FAST AND EASY CROWDSOURCED PERCEPTUAL AUDIO EVALUATION

Mark Cartwright [*], *Bryan Pardo*[†]

Northwestern University
{mcartwright@u., pardo@}northwestern.edu

Gautham J. Mysore, Matt Hoffman

Adobe Research
{gmysore, mathoffm}@adobe.com

ABSTRACT

Automated objective methods of audio evaluation are fast, cheap, and require little effort by the investigator. However, objective evaluation methods do not exist for the output of all audio processing algorithms, often have output that correlates poorly with human quality assessments, and require ground truth data in their calculation. Subjective human ratings of audio quality are the gold standard for many tasks, but are expensive, slow, and require a great deal of effort to recruit subjects and run listening tests. Moving listening tests from the lab to the micro-task labor market of Amazon Mechanical Turk speeds data collection and reduces investigator effort. However, it also reduces the amount of control investigators have over the testing environment, adding new variability and potential biases to the data. In this work, we compare multiple stimulus listening tests performed in a lab environment to multiple stimulus listening tests performed in web environment on a population drawn from Mechanical Turk.

Index Terms— audio quality evaluation, crowdsourcing

1. INTRODUCTION

A goal of much research into audio processing and synthesis algorithms (e.g. audio source separation) is to create algorithms that produce output that "sounds good" to a person. In these cases, human perception of quality is the gold standard. The ITU standard methodology for subjective evaluation of audio with "intermediate impairments" (such as in source separation) is ITU-BS.1534-2, a.k.a. MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) [1]. MUSHRA specifies each evaluation should use at least 20 expert participants in a controlled lab setting that adheres to specific acoustical criteria. Since this is difficult and time consuming, researchers often use quick, easy, automated quality measures.

Automated quality measures for audio source separation either were not designed to emulate output human quality evaluations (e.g. BSS-Eval [2]) or were optimized to correlate their output to a fixed set of known human evaluations that are task-specific and limited in size [3, 4]. Objective quality measures for audio codecs that correlate well with human ratings [5, 6] on minor sound degradations are typically not useful on the much larger range of errors and artifacts caused by audio source separation [4].

One can reduce the effort required for conducting a perceptual evaluation of audio by moving from the lab to the web [7, 8, 9, 10, 11]. Web platforms allow automating conducting perceptual studies and recruiting participants on Amazon's Mechanical Turk or LabintheWild [12]. Often, studies with many subjects can be completed in hours (instead of days or weeks for lab studies) with relatively little researcher effort.

There have been web-based media quality evaluation frameworks whose results have been tested [13, 14, 15, 16, 8]. Two of these studies tested perceptual audio evaluations [16, 8], but no study has evaluated MUSHRA listening tests in a web context. Neither has anyone addressed the variability (levels of expertise, listening environments, listening devices, and hearing abilities) introduced when performing MUSHRA-like[1] listening tests on the web. Since MUSHRA is a very popular protocol, it is worth determining the relationship between the results of lab-based and web-based MUSHRA-like evaluations.

In this work, we compare the results from web-based MUSHRA-like listening tests performed on Mechanical Turk to those of MUSHRA performed in a controlled lab environment and to the widely-used objective quality measures used for source separation in BSS-Eval [2]. We also present simple additions to the MUSHRA protocol to account for variability in listening environments. Source separation is our audio task of interest, but our results should be relevant to the evaluation of any audio task for which there are "intermediate impairments" (i.e. significant degradation that can be heard in most listening environments).

2. OVERVIEW OF MUSHRA

MUSHRA is a protocol for the subjective assessment of intermediate audio quality. In a single MUSHRA trial, 3 to 12 stimuli are rated in comparison to a reference and each other on a 0-100 quality scale using a set of sliders. One of

[*] This work was partially performed while interning at Adobe Research.

[†] Thanks to support from NSF Grant 1420971

[1] Some of the specifications of the MUSHRA protocol are not feasible on the web. Therefore, we refer to such tests performed on the web as "MUSHRA-like" tests.

these stimuli is the hidden/unlabeled reference (the desired sound) and at least one other is an unlabeled anchor (a very bad sound). The remaining stimuli are outputs from the systems under test. Since these stimuli are rated in comparison to the reference, we expect the reference to be rated as excellent. Anchors are stimuli designed to be rated as poor. The stimuli and the reference can be played and rated unlimited times in each trial before submission. MUSHRA also specifies the listening environment, training procedure, and participant selection (participants must be experienced, normal-hearing listeners, trained in subjective quality evaluation).

3. THE PEASS DATA SET

The developers of the PEASS Toolkit for automatic source separation evaluation, [4] trained their models on human ratings of source separation algorithm outputs collected using a MUSHRA protocol. Humans rated source separation of 10 mixtures: 5 of speech and 5 of music. All mixtures were 5 seconds long. For each mixture, 8 test sounds were generated: the ground truth target source (the reference), 3 anchors, and 4 outputs, each from one of a variety of source separation algorithms. Participants were 20 normal-hearing experts in general audio applications. Each participant was consented via script and performed a MUSHRA trial for 4 different quality scales on each of the 10 sets of test sounds. All participants listening over the same model of headphones in a quiet environment. The quality scales were labeled *global quality*, *preservation of the target source*, *suppression of other sources* and *absence of additional artificial noises*. We use the same test material (audio stimuli) used to generate the PEASS training data, and compare ratings collected in a web-based listening tests to the ratings collected in the lab-based listening tests in the PEASS data set.

4. CROWDSOURCING MUSHRA-LIKE TESTS

To perform our MUSHRA-like listening test on the web, we used Amazon's Mechanical Turk (AMT) to recruit and pay subjects. In the original PEASS data collection, participants performed 40 different MUSHRA trials, one for each mixture / quality-scale pair—this took each participant several hours to complete. However on AMT, tasks typically take only a few minutes or less to complete [17]. We therefore limited each AMT task to be one MUSHRA trial. We randomly assigned participants to one of the four quality scales, and allowed them to perform up to 10 MUSHRA trials, one for each mixture (randomly ordered).

4.1. Accounting for varied hearing abilities and listening environments

The MUSHRA protocol specifies all participants must listen in the same controlled listening environment. When running tests via AMT, one must account for participants listening in a large variety of environments. Therefore, participants were asked to complete a survey on demographics and listening conditions. We also asked participants to listen over headphones and perform two different hearing tests in addition to the MUSHRA trial. Hearing tests only had to be performed once by any participant regardless of the number of MUSHRA trials they did.

The first hearing test—the *hearing screening*—ensured participants listened over devices with an adequate frequency response (e.g. *not* laptop speakers) and followed instructions (since the answer is objective and known). In this test, participants were first asked to adjust the volume of a 1000Hz sine tone to a comfortable level and to not change the level thereafter. Participants listened to two 8s audio clips and counted how many tones were heard. Each clip contained a 55Hz tone, a 10kHz tone, and between 0 and 6 tones of random frequencies between 55Hz and 10kHz. Tones were 750ms sine waves spaced by 250ms of silence. Tones were scaled to be of approximately equal loudness, and were presented in random order with silence replacing tones when there were less than 8 tones. Participants had two chances to answer correctly.

The second hearing test—the *in situ hearing response estimation*—obtained an overall estimate of hearing thresholds at a range of frequencies, treating the combination of the frequency response of the environment, their hearing, and their listening device as one unit. Participants again listened to audio clips and counted tones of the same duration as the hearing screening. There were eight 12s audio clips in this test. One clip contained only silence. In the remaining clips, the frequency of the sine tones was constant throughout the clip, but the amplitude of the tones varied. The 7 frequencies were log-spaced between 23Hz and 16.8kHz. Each clip contained tones of six 15dBFS-spaced levels (-90 to -15dBFS) and up to 3 additional repeated tones. The remaining time consisted of silent beats. Ordering of tones and beats was random. Based on a participant's tone count for a particular frequency, we determined their in situ hearing threshold at that frequency. If a participant's tone count was 0 for all frequencies or was more than 1 higher than the actual number of tones, their response was marked as rejected.

4.2. Modification to PEASS Instructions

We discovered in a pilot study that some language used in the PEASS MUSHRA instructions was inappropriate for novice participants. To reduce participant confusion, we simplified the instructions from the instructions used in the PEASS data collection. We rephrased and reworded the instructions to be as clear as possible to a novice unfamiliar with source separation. We also added an additional training step for participants in which we played example anchors and references (that were not used in the rest of the study) and informed them of the clips' expected ratings.

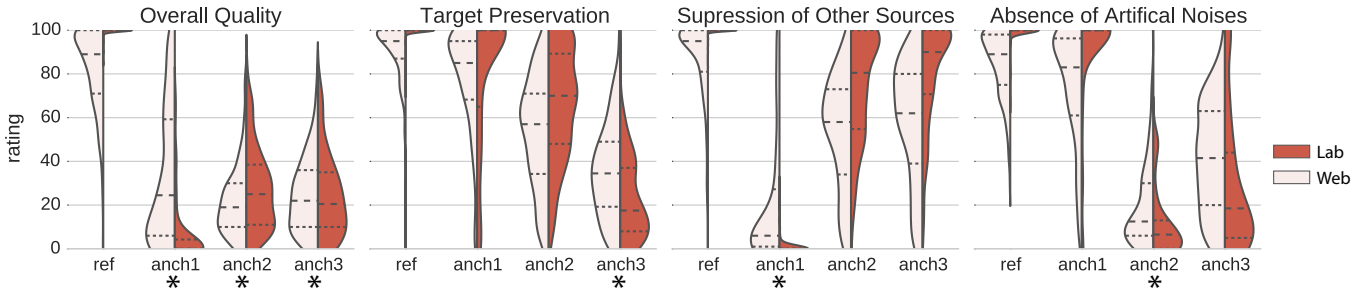

Fig. 1. Distribution of reference (ref) and anchor ratings (anch1-3) for the 4 quality scales pooled over all mixtures. The asterisks below the anchors indicate which are expected to be rated low for that quality scale. The dotted lines are the quartile and median markings. Anchor 1: the sum of all sources (the mixture); Anchor 2: target + 'spectral noise artifacts'; Anchor 3: low-passed target with 20% of time-domain frames missing (see [4] for more details)

.

5. RESULTS

5.1. Data overview

Using participants recruited on Mechanical Turk, we collected *at least* 20 MUSHRA trials for each condition (mixture / quality-scale pair). If we limit the data to participants that passed the hearing screening, there were mean=23.6, max=37 trials per condition. Including those who failed the hearing screening resulted in mean=34.4, max=55 trials per condition. We paid participants $0.80 for the first trial, which included the hearing tests, and $0.50 for subsequent trials. Only people with at least 1000 prior AMT assignments and a 97% approval rate were allowed to participate. In total, we obtained 1763 trials from 530 unique participants, 336 of whom passed the hearing screening. This led to 1147 trials by participants who passed the hearing screening. The mean trials per participant was 3.3 (min=1, max=10). All of the data was collected in 8.2 hours. Note that it would be very difficult to collect data from such a large number of participants in any reasonable time frame in a lab setting.

According to the participant survey, the distribution of reported listening devices was 72% headphones, 16% laptop speakers, 10% loudspeakers. For those who passed the hearing screening, this distribution changes to 84% headphones, 3% laptop speakers, 11% loudspeakers. In addition, 44% of survey respondents reported being able to hear non-test sounds (e.g. environmental sounds) during the test, but only 7% found these sounds distracting. 85% of participants reported this was their first perceptual audio study.

5.2. Ascertaining which scales participants understood

We expect participants who understood the task and quality scale to rate quality scale's anchor(s) (indicated by the asterisks in Figure 1) in the lower half of the scale, and the other quality scales' anchors in the upper half of the scale. We also expect references to be rated very high, near 100.

Figure 1 shows distributions of participant ratings of the reference and anchor sounds from the PEASS data. Each subfigure corresponds to a quality scale (e.g. absence of artificial noises). Within a quality scale there are four violin plots. The left half of each plot gives the distribution of web participants, the right half gives the distribution of lab participants. An asterisk below a plot indicates that stimulus should be rated low, if the participant understands the task.

In general, both web and lab participant distributions of ratings indicate they understand the task, as distributions skew low for stimuli marked by an asterisk and high for the others. The main exception to this is on the *absence of artificial noise* quality-scale, the distribution of *anch3* is centered on the wrong side of the scale for the lab participants. While the use of training examples mitigated this effect for the web participants, the median of the distribution is still below 50. It seems that listeners hear any distortion (subtractive or additive) to the target as simply a distortion. This conflation makes the *target preservation* and *absence of artificial noise* quality scales problematic. We believe using a single quality scale that is inclusive of all distortions (e.g. *lack of distortions to the target*) would eliminate this confusion.

5.3. Weighting participant ratings by hearing response

For the *hearing response estimation* described in Section 4.1, only 10 of the 530 participants' responses were rejected according to the criteria in Section 4.1. Figure 2 shows the mean and standard deviation of the remaining responses and is encouragingly resemblant of minimum audible sound level curves [18]. We combined a participant's resulting threshold curve from the later test with the power spectral densities of the stimuli; thereby creating a weight that is higher when the stimulus contains audible frequency content and lower when it contains inaudible frequency content. To do so, we first subtracted the log hearing threshold (linearly interpolated to N (2048) frequency bins), $\mathbf{H}_k$, of the kth partici-

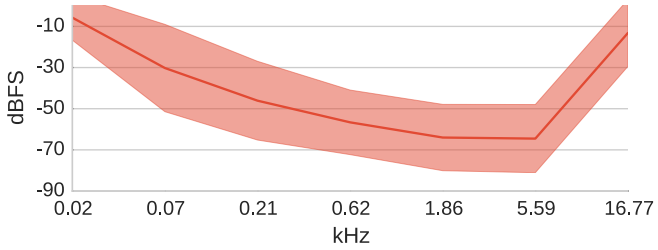

Fig. 2. Mean in situ hearing responses of web participants (N=520). Shaded region is +/- sample standard deviation.

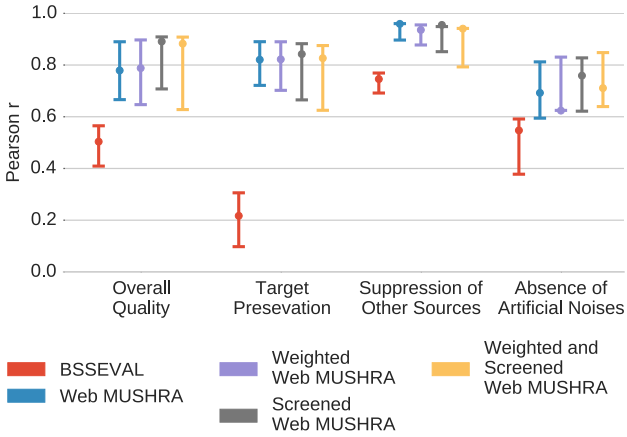

Fig. 3. Pearson correlation of web-MUSHRA and BSS-Eval scores with the lab-MUSHRA scores for the 4 quality scales. Scores were limited to the systems under test (i.e. excluding the reference and anchors) and estimated using the median of ratings from a sample size of 20 participants per mixture. Scores for all mixtures were concatenated before calculating the correlation for each quality scale. Bars represent 95% CIs.

pant, from the log power spectral density, S_m of the mth mixture. This inverse filters the power spectral density, emphasizing the frequencies that the participant can hear well. We then take the log-RMS of this difference to obtain a weight, $w_{m,k} = 20 \log_{10} \sqrt{\frac{1}{N^2} \sum_n^N 10^{(\mathbf{S}_{m,n} - \mathbf{H}_{k,n})/10.0}}$.

5.4. Comparing to the lab-based listening test results

To establish if a Mechanical Turk-based listening test can act as a proxy to a lab-based test, we measured the Pearson correlation between web-MUSHRA scores with lab-MUSHRA scores. As recommended in the MUSHRA standard, we used the median to aggregate participants' ratings into stimulus scores. Figure 3 details the process and displays 95% confidence intervals and point estimates of the correlation. The weighted MUSHRA scores were calculated using a weighted median with the weights described in Section 5.3. The screened MUSHRA scores were calculated only with responses from participants who passed the hearing screening. From the figure, we see that scores calculated from MUSHRA-like tests on the web correlate well to lab-MUSHRA scores (Overall Quality: r=0.78; Target Preservation: r=0.82; Suppression of other Sources: r=0.96; Absence of Artificial Noise: r=0.69). For comparison, we also correlated the corresponding the BSS-Eval objective measures (SDR, ISR, SIR, and SAR [4])—all of which were less correlated to the lab scores than the web scores were.

Neither the hearing screening nor the hearing response weighting seem to affect the scores for these stimuli. A possible explanation for this is that the source separation algorithms under test have such easily detectable impairments that they can be heard and assessed in both good and poor listening conditions. Therefore, for the remainder of the paper we will simply focus on the web-MUSHRA results without the hearing test filtering or weighting.

