Lab Experiment vs. Crowdsourcing: A Comparative User Study on Skype Call Quality

Yu-Chuan Yen†, Cing-Yu Chu‡, Su-Ling Yeh*, Hao-Hua Chu§, Polly Huang†
†Department of Electrical Engineering, National Taiwan University
*Department of Psychology, National Taiwan University
§Department of Computer Science and Information Engineering, National Taiwan University
‡Department of Electrical and Computer Engineering, New York University – Polytechnics
{b99901079, r99921035, suling, haochu, pollyhuang}@ntu.edu.tw

ABSTRACT
To deliver voice over the Internet in a cost-effective way, it is essential to quantify the quality of user experience (i.e., QoE) of a voice service at various provisioning levels. Conducting user studies is an inevitable step facilitating quantitative studies of QoE. The two experimental methods – lab experiment vs. crowdsourcing via Amazon Mechanical Turk [1] – are compared in this study. We find that, for the study of Skype call quality, the crowdsourcing approach stands out in terms of efficiency and user diversity, which in turn strengthens the robustness and the depth of the analysis.

Categories and Subject Descriptors
C.2.2 [Computer-Communication Networks]: Network Protocols – applications.

General Terms
Performance, Experimentation, Human Factors.

Keywords
VoIP, QoE, Skype, Rate Control, Mechanical Turk

1. INTRODUCTION
As Skype, FaceTime, and Hangout’s Android/iOS ports become available, these IP telephony services are growing increasingly popular. It is predicted [2] that one out of seven mobile subscribers will be VoIP users. From a network engineer’s point of view, the higher bitrate we send the multimedia content, the better quality the users perceive. From a consumer’s point of view, many prefer and are paying the Internet access with a flat fee. Sending more does not seem to cost (the users) more and therefore a sensible thing to do as well. The trouble is network bandwidth is really not free. The more we spend, the more it costs to maintain the infrastructure. From an ISP’s point of view, and for the economic wellbeing of the Internet, it is evident that knowing how we can allocate network resources to the voice streams such that the paying users are mostly happy is crucial.

Towards a better and quantitative understanding of user perception, we investigated in [3] how users perceive Skype calls at different bitrates, i.e., different levels of bandwidth required to stream the calls. In that, we conducted a user study where 30 users were recruited to the lab to score audio tracks of a speech encoded in different bitrates. We were able to derive a logarithmic relationship in user score vs. bitrate, which indicated that allocating residual bandwidth to low-bitrate calls, as opposed to high-bitrate ones, would be more rewarding overall, considering the entire user population.

Untold of the study is that the user study, merely to collect user scores for a speech encoded in 9 bitrates, takes weeks of time from preparing the tests, recruiting users, scheduling time slots to supervising the experiments. Relative to average system experiments of the same scale, user studies are substantially more time and labor consuming and pose a significant challenge investigating user perception of VoIP services at scale.

Another issue is that needing users physically in the lab to score the audio tracks limits the diversity of the users in the experiment. In [3], most of the 30 users are students in the engineering school. It is not surprising that most of the users are male and in their late teens or early twenties. This poses also a challenge conducting fine-grained analysis of the user perception to factors such as gender and age, which may have a role in how users score.

Lab experiment being (1) time and labor consuming and (2) limited in user diversity over-shadows the pursuit of network engineering for user experience. In this study, we explore the alternative – crowdsourcing. Crowdsourcing, although not having the shortcomings of the lab experiment, is controversial [4]. It is largely due to the fact that researchers could not quite administer the users during the tests and the test results take time and effort to verify.

To cross compare, we conducted a crowdsourcing-based user study via Amazon Mechanical Turk (MTurk). MTurk is an experimental platform already used by studies in
social science [5], behavioral science [6], psychology [7], political science [8] and communication [9]. We find that in terms of the process: (1) the MTurk experiment is more time, labor, and money efficient. (2) Verifying results from MTurk is indeed an issue but, for call quality evaluation, it can be resolved by randomly planting reference audio tracks amongst the tests. In terms of the user scores, (1) the two sets of user scores are comparable. (2) Having a more diverse set of users joining the experiment, MTurk facilitates better in-depth analysis.

In the following sections, we (1) describe in detail the two experiments, (2) compare the time, labor, and cost of conducting lab vs. MTurk experiments, (3) confirm the logarithmic relationship in user score and bitrate still holds, (4) discuss whether the choice of experimental platform play a role analyzing fine-grained factors, and (5) examine factors that are not included in the lab test study.

2. USER STUDY
We begin in this section by describing the problem and the user study designed to tackle the problem. We then introduce the settings of the user study in two different experimental platforms: laboratory test and Amazon Mechanical Turk.

