FAST AND EASY CROWDSOURCED PERCEPTUAL AUDIO EVALUATION

Mark Cartwright and Bryan Pardo — Northwestern University
Gautham Mysore and Matt Hoffman — Adobe Research

mcartwright@u.northwestern.edu   http://music.eecs.northwestern.edu

Introduction

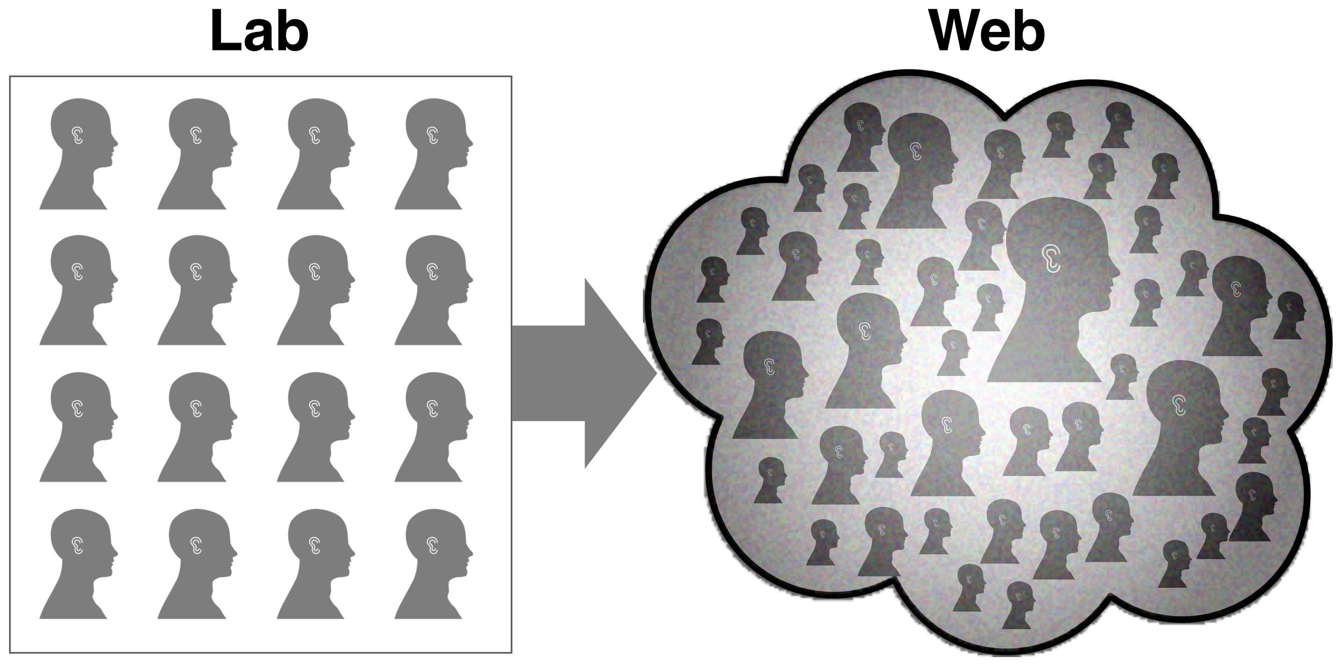
A goal of much research into audio processing and synthesis algorithms (e.g. audio source separation) is to create algorithms that produce output that “sounds good” to a person. In these cases, human perception of quality is the gold standard. Current methods for audio evaluation either require a lot of effort by the investigator or are poorly correlated to human judgments of quality. We need a method of evaluating audio quality that is both accurate and easy for investigators to perform. We propose to move listening tests from the lab to the web, and we compare our web-based results to gold-standard lab-based results.

Current Audio Evaluation Methods

	Pros	Cons
Listening tests e.g. participants listen to and rate audio stimuli	• “If it sounds good, it is good” • No ground truth required	• Slow • Expensive • Require a lot of effort by the investigator
Signal measures e.g. machines estimate audio quality based on signal properties	• Fast • Cheap • Require little effort by the investigator	• Poorly correlated to human judgments of quality • Require ground truth • Difficult to develop new measures

Our Approach

Crowdsource listening tests by moving them from the lab to the web in order to reduce the effort required by the investigator



Potential Benefits	Challenges
• Speed • Minimal effort for investigator • Human judgments of quality • Large, diverse population of participants • Can easily customize evaluation measures	• Varied reliability of assessors • Varied levels of expertise • Varied listening environments • Varied listening devices • Varied hearing abilities

Evaluating Our Approach

We conducted MