity Satisfaction Rate (0~100): 80
Test URL: http://mmnet.iis.sinica.edu.tw/proj/dualvideo/dv.php?pmode=ext&profile=test

Download Log Files

Register New Profile | Logout

(c)

Fig. 1: Quadrant of Euphoria's portal and administrative pages. (a) The Portal. The upper pane directs researchers to administrative pages; the lower pane displays for subjects the catalogue of open experiments. (b) Profile registration page. (c) Profile update page.

Platform Design

The portal to Quadrant of Euphoria (`http://mmnet.iis.sinica.edu.tw/link/qoe/`) is of a role-based layout that serves researchers and participants of their experiments alike (Figure 1a). From the upper pane researchers are directed to the administrative pages, where they can register or update their experiment "profiles," and download results, or logs, for further analysis. In contrast, an Internet user may browse on the lower pane through the catalogue of open experiments on our website and find the ones that interest him to participate. The catalogue informs the user of the experiments' name, type (Image, Audio, or Video as currently supported), description, and payment level.

To Set Up and Conduct an Experiment

Experiments are maintained on Quadrant of Euphoria as *profiles*. The procedure by which a researcher sets up a profile and conducts the experiment is shown in Figure 2. Stimuli must be prepared beforehand and uploaded upon profile registration. Once registration is successful, the researcher is given the hosted experiment URL, which he is free to publish to any Internet community to gather the subject crowd. If monetary reward is involved, a micro-payment platform such as MTurk is recommended. The researcher now awaits the crowd to perform the experiment. Subjects may receive unique verification codes for complete and qualified results and use them to prove to the researcher their eligibility for payment. Issuance of the codes can be set during profile registration. Finally, the researcher decides on paying whom how much and collects data for further analysis.

Experiment Interface

When a subject enters an experiment Web page hosted on Quadrant of Euphoria, he sees an Adobe Flash application with a large upper pane (Figure 3) and immediately begins the first paired comparison. Depending on the context of the experiment, the participant will be able to view an image, hear an audio clip, or watch some visual content displayed on the upper pane. Upon pressing and holding the spacebar, the (objective) quality of the content changes. Releasing the spacebar restores the original quality at the start of the comparison. The subject then makes a judgment on which spacebar state (pressed or released) corresponds to better (perceived) quality by mouse-clicking one of the buttons beneath the display pane or by pressing left or right arrow keys. The next paired comparison commences right after the decision, or the experiment ends after all $\binom{n}{2}$ comparable pairs are exhausted.

What the subject hears or sees is actually a *dynamic interweaving* of a pair of stimuli. One stimulus in the pair is played out first. If the participant presses and holds the spacebar at, say, the fifth second, the other stimulus will take over *seamlessly* and start playing from its