2.1 Skype Call QoE
Traditionally, an Internet service is evaluated by metrics such as throughput, loss rate, delay, and delay jitter. These metrics are referred to as the Quality of Service (QoS). In this QoS paradigm, VoIP services are predominantly measured by the loss rate, delay, and delay jitter. A recent trend is to weigh in more how exactly users feel about the played voice. This is referred to as the Quality of Experience (QoE). Despite their differences, QoS and QoE are not in competition to each other. Rather, they are complementary pieces in the longstanding puzzle of assessing the quality of multimedia services. By mapping the objective, network-centric QoS to the direct, user-centric QoE, one can assess realistically high-level user experience by measuring low-level QoS metrics in real time.

Seeking to find out how different sending bitrates influence the perceived quality of VoIP services, we have discovered in [3] that the relationship in between the sending bitrate and user score for Skype calls is logarithmic. The relationship is shown in Figure 1, where the x-axis depicts the bitrate and the y-axis the mean opinion score (MOS). One could observe that as the call bitrates increases, the perceived quality improves. One could also observe that given the same amount of extra bandwidth – Δ, a user of low-rate call feels stronger the improvement, as opposed to that of high-rate call.

One major challenge hindering the pursuit of better user experience is user-study methodology. Conducting user studies in laboratory environment is time-consuming and the participant diversity is low. It took us a few weeks to recruit 30 participants just to complete scoring of 9 audio tracks – one speech encoded in 9 bitrates. The time and effort grow exponentially as the number of QoS metrics we would like to jointly investigate increases. Although we would have fidelity in the score collected being able to administer users during the tests in person, lab experiment obviously would not scale to the number of QoS metrics to investigate. The alternative methodology, crowdsourcing, is not free of its own problems. One of the major technical issues concerning crowdsourcing is the fidelity of the user responses. In this study, we compare the measured QoE of Skype calls using Amazon Mechanical Turk, a popular crowdsourcing platform, to that of the lab experiment.

2.2 Experimental Design
The subjective Skype call quality assessment is conducted on two different user-study platforms. This allows us to compare the effectiveness and efficiency of conducting user studies on both platforms. Specifically, we encode the speech using SILK, the audio codec implemented in the latest version of Skype. The speech is encoded into multiple audio files of different encoding bitrates and each user is asked to score all audio files. The details are elaborated below.

Audio Source: Following recommendations of ITU-T P.830 [10], our source material consists of a number of simple, short, meaningful sentences which do not have obvious contextual connections. Two female and two male speakers are recruited to produce the voice of the audio source. Limited by the location of our laboratory test and potential participants, the audio source is recorded in Mandarin and all the speakers are native speakers. Length of the source material is 30 seconds and the sampling rate of recording is the standard 44.1 kHz.

Test Tracks: 9 different rates are selected. Hinted by the observed logarithmic QoE relationship of various Internet services from previous works [13][14], we choose to separate the 9 different rates exponentially to ensure an even sampling of resulting perceptual qualities. The density of rate choosing is higher at low quality region and lower at high quality region to ensure we get enough samples at the low rates where perceptual quality changes more rapidly.
The sending rates are 40.6, 27.7, 19.4, 14.1, 10.7, 8.2, 7.1, 6.6, and 5.6 kbps.

2.3 Platform #1 -- Laboratory Test
We first conducted the user study in a controlled laboratory environment with a pre-prepared computer and a headphone. In order to recruit participants, we needed to provide a certain reward so that people would be willing to commute to the laboratory and do the test. To this end, we bought an iPhone 4 and held a lucky draw, so every participant had the chance to win the iPhone 4. Limited by the amount of testing equipment and administrative staff, we could not allow multiple participants doing the tests in parallel. We had to schedule the arrival of each participant and run the tests in sequence. Based on the availability information we received from each participant, we scheduled and allocated time slots for them. When the participants came at the assigned time, we first explained the tasks they needed to complete, and then they started to score the quality of each audio file. In the end of the laboratory experiment, we recruited 30 participants in total.

During the scoring process, we first played the original audio source to the participants as a reference. Following by the reference track, all the testing tracks were played in a random order. The scoring standard adopted is the 5-point Mean Opinion Score (MOS) [11] where ‘5’ represents the most desirable quality and ‘1’ represents the least.

2.4 Platform #2 -- Amazon Mechanical Turk
The crowdsourcing platform we chose is Amazon Mechanical Turk [1]. The users of Amazon Mechanical Turk are composed of workers and requesters. The platform allows requesters to publish various types of subjective tests, and workers from all over the world can complete the assigned tests collaboratively. However, the default templates provided by Amazon Mechanical Turk are not suitable for audio quality assessment. To address this problem, authors of [12] made available CrowdMOS, which was a set of scripts and tools supporting subjective multimedia quality assessment via Amazon Mechanical Turk. Being a flexible tool and matches our needs, we used CrowdMOS, to manage our subjective tests on Amazon Mechanical Turk.

