CROWDMOS: AN APPROACH FOR CROWDSOURCING MEAN OPINION SCORE STUDIES

Flávio Ribeiro[1], *Dinei Florêncio*[2], *Cha Zhang*[2] *and Michael Seltzer*[2]

[1] Electronic Systems Engineering Department, Universidade de São Paulo, Brazil
[2] Microsoft Research, One Microsoft Way, Redmond, WA, 98052

ABSTRACT

MOS (mean opinion score) subjective quality studies are used to evaluate many signal processing methods. Since laboratory quality studies are time consuming and expensive, researchers often run small studies with less statistical significance or use objective measures which only approximate human perception. We propose a cost-effective and convenient measure called crowdMOS, obtained by having internet users participate in a MOS-like listening study. Workers listen and rate sentences at their leisure, using their own hardware, in an environment of their choice. Since these individuals cannot be supervised, we propose methods for detecting and discarding inaccurate scores. To automate crowdMOS testing, we offer a set of freely distributable, open-source tools for Amazon Mechanical Turk, a platform designed to facilitate crowdsourcing. These tools implement the MOS testing methodology described in this paper, providing researchers with a user-friendly means of performing subjective quality evaluations without the overhead associated with laboratory studies. Finally, we demonstrate the use of crowdMOS using data from the Blizzard text-to-speech competition, showing that it delivers accurate and repeatable results.

***Index Terms*—** crowdsourcing, subjective quality, mean opinion score, MOS, MUSHRA, mechanical turk, crowdMOS.

1. INTRODUCTION

Subjective listening tests are generally regarded as the most reliable and definitive way of assessing audio quality. In the 1990s, several types of listening tests commonly used in telephony were standardized in ITU-T P.800 [1]. The most popular of these is the absolute category rating (ACR) test, in which a volunteer is asked to rate audio files using the discrete 1-5 scale presented in Table 1. This ACR test became the standard for subjective audio quality, and is commonly referred to as the MOS (mean opinion score) test. Since then, other ACR tests were proposed with different scales and for other domains (e.g. [2, 3]).

In general, subjective quality measures require that:

1. there are enough listening subjects of sufficient diversity to deliver statistically significant results;
2. experiments are conducted in a controlled environment with specific acoustic characteristics and equipment;
3. every subject receives the same instructions and stimuli.

The degree to which one follows these requirements strongly determines the accuracy and repeatability of a study. However, the motivation to produce an accurate MOS test is often outweighed by the pressure to make it affordable and practical. Indeed, it is not trivial to assemble panels of pre-screened listeners, and have them rate samples in a laboratory conforming to the recommendations. Thus, the costs associated with laboratory MOS testing often lead researchers to only run informal subjective opinion tests or to use objective quality measures, such as PESQ [4].

Table 1. MOS (ACR) scores

Rating	Quality	Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying and objectionable

Objective quality measures are convenient because they do not have the costs associated with human subjects. On the other hand, each objective measure is only designed to estimate specific quality aspects and is tuned to map to a unique dataset. Thus, an objective measure will only produce predictable results for the listening environment, error conditions and impairments it was developed for. In particular, conventional PESQ was developed for telephone speech applications, and variants are needed for other domains; also, it has been shown to incorrectly estimate MOS in the presence of jitter buffer adjustment, packet loss concealment and other nonlinear processing [5, 6].

In this paper, we propose a class of subjective listening tests which are obtained by relaxing requirement (2) listed above. Instead of running a MOS test in a controlled environment, we outsource this task to workers from an internet crowd, obtaining a measure we call crowdMOS. Several companies provide elements to help facilitate or intermediate this crowdsourcing, with the most famous probably being Amazon Mechanical Turk [7], which we use in our approach. Recruited workers conduct listening experiments at their leisure, using their own hardware. They are typically non-experts drawn from a pool of hundreds of thousands of individuals distributed around the world. Thus, a crowdMOS study can easily have a much larger and more diverse pool than a traditional MOS study, at a significantly lower cost. On the other hand, crowdMOS has to deal with the workers' lack of supervision and uncontrolled environment. Even though we provide financial incentive for workers to deliver high quality results, they have no obligation to do so. Thus, we propose a screening process to detect and discard inaccurate or malicious submissions.