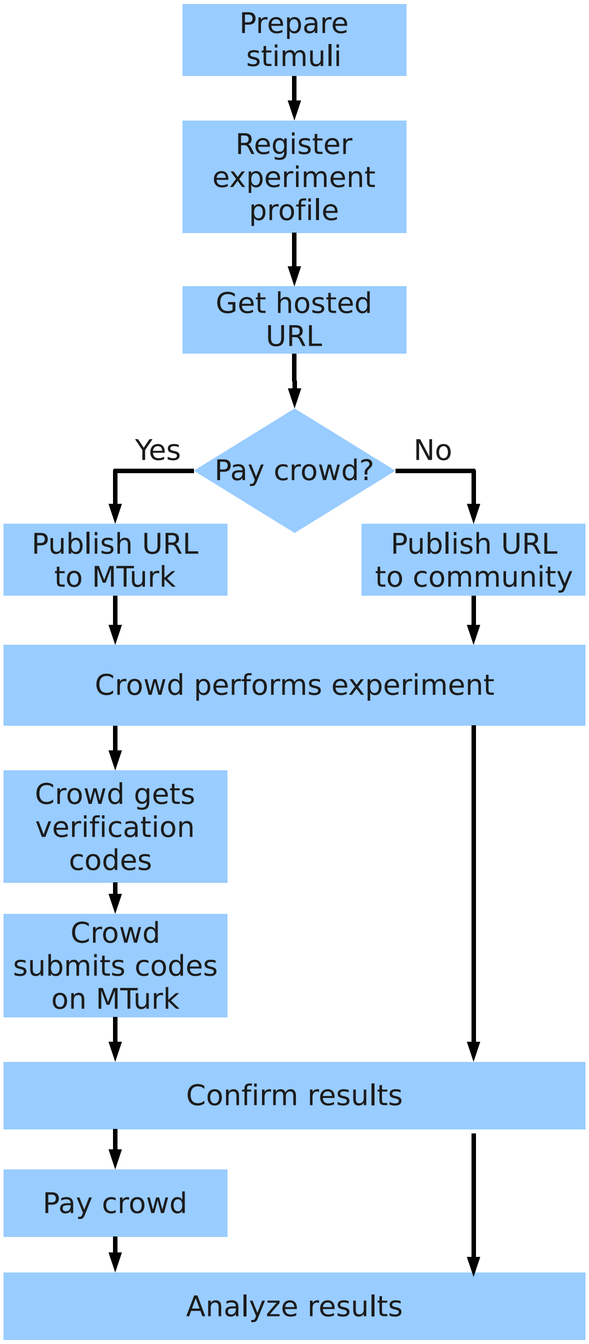
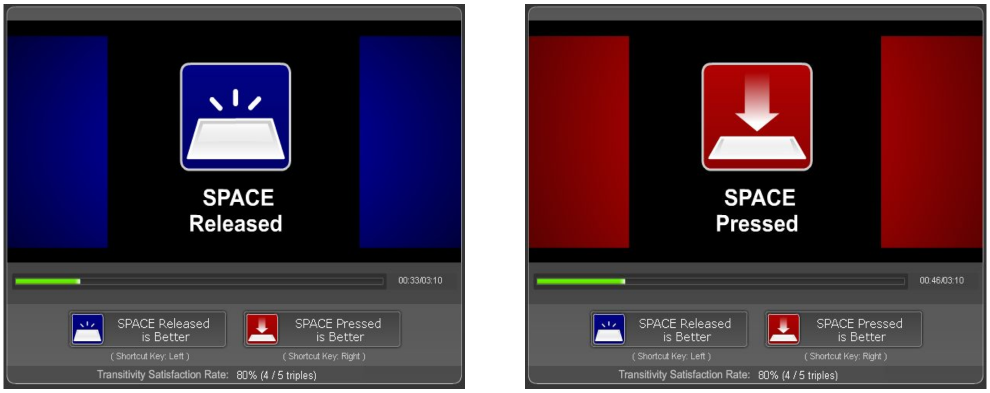


Fig. 2: Flowchart of a researcher conducting an experiment.

fifth second and onwards, giving the participant the illusion that an audio clip is played with adjustable quality levels. The decision flow that a subject goes through is illustrated in Figure 4.



(a) Acoustic QoE assessment: spacebar released.

(b) Acoustic QoE assessment: spacebar pressed.

Fig. 3: The experiment interface as seen by participants under both spacebar states.

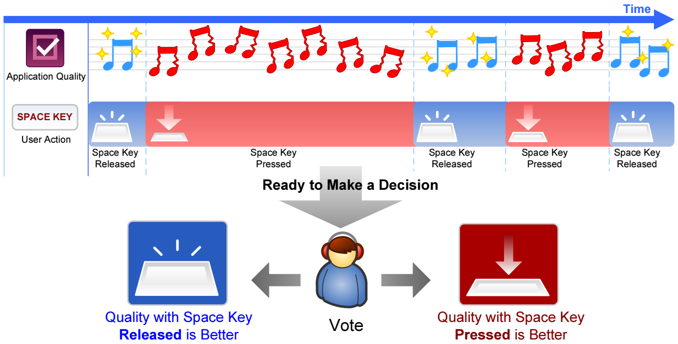


Fig. 4: The decision flow of a subject in an acoustic paired comparison.

Administration Interface

A profile is identified by its name, which along with a password is required when a researcher logs in to manage his experiment. When registering a profile (Figure 1b), the researcher provides URLs of the stimuli and an e-mail address for password recovery. He can also set up a TSR threshold to fend off disqualified results, and opt to show unique verification codes on qualified ones. All but the profile name and the backup e-mail address are modifiable afterwards (Figure 1c).

The logs of the experiment are zipped and available for download on the profile management page. We also bundle in the ZIP archive the source code to infer QoE scores and draw diagrams like Figures 5 and 6. The result of each participant makes up a text file with the name

`datetime_ipaddr_profile_sname_vcode.txt`,

where `sname` and `ipaddr` are respectively the name and IP address of the participant, `datetime`

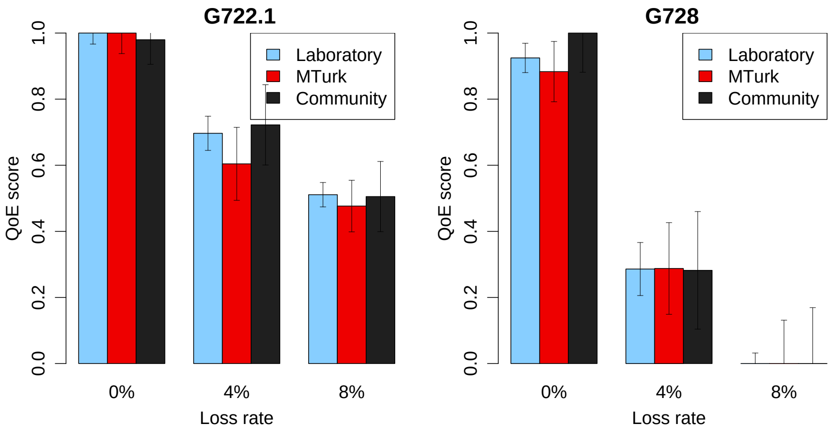


Fig. 5: QoE scores of VoIP recordings encoded by two codecs at various packet loss rates.

the POSIX timestamp of the result down to milliseconds, and `vcode` the verification code unique to the result. Each line within the log represents a judgment between a comparable pair with the format

`stimulus_A stimulus_B (A|B) time`,

where `A|B` represents a judgment of preferred QoE between stimuli A and B and `time` denotes the time spent on making the judgment in seconds.

Case Studies and Evaluation

We demonstrate Quadrant of Euphoria with two network-related case studies which, along with others detailed in [8], were carried out on the platform in three ways: in a physical laboratory, crowdsourced from MTurk, and crowdsourced from another populous Internet community. Such an arrangement provides us the stepping stone to evaluate the framework and explore the exciting possibilities of it.

Effect of Packet Loss on VoIP Quality

A three-minute long uncompressed recording of speech acquired from the Open Speech Repository was encoded into two voice packet streams by audio codecs G.722.1 and G.728 respectively. To simulate loss events, the streams were put into a Gilbert-Elliott channel [9], whose two states dictate whether a packet passing through is stripped off or not. The probability of the channel going from *allowing* to *blocking* state is p, and q vice versa. Juggling with the formulated loss rate, $\frac{p}{p+q}$, we were able to drop at will 0%, 4%, or 8% of incoming packets and decode the streams back into a total of six degraded recordings ready to be paired.

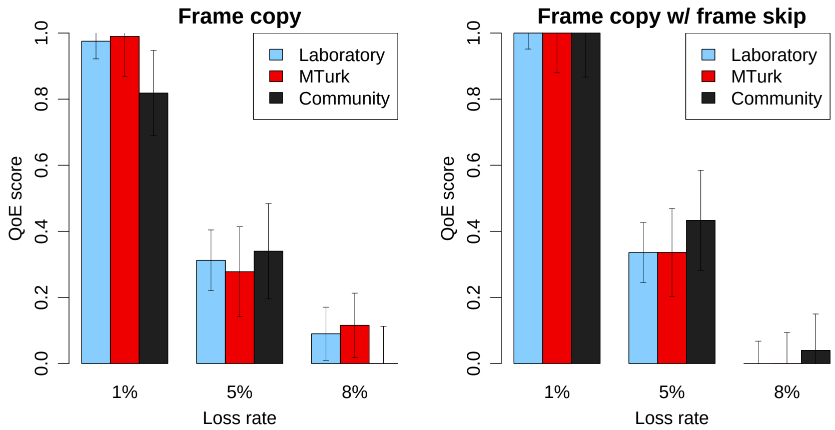


Fig. 6: QoE scores of a video clip repaired with two loss concealment schemes at various packet loss rates.

$1,545$ comparisons were performed by 62 subjects, including 10 for the laboratory, 15 from MTurk, and 37 from the other Internet community. The inferred QoE scores of the six recordings in Figure 5, properly normalized so that they are cross-comparable, exhibit general agreement among the three settings. They also conform with our expectations that 1) higher loss rates lead to lower QoE scores, and 2) G.722.1, operating at higher bitrates than G.728, is dominantly more robust even with higher loss rates. Also, the graph manifests that subjects from different participant sources reveal statistical equivalent perceptions in terms of their preferences for audio codecs operating at different bitrates and loss rates.