The flow of our crowdsourcing test is similar to the laboratory test, but the instructions are all in written form and on the Web-based questionnaire. Two reference tracks with the best quality were inserted. One of them was presented to the participants at the beginning of the test to provide a scoring basis and it is labeled with the best...
quality (i.e., MOS = 5), the other one was a hidden reference which is used to verify if the results from a certain participant is valid.

3. PROCESS

3.1 User Demography

30 people are recruited in the lab experiment. As Figure 2 (a) shows, they age from 19 to 27 years old, and the average age is 23.1 years old. As to the gender distribution, 22 of them are males and 8 females. All of them took the test in a quiet lab room. Each of them listened to the audio tracks using the same set of headphone and computer.

In comparison, we approved 60 workers’ results in the MTurk experiment. As Figure 2 (b) shows, the age ranges from 18 to 55 years old and the average is 29. The age span of the user pool is significantly wider and the average age is higher. The gender distribution is more even. 32 of them are males and 28 females.

One slight difference in the MTurk experiment is the control of listening devices. The users are free to choose whichever listening device they prefer. They, however, are asked to specify the equipment they used, namely desktop speaker, laptop speaker, in-ear headphone and over-the-ear headphone respectively. As the pie chart in Figure 3(a) shows, the distribution of the users amongst different listening devices is quite even. To assess how easy/hard MTurk allows collection of supplementary information about the users, we ask the users to specify how frequent they use Skype. The bar chart in Figure 3(b) illustrates the frequency of the 60 workers using Skype on a regular basis.

3.2 Time

Here we separate the entire experimental process into 5 phases and discuss the differences between the two platforms in terms of time in each of the phases. The time efficiency of these two platforms is summarized in Table 1.

As one can observe from this table, crowdsourcing does save time and effort needed in particularly in user scheduling, test supervising and data collection.

<table>
<thead>
<tr>
<th></th>
<th>Lab Exp</th>
<th>Amazon MTurk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preparation</td>
<td>Same</td>
<td></td>
</tr>
<tr>
<td>Scheduling</td>
<td>2-3 days</td>
<td>none</td>
</tr>
<tr>
<td>Participation</td>
<td>8 minutes</td>
<td>8 minutes</td>
</tr>
<tr>
<td>Supervising</td>
<td>8 mins/submission</td>
<td>1 min/submission</td>
</tr>
<tr>
<td>Data Collection</td>
<td>5 hours</td>
<td>1 hour</td>
</tr>
</tbody>
</table>

Preparation Time: Preparation time is the time that researchers spend on tasks that should be finished before publishing the test. These tasks generally include the production and organization of audio files and the design of scoring questionnaire. The preparation time for both platforms is fairly the same.

Scheduling Time: Scheduling time is the time spent on recruiting people. In the laboratory experiment, it took two to three days before we could schedule the time slots for each participant and start the test. In contrast, there is no scheduling time needed for the MTurk experiment since there are already thousands of workers available on this platform and the tests can be performed in parallel. There is really no need of scheduling the workers for the tests.

Participation Time: By participation time, we refer to the time needed for each user to finish the test. The total number of audio files every user needed to score is 11 (2 reference tracks and 9 testing tracks) and each audio file lasts for 30 seconds. This results in 5.5 minutes in listening and scoring audio files. Considering the time needed for each user to understand the test and fill in their basic information (i.e., age, gender, etc.), the total participation time should be around 8 minutes. The time spent in this phase is about the same for both platforms.

Supervising Time: Supervising time means the time a researcher spends on supervising the users and approving the result. For the laboratory experiment, all the users were invited to the laboratory. The researcher needed to provide...
instructions and administer the process while the users were doing the test. This makes the supervising time equivalent to the participant time per submission. As for the MTurk experiment, the only thing the researcher needed to do is to approve or reject the results submitted by the workers. This process normally involves checking the answers for the basic questions and the score for the hidden reference track. This takes about 1 minute for each submission.

Data Collection Time: We define the data collection time as the period between the arrival of the first and last users. As the estimated participant time is around 8 minutes and we cannot parallelize the tests in laboratory experiment, a reasonable size of time slot assigned to each user would be at least 10 minutes. We can only schedule six users every hour. For the targeted number of users (i.e., 30), it would take 5 hours to finish the data collection. Note that it is the ideal case. The actual data collection time is significantly longer, e.g., weeks, given potential schedule conflicts in users’ availability. As for the MTurk experiment, it took only 1 hour to complete the data collection process for 60 users, which is twice the number of data collected in the lab experiment. The observed data collection time of MTurk experiment is approximately 1 order of magnitude shorter than the lab experiment.