A useful byproduct of crowdMOS is a two-way random effects model for MOS, which we propose to reliably determine confidence intervals (CIs). While performing a literature review, we noticed that it is not widely acknowledged how to compute CIs for MOS. Many studies present no CIs, and others make incorrect assumptions about the independence of scores, producing CIs that are extremely optimistic. With crowdMOS, one can easily and reliably compute CIs for a wide range of experiments.

Even though crowdsourcing has become quite popular for user studies, to our knowledge the previous work involving subjective quality assessment has only focused on paired comparison [8]. Our preference for MOS is three-fold: (i) while paired comparison can be very useful to determine the relative performance of two methods, if the objective is to have absolute scores on an 1-5 scale, then using paired comparisons will require much more data to obtain results

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

with the same statistical significance; (ii) in a MOS test, listeners can be instructed to rate according to given references, thus producing calibrated scores; (iii) MOS is the most popular measure for subjective quality, and we aim to maintain compatibility with previous research. Our proposal is also unique since it provides an open-source set of tools[1], making it more accessible and customizable. For example, we recently extended the crowdMOS tools for image quality assessment [9] and region of interest determination [10].

Our experiments show that crowdMOS is a very reliable method for evaluating subjective audio quality when expert training is not required. By modeling preference variation across users, intrinsic quality variation across test files and subjective uncertainty, crowdMOS provides an accurate measure of statistical significance, which can be combined with the scalability of crowdsourcing to design studies with a desired level of accuracy.

The remainder of this paper is organized as follows: Section 2 shows how we design crowdsourcing experiments to maximize accuracy and worker throughput. Section 3 describes test methodologies, and how scores are modeled and analyzed for statistical significance. Section 4 presents an application using data from the Blizzard text-to-speech challenge [11], in which we compare our crowdsourced subjective measures with Blizzard's laboratory study. Finally, Section 5 has our conclusions and future work.

2. CROWDSOURCING WITH MECHANICAL TURK

2.1. Elements of Mechanical Turk

Amazon Mechanical Turk (MTurk) [7] is a service designed for connecting prospective workers and requesters using a web interface. Jobs are known as human intelligence tasks (HITs), and are typically designed to be very simple. Indeed, most HITs can be completed in under a minute, and workers are rewarded per submitted HIT using a micropayment scheme. Submitted HITs can be rejected, in which case the worker does not get paid. Requesters can also award bonuses to workers who have submitted high quality work.

Each HIT is typically designed to look like a web page, with form elements where the worker can provide answers. For crowdMOS, the HIT contains instructions, followed by audio players which are used to reproduce the samples. Next to each audio player there are controls to enter scores. We typically design HITs to require between one and two minutes of working time, which is in line with most MTurk tasks.

2.2. Effective crowdsourcing

The procedure to run a study using the crowdMOS tools [12] consists of: (1) obtaining a set of files which one wishes to score on a subjective ACR scale; (2) choosing the experiment design, as described in Section 3.1; (3) setting parameters such as HIT reward and bonus; (4) using the tools to automatically create HITs on MTurk; (5) using the tools to retrieve, update, analyze and automatically approve/reject submitted HITs.

In MTurk, workers have thousands of jobs to choose from. Thus, to obtain a significant number of answers, we design HITs with a low entry barrier and provide adequate incentive for participation. This section describes the experiment design used by the crowdMOS tools to maximize the quantity and quality of the answers.

A MTurk requester can require workers to pass a qualification test before working on his HITs. When designing crowdMOS, we experimented using qualification tests, which required workers to score two samples (one obviously good, the other obviously bad) according to instructions and reference samples. This test discouraged most workers from accepting our HITs, since there are numerous other HITs which offer no such barrier. Since it was very easy, it probably did nothing to improve the quality of our results. Thus, crowdMOS uses no qualification tests, and HITs only feature a link to a page with instructions.