Regardless of the correlation, it may be that scores calculated from MUSHRA-like tests on the web are noisier than those of lab-based MUSHRA tests. We calculated the widths of the 95% confidence intervals (calculated via bootstrapping with 1000 iterations) for the scores of the systems under tests for all four quality scales and pooled them together into one distribution for each MUSHRA type (lab and web). Since there were 20 lab participants, we limited ourselves to only using the first 20 web participants. The distributions of the CI widths for the web and lab scores are quite similar with almost identical sample means (web: 17.5, lab: 17.1) and with the standard deviation of the web score widths actually a bit smaller than the lab score widths (web: 3.9, lab: 5.2). This implies that scores from MUSHRA-like tests on the web are not noiser than those from MUSHRA in the lab, and that 20 web participants may be adequate to obtain scores of comparable confidence as that of 20 lab participants.

6. CONCLUSION

We compared MUSHRA performed in a controlled lab environment to a MUSHRA-like test performed in an uncontrolled web environment on a population drawn from Mechanical Turk. The web data was collected from 530 participants in only 8.2 hours. The resulting perceptual evaluation scores were comparable to those estimated in the controlled lab environment. Two procedures were proposed to account for varied hearing abilities and listening environments, but these procedures did not improve correlations with the lab data for our stimuli. This could be due to the size of impairments in the output of source separation algorithms, which may be easily detectable in both good and poor listening conditions. While such hearing tests may prove useful on more subtly different stimuli, it is encouraging that these additional tests may not be necessary for many types of stimuli.

7. REFERENCES

[1] ITU, "Recommendation ITU-R BS.1534-2: Method for the subjective assessment of intermediate quality level of audio systems," 2014.

[2] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[3] Brendan Fox, Andrew Sabin, Bryan Pardo, and Alec Zopf, *Modeling perceptual similarity of audio signals for blind source separation evaluation*, pp. 454–461, Springer, 2007.

[4] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[5] Thilo Thiede, William C Treurniet, Roland Bitto, Christian Schmidmer, Thomas Sporer, John G Beerends, and Catherine Colomes, "PEAQ—the ITU standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.

[6] R. Huber and B. Kollmeier, "PEMO-Q: A new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[7] Sebastian Kraft and Udo Zölzer, "BeaqleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality," in *Proceedings of the Linux Audio Conference, Karlsruhe, DE*, 2014.

[8] F. Ribeiro, D. Florencio, Zhang Cha, and M. Seltzer, "CROWDMOS: An approach for crowdsourcing mean opinion score studies," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 2416–2419.

[9] Michael Schoeffler, Fabian-Robert Stöter, Harald Bayerlein, Bernd Edler, and Jürgen Herre, "An experiment about estimating the number of instruments in polyphonic music: A comparison between internet and laboratory results," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2013.

[10] Michael Schoeffler, Fabian-Robert Stöter, Bernd Edler, and Jürgen Herre, "Towards the next generation of web-based experiments: A case study assessing basic audio quality following the ITU-R recommendation BS. 1534 (MUSHRA)," in *Proceedings of the Web Audio Conference*, 2015.

[11] Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *Proceedings of the ACM International Conference on Multimedia*. 2009, ACM.

[12] Katharina Reinecke and Krzysztof Z. Gajos, "LabintheWild: Conducting large-scale online experiments with uncompensated samples," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 2015, ACM.

[13] O. Figuerola Salas, V. Adzic, and H. Kalva, "Subjective quality evaluations using crowdsourcing," in *Proceedings of the Picture Coding Symposium (PCS), 2013*, 2013.

[14] Oscar Figuerola Salas, Velibor Adzic, Akash Shah, and Hari Kalva, "Assessing internet video quality using crowdsourcing," in *Proceedings of the Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*. 2013, ACM.

[15] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "QualityCrowd: A framework for crowd-based quality evaluation," in *Proceedings of the Picture Coding Symposium (PCS), 2012*, 2012.

[16] Chen Kuan-Ta, Chang Chi-Jui, Wu Chen-Chi, Chang Yu-Chun, and Lei Chin-Laung, "Quadrant of euphoria: a crowdsourcing platform for QoE assessment," *Network, IEEE*, vol. 24, no. 2, pp. 28–35, 2010.

[17] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis, "Running experiments on Amazon Mechanical Turk," *Judgment and Decision making*, vol. 5, no. 5, pp. 411–419, 2010.

[18] Harvey Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *The Journal of the Acoustical Society of America*, vol. 5, no. 2, pp. 82–108, 1933.