3.3 Cost
In the lab experiment, the only cost is an iPhone 4 for the lottery. The market price is around 670 USD. On the other hand, the questionnaires put on the MTurk are rewarded 1 US dollar for each set of approved results. In addition, MTurk charges a 10% service fee from the requesters. Hence, in this experiment, 66 US dollar is spent in total.

4. RESULTS
Figure 4 depicts the MOS with respects to the bitrates of the audio files. Drawn on the figure for each of the 9 bitrates are the average and the score range plus and minus one standard deviation. From the two curves showing results of lab and MTurk experiment, it is clear that the data sets show the same trend. Although when the bitrate is at the extreme low or high, the average MOS of MTurk experiment is slightly lower that of Lab experiment, the standard deviation ranges overlap quite significantly. This suggests that MTurk is plausible as an alternative experimental platform, and, perhaps equally encouraging, the new MTurk experiment echoes the finding in [3] where the relationship between call bitrate and user experience is shown logarithmic.
ANALYSIS

After showing MTurk experiment being more efficient as an experimental methodology and effective in collecting user opinions, we analyze further the data sets acquired from the lab and MTurk experiment. In particular, we are interested in whether gender or age plays a role in the user scores. We are also interested in comparing the two experimental platforms for finer-grained analysis.

5.1 Gender Factor

Figure 5 shows two plots. The left plot indicates the results from the lab experiment and the right plot from Amazon MTurk. Depicted on the plots are the average MOS by the male vs. female users. We can observe that the average MOS by males vs. females are very close, and the observation holds true for both the lab and MTurk results. This suggests that in terms of user experience gender difference is insignificant. Furthermore, we find that the two experimental platforms are equally effective showing the gender difference for the user study at hand. MTurk experiment is as credible as the lab experiment.

5.2 Age Factor

Figure 6 plots the results from the lab experiment on the left and the MTurk experiment on the right. Average MOS from different age groups are drawn separately for comparison. Generally speaking, the logarithmic relationship holds true across age groups and experimental platforms. There is also no significant difference across age groups. However, it is obvious that the narrow age range in the lab experiment lowers the degree of generality of the statement on age difference. Conversely, MTurk allows a higher variety of users, spanning a wide age range, which shows the advantage of crowdsourcing for analysis in depth.

In particular, in the MTurk result, we observe that, for the ‘10-19’ and ‘50-59’ age groups, there exist a number of odd points. It is due to the numbers of people recruited in these two groups are extremely low, and therefore the curves might not reveal the average that are statistically significant. We do not think this is the shortcoming of crowdsourcing in general, but a problem of this particular MTurk data collection. This, in fact, stresses the need of crowdsourcing towards data analysis of statistical significance.

6. FURTHER ANALYSIS

From the MTurk data, we are able to analyze additional factors such as the Skype usage and the listening device. Here we discuss how these two factors influence the user experience of Skype calls.

6.1 Usage Factor

Figure 7 shows the bitrate-MOS relationship across different usage frequencies. It is observable that the logarithmic trend remain pronounced and is consistent
independent of the usage frequency. This suggests that how frequent the users make calls over Skype would not impact the perceived quality. In other words, we (i.e., the average user) are born with the basic ability to perceive and differentiate the quality of audible speeches.

6.2 Device Factor

Figure 8 shows the bitrate-MOS relationship across users listening via speakers vs. headphones. Again, we do not observe a significant difference between the two groups and the logarithmic relationship remains strong. We have conjectured that headphone users are likely to score higher than the speaker users, but, in contrast, listening device does not make a significant difference to the average perceived quality of audible speeches. What we find particularly interesting is that the logarithmic fit of the headphone user scores is smoother than that of the speaker users. Realized here is that headphones make clear not just the soundness but also the flaws in audible speeches. Hence, the headphone users differentiate better the audio quality. The headphone users do not score higher than the speaker users in general.

7. CONCLUSION

Shown in this study are that MTurk, as a platform for user studies, stands out in the ease of recruiting a high number and diverse set of users. For the Skype call quality analysis, the time consumption and cost are each approximately 1 order of magnitude better. In the meantime, the MOS-bitrate relationship shown from MTurk results echoes that from the lab experiment. Furthermore, having a diverse user pool and being easier to acquire finer information about the users give MTurk an edge in finer-grained analysis.

Note though no one experimental tool would be able to prove any hypothesis. Even with a powerful tool such as MTurk, the researchers are ultimately responsible for the credibility of the studies. The experimental design, i.e., how the tools are used, is crucial. In the case of MTurk experiment, extra cautions are needed, depending on the complexity of the problem at hand, to ensure the worker responses are valid. Speaking from the experience, it helps the response verification process tremendously to embed randomly a small number of cheat-proof tests in a request. On top of that, careful examination of the statistical significance excludes the impact of outliers (i.e., intelligent cheaters) and enhances the technical robustness of the analysis.

8. REFERENCES