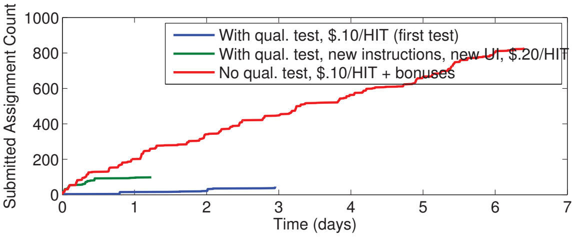

Fig. 1. HIT throughput for different strategies

Workers are often hesitant to participate in subjective studies out of fear their work might be rejected automatically by statistical analysis software. Unfair requesters can easily get a bad reputation, since workers have forums and social networks. Thus, crowdMOS uses very conservative thresholds for automatically rejecting HITs, as described in Section 3.2.

After some experimentation, we designed a user interface to maximize worker throughput. For example, we do not allow submission of incomplete HITs, use radio buttons instead of drop-down boxes, since with radio buttons all options are visible, and selection requires only one click.

We successfully ran studies with rewards ranging from $.05/HIT to $.10/HIT, for HITs requiring about 90 seconds to complete. To motivate users to continue working, we paid a bonus of $.05/HIT if a user submitted more than a given number of HITs. To promote quality, we ranked scores according to their correlation coefficient, and paid a bonus of $.10/HIT if they were in the top 50%, and an additional $.10/HIT if they were in the top 10%. A careful worker could thus potentially earn up to $.30/HIT or around $12/hour, which is a generous payment for MTurk. While it would have been possible to use smaller bonuses and thus lower costs, this would have impacted throughput, which we highly valued for these studies.

Fig. 1 shows the HIT submission throughput for a few experiments. One obtains a dramatic effect by awarding bonuses, having clear instructions and a well designed user interface, as provided by the crowdMOS tools.

3. EXPERIMENT DESIGN AND ANALYSIS

3.1. Experiment methodologies

A given subjective study is meant to detect specific types of impairments, whose intensity must be within the discrimination capacity of the listeners. Current consumer-level equipment easily allows one to evaluate impairments for low and intermediate quality audio, enabling a wide variety of studies. In this section we briefly describe the approaches implemented for subjective evaluation, which draw heavily from existing recommendations.

CrowdMOS is inspired by the ACR test from ITU-T P.800 [1]. An algorithm's crowdMOS is intended to be an absolute measure of subjective quality. It is implemented by having listeners rate samples drawn at random from a pool, under the constraint that the same HIT never contains two samples created from the same test signal (for example, the same utterance). Thus, the listener is not tempted to give relative scores. An instructions page presents Table 1 and also shows examples of files with scores ranging from 1 to 5 for the

[1] Available online at http://research.microsoft.com/crowdmos/.

application of interest (text-to-speech, speech coding, dereverberation, etc.), which are intended to anchor each worker's scores.

We also implement a crowdsourced version of ITU-R BS.1534-1 [3], also known as MUSHRA (multi-stimulus test with hidden reference and anchor). MUSHRA was designed to deliver better discrimination capacity by having listeners always compare samples created from the same test signal. Furthermore, the unprocessed (reference) sample is presented to the user before the processed files, and is labeled with a 5.0 score. An anchor is generated by low-pass filtering the reference, and the test features a hidden copy of the reference and of the anchor hidden among the other samples.

3.2. Score determination and screening

Since MTurk workers are unsupervised, post-screening is needed to validate their answers. Fortunately, workers have little incentive to submit intentionally misleading results, because their percentage of rejected HITs is used as a qualification requirement by almost every requester. Furthermore, a malicious worker can be easily blocked by requesters and automatically banned by MTurk. Thus, workers have a strong incentive to produce large amounts of consistent results.

Nevertheless, screening is important because listeners may not be working in a suitable environment or using appropriate audio hardware. At the beginning of each HIT, the worker is asked if he is using in-ear headphones, over-the-ear headphones, desktop speakers or laptop speakers. Workers using loudspeakers generally have smaller discrimination capacity than users wearing headphones. One cannot require workers to always wear headphones, as there is no way to enforce this requirement. Thus, we only inform workers that it is much easier to deliver high quality results (and receive a bonus) if they wear headphones.

Assume we are evaluating K algorithms using M sentences. All sentences are processed by all algorithms, producing KM samples which are rated by N workers. In laboratory user studies, volunteers can be instructed to always score a precise subset of samples. In crowdMOS, we have no such choice, since HITs are assigned randomly, and a worker can quit at any time. To measure the statistical significance of the result we use 95% confidence intervals (CIs) for the algorithm mean score. To obtain CIs we use the approach described below.

Consider a fixed algorithm of interest whose mean score μ we wish to estimate. Let μ_{mn} be the score given to sentence m by worker n, with $1 \leq m \leq M$ and $1 \leq n \leq N$. To determine CIs for a wide variety of experiments, we use the two-way random effects model given by

$$\begin{aligned} \mu_{mn} &= \mu + x_m + y_n + z_{mn} \\ x_m &\sim \mathcal{N}\left(0, \sigma_s^2\right) \\ y_n &\sim \mathcal{N}\left(0, \sigma_w^2\right) \\ z_{mn} &\sim \mathcal{N}\left(0, \sigma_u^2\right), \end{aligned}$$

where σ_s^2, σ_w^2, and σ_u^2 model the diversity of intrinsic sentence quality, diversity of worker preference, and subjective uncertainty (σ_s^2, σ_w^2 and σ_u^2 depend on the algorithm). To simplify this discussion and the notation, assume there are no missing scores (the crowdMOS tools make no such assumption, and also account for cases when the formulas below degenerate). We then have the estimates

$$\begin{aligned} \hat{\mu} &= \frac{1}{MN} \sum_{n=1}^{N} \sum_{m=1}^{M} \mu_{mn} \\ \hat{\sigma}_w^2 + \hat{\sigma}_u^2 &= \frac{1}{M} \sum_{m=1}^{M} \operatorname{var}\left(\mu_{m,1}, ..., \mu_{m,N}\right) \\ \hat{\sigma}_s^2 + \hat{\sigma}_u^2 &= \frac{1}{N} \sum_{n=1}^{N} \operatorname{var}\left(\mu_{1,n}, ..., \mu_{M,n}\right) \\ \hat{\sigma}_s^2 + \hat{\sigma}_w^2 + \hat{\sigma}_u^2 &= \operatorname{var}\left(\mu_{1,1}, ..., \mu_{M,N}\right). \end{aligned}$$

$\hat{\sigma}_s^2$, $\hat{\sigma}_w^2$ and $\hat{\sigma}_u^2$ can be obtained from the above using a least-squares estimate (if there are missing scores, $\hat{\sigma}_w^2 + \hat{\sigma}_u^2$ and $\hat{\sigma}_s^2 + \hat{\sigma}_u^2$ can be determined by averaging sample variances over smaller blocks of fixed size). An estimate of the mean score variance is given by

$$\operatorname{var}\left[\hat{\mu}\right] = \frac{\hat{\sigma}_s^2}{M} + \frac{\hat{\sigma}_w^2}{N} + \frac{\hat{\sigma}_u^2}{MN}.$$

(Note that the factors $1/M$, $1/N$, and $1/MN$ change if there are missing scores.) To exactly determine the 95% CI for $\hat{\mu}$, one must integrate the PDF of the sum of 3 scaled t-distributed random variables with $M-1$, $N-1$ and $MN-1$ degrees of freedom, which is quite inconvenient to determine. Instead, the crowdMOS tools use a slightly more conservative CI for $\hat{\mu}$ given by

$$\left[\hat{\mu} - t\sqrt{\operatorname{var}\left[\hat{\mu}\right]}, \hat{\mu} + t\sqrt{\operatorname{var}\left[\hat{\mu}\right]}\right],$$

where t is the appropriate percentile from a t distribution with $\min\left(N, M\right) - 1$ degrees of freedom.

To validate this approach, we used an experiment where all the scores were available and compared the obtained CIs with those produced by percentile bootstrap resampling [13] (a non-parametric method), with very similar results. Our approach is much more convenient, since unlike bootstrap resampling, it can be easily extended to work with missing values and does not require a computationally intensive procedure.