Comparison of IPTV Loss Concealment Schemes

A benchmark video clip “Cheerleaders” was encoded with JM [10], the H.264/AVC reference software, and again put into a Gilbert-Elliott channel to simulate packet loss events at rates 1%, 5%, and 8%. The three resulting streams were then decoded back and repaired with two loss concealment schemes: frame copy (FC) and frame copy with frame skip (FCFS). FC hides errors in a video frame by replacing a corrupted block with one at corresponding positions in previous frames. FCFS works exactly as FC until the percentage of corrupted blocks in a frame exceeds a certain threshold (10% in our experiments), then it simply drops that frame.

The inferred QoE scores in Figure 6 are properly normalized so that those of the same clip are comparable. While it is unsurprising that QoE scores are negatively correlated to loss rates, we note that there is no significant resultant discrepancy between the two loss concealment schemes.

TABLE I: Cost and performance summary of laboratory and crowdsourcing strategies. Despite lower Qualified Rates, with consistency assurance crowdsourcing can still yield results as fine as laboratory experiments (average TSR ≥ 0.96).

Case Study	Experiment Strategy	Total Cost ($)	# Runs	# Subjects	Qualified Rate	Cost/Run (¢)	Time/Run (sec)	Avg. TSR
MP3 Bitrate	Laboratory	50.97	1,440	10	67%	3.54	16	0.96
	MTurk	7.50	750	24	47%	1.00	9	0.96
	Community	1.03	1,470	93	54%	0.07	25	0.96
VoIP Quality	Laboratory	22.95	675	10	67%	3.40	16	0.98
	MTurk	3.00	300	15	74%	1.00	19	0.98
	Community	0.40	570	37	86%	0.07	24	0.98
Video Codec	Laboratory	23.73	1,500	10	80%	1.58	7	0.98
	MTurk	4.95	495	23	65%	1.00	17	0.98
	Community	0.95	1,350	88	71%	0.07	11	0.97
Loss Concealment	Laboratory	51.93	1,260	11	69%	4.12	19	0.96
	MTurk	5.85	585	21	36%	1.00	25	0.97
	Community	0.63	900	59	35%	0.07	21	0.96
Overall		173.88	11,295	298	59%	1.54	17	0.97

The subjects from different participant sources again show remarkably similar assessments of the impact of loss concealment schemes on video clips. We remark that laboratory subjects seem to have more consistent scores (represented by narrower error bars) than the crowd subjects. We ascribe the fact to the effect of practice: MTurk and community participants on average attend to 28 and 15 runs respectively, whereas in the laboratory subjects complete more than a hundred runs, enabling them to discern the subtle difference of the video clips more easily. We will revisit this argument shortly.

Crowdsourcing Evaluated

In addition to the studies above, we also asked the participants in all three settings (laboratory, MTurk, and community) to rate MP3 compression levels and various video codecs at different bitrates. Due to the article's length limit, we are obliged not to describe those studies in detail but do include their statistics in Table I to give a hologram of how crowdsourcing jostles against laboratory experiments.

Cost. We hired part-timers for laboratory experiments with an hourly pay of $\$8$. They were asked to perform the tests repeatedly within work hours. We also announced the recruitment on the MTurk website and another Internet community with 1.5 million users. Only qualified results (TSR ≥ 0.8) were rewarded for MTurk and the community. The compensations were $15¢$ and virtual currencies worth $1¢$ respectively.

We estimate that laboratory experiments consumed 86% of our budget while producing only 43% of the judgments ($4,875$ out of $11,295$). The cost per judgment was $3¢$ on average, a lot more expensive than that measured for MTurk ($1¢$) and community ($0.07¢$).

Quality. We define the Qualified Rate as the ratio of results in an experiment that yield a TSR higher than 0.8. The Qualified Rates observed are usually around 60%–70%. The laboratory experiments achieved the highest Qualified Rates in all cases except in the VoIP quality study. Moreover, in the study of loss concealment schemes, the laboratory setting boasted a 69% Qualified Rate, almost twice as high as those attained by both crowdsourcing strategies. Such polarized statistics indicate that the quality of the video stimuli in this experiment are more difficult to be differentiated. We attribute the superiority of laboratory subsjects in this case to their proficiency acquired during the course of the experiment, since on average each one of them made 115 paired comparisons, as opposed to a merely 18 by their crowdsourced counterparts. Despite the sometimes dismal Qualified Rate, crowdsourcing can still produce results of the same standard as laboratory runs after the removal of disqualified submissions. That our consistency assurance is ticking is evident in the Avg. TSR column of Table I, where an unanimous 0.96 is reached or surpassed.

Participant diversity. In addition to evaluating quality and cost aspects, we also emphasize the diversity of participants in QoE-assessing experiments. Since the purpose of these studies is to understand human perception of certain stimuli, e.g., multimedia content, a subject set as diverse as possible enables us to collect broader opinions and infer more *real* QoE scores. From this perspective, crowdsourcing is especially suitable for assessing QoE as it greatly increases the participant diversity. In our experiments, crowdsourcing strategies contributed 97% to a total of 298 subjects while costing only $\$24.31$ or 14% of the budget therein.

While we argue that crowdsourcing is a viable substitute to laboratory experiments, there are a few concerns which we cannot ignore without compromising the completeness of this article.

Environment control. Participants in a crowdsourcing setting experience the stimuli under a wide variety of conditions. Whether or not this is disadvantageous is subject to discussion. On the one hand, if the experiment is to assess QoE in a specific scenario, then crowdsourcing might not be the choice. On the other hand, crowdsourcing experiments are carried out in the *real world*, where there are subpar equipment, inevitable ambient interference, and no "typical" user environment.

Type of device. Since in the most case people connect to the Internet with their personal computers, the crowdsourcing strategy is most suitable for assessing QoE on such platform. When it comes to televisions, mobile phones, or electronic paper, the feasibility of crowdsourcing admittedly depends on these alternative devices' networking and computing capability, which fortunately is foreseeable in the very near future.

Type of media. For the time being, Quadrant of Euphoria supports only assessments of static image, audio, and video. While work needs to be done on interactive and streamed multimedia, a provisional solution is to let experiment participants download "perfect" stimuli and simulate network-related impairments such as delay or packet loss on the client side given that the partipants' machines are powerful enough.

Demography. As we do not physically recognize any one of the crowd, it is difficult to relate experiment results to gender, age, race, nationality, etc. with confidence, as a certain degree of virtuality is inherent to the Internet.

Platform Outlook

Quadrant of Euphoria, as its name suggests, strives to bring about an ease of mind to network and multimedia researchers who wish to be relieved from the annoyances of subjective QoE experiments. The foundations and evaluation of the platform we do not iterate here, but stress that it achieves participant diversity at a lower cost, along with enhancements in the theoretical accuracy and correctness of QoE score inference. In the future, we plan to keep on developing the platform toward the following directions:

- Streaming support for acoustic and visual QoE assessments;
- QoE assessment for interactive applications, such as video-conferencing and online game;

- Integrated micro-payment mechanism;
- User facilities like a search box for open experiments or personalized participation tracking;
- Multi-dimensional consistency quantification superior to TSR;
- A "neutral" or "indifferent" option for paired comparison and subsequent model changes.

The ultimate goal, of course, is to render the free, open-access Quadrant of Euphoria as much assisting to the research community as we are capable of.

ACKNOWLEDGEMENT

This work was supported in part by Taiwan E-learning and Digital Archives Programs (TELDAP) sponsored by the National Science Council of Taiwan under the grants NSC98-2631-001-011 and NSC98-2631-001-013. It was also supported in part by the National Science Council of Taiwan under the grants NSC98-2221-E-001-017.

REFERENCES

[1] ITU-T Recommendation P.10/G.100/Amd.2, "New definitions for inclusion in recommendation p.10/g.100," 2008.

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

[3] P. Rossi, Z. Gilula, and G. Allenby, "Overcoming scale usage heterogeneity: A bayesian hierarchical approach," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 20–31, 2001.

[4] 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.

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

[6] D. C. Brabham, "Crowdsourcing the public participation process for planning projects," *Planning Theory*, vol. 8, no. 3, pp. 242–262, 2009.

[7] R. Allio, "CEO interview: the InnoCentive model of open innovation," *Strategy & Leadership*, vol. 32, no. 4, pp. 4–9, 2004.

[8] 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.

[9] E. O. Elliott, “Estimates of error rates for codes on burst-noise channels,” *Bell System Technical Journal*, vol. 42, pp. 1977–1997, 1963.

[10] “H.264/AVC reference software JM 15.1,” http://iphome.hhi.de/suehring/tml/.