Once enough scores have been submitted, the tool can perform a post-screening procedure to automatically approve and reject HITs. To do so, it computes the global MOS values $\hat{\mu}^k$ and the worker MOS values $\hat{\nu}_n^k = \frac{1}{M}\sum_{m=1}^{M} \mu_{mn}^k$, for $1 \leq k \leq K$ and $1 \leq n \leq N$. Let

$$r_n = \frac{\operatorname{cov}\left(\hat{\mu}^1, ..., \hat{\mu}^K; \hat{\nu}_n^1, ..., \hat{\nu}_n^K\right)}{\sqrt{\operatorname{var}\left(\hat{\mu}^1, ..., \hat{\mu}^K\right)}\sqrt{\operatorname{var}\left(\hat{\nu}_n^1, ..., \hat{\nu}_n^K\right)}},$$

which is the sample correlation coefficient between the MOS estimates from worker n and the global MOS estimates. If $r_n < 0.25$ (which is an arbitrarily chosen, conservative threshold), all HITs from worker n are rejected. HITs which were submitted too quickly to have been listened to are also rejected. All r_n values are recomputed for the remaining HITs, and workers are ranked in decreasing order of r_n. Workers are then awarded the bonuses described in Section 2.2.

4. APPLICATION: THE BLIZZARD TTS CHALLENGE

The Blizzard Challenge [11] is an evaluation of corpus-based speech synthesizers. For this experiment, we compared 17 speech synthesis algorithms and a human speech reference, corresponding to Blizzard's EH1 task (English speech generated from a 15 hour dataset). We restricted this comparison to Blizzard's MOSnews listening test, in which users rate the naturalness of synthesized speech for the "news" domain, using a discrete 1-5 scale. The comparison is made using 18 sentences, such that the full set consists of 324 sentences lasting 3-5 seconds each. We compare our results with two Blizzard studies: one generated by paid UK undergraduates in a controlled environment and the other generated by online volunteers (not related to MTurk).

In our studies, users scored 8-10 samples per HIT, and could work until they scored all 324 sentences. Results are presented in

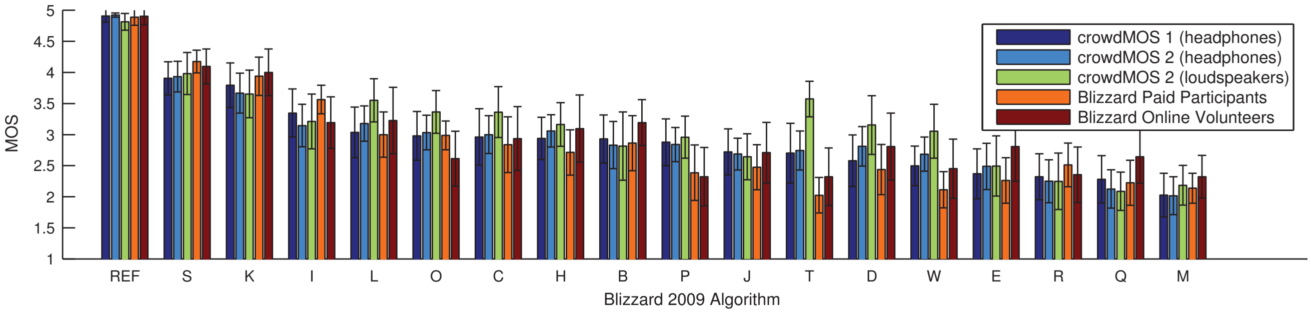

Fig. 2. Comparison between crowdMOS and official Blizzard scores, with 95% confidence intervals

Table 2. Experiment sizes

Experiment	# of Scores	# of Listeners
crowdMOS 1 (headphones)	4410	20
crowdMOS 2 (headphones)	8307	35
crowdMOS 2 (loudspeakers)	3831	17
Blizzard Paid Participants	1440	80
Blizzard Online Volunteers	558	31

Table 3. Correlation with Blizzard Paid Participants

Experiment	r
crowdMOS 1 (headphones)	0.95
crowdMOS 2 (headphones)	0.92
crowdMOS 2 (loudspeakers)	0.78
Blizzard Online Volunteers	0.94

Fig. 2 with 95% confidence intervals. To measure repeatability, we ran two crowdMOS studies (crowdMOS 1 and crowdMOS 2), each with a budget of approximately \$200. Note that this comes out to $\$200/18 \approx \$11/\text{algorithm}$. Thus, typical MOS studies that only compare a few algorithms would be much cheaper to perform. 64% (76%) of the scores in crowdMOS 1 (crowdMOS 2) were generated by workers who did not participate in crowdMOS 2 (crowdMOS 1). Nevertheless, results from the two studies are very closely matched, with a correlation coefficient of 0.99. Scores are also consistent with those from Blizzard's paid workers, except for algorithms T and W.

By listening to the Blizzard dataset, one notices that samples from algorithms T and W distinguish themselves by having a very narrowband sound, but by otherwise having no glaring artifacts or cadence problems. Untrained listeners equipped with commodity hardware should find algorithms T and W quite acceptable. On the other hand, the narrowband characteristic would definitely set them apart if compared using laboratory equipment. Results from crowdMOS 2 from workers listening with loudspeakers support this conclusion. Indeed, their scores for algorithms T and W are significantly higher, indicating that they do not notice the low-pass characteristic and thus do not consider the speech to be excessively unnatural.

5. CONCLUSION

In this paper we described crowdMOS, a crowdsourced measure for subjective audio quality. It was designed to deliver statistically meaningful results with costs which are at least an order of magnitude smaller than laboratory MOS. Indeed, we ran experiments with a cost of about \$10/algorithm.

CrowdMOS applies to studies where impairment can be detected without high-end hardware, and expert training is not required. Thus, it can be used to complement or replace objective quality measures in preliminary quality assessments, where a rigorous laboratory MOS test is not yet justified, yet a subjective assessment is highly desirable. It also has the characteristic (and potential advantage) of having random real world users with commodity hardware in their own environments. Because these are typical users, the differences between laboratory MOS and crowdMOS are likely to highlight what is most important or noticeable to real users.

This work also provides an open source set of tools designed to carry out subjective opinion experiments in a customizable and user-friendly way, completely shielding the researcher from the details surrounding crowdsourcing or Mechanical Turk, and from the bookkeeping required for user studies. We have briefly illustrated the use of these tools to obtain subjective quality measures for the Blizzard Challenge 2009 dataset. They can be easily modified to run other subjective experiments for signal processing, which are the subject of ongoing work.

6. REFERENCES

[1] "Methods for subjective determination of transmission quality," ITU-T Recommendation P.800, Aug. 1996.

[2] "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," ITU-R Recommendation BS.1116-1, Oct. 1997.

[3] "Method for the subjective assessment of intermediate quality level of coding systems," ITU-R Recommendation BS.1534-1, Jan. 2001.

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

[5] S. Pennock, "Accuracy of the perceptual evaluation of speech quality (PESQ) algorithm," *Proc. Of MESAQIN*, 2002.

[6] Z. Qiao, L. Sun, and E. Ifeachor, "Case study of PESQ performance in live wireless mobile VoIP environment," in *Proc. of PIMRC*, 2008.

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

[8] K.T. Chen, C.C. Wu, Y.C. Chang, and C.L. Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *Proc. ACM Multimedia 2009*. ACM, 2009, pp. 491–500.

[9] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing Subjective Image Quality Evaluation," submitted.

[10] F. Ribeiro and D. Florencio, "Region of Interest Determination Using Human Computation," submitted.

[11] S. King and V. Karaiskos, "The Blizzard Challenge 2009," in *Proc. of the Blizzard Challenge*, 2009.

[12] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, "crowdMOS Standalone Tools," available at http://research.microsoft.com/crowdmos/.

[13] B. Efron, R. Tibshirani, and R.J. Tibshirani, *An introduction to the bootstrap*, Chapman & Hall/CRC, 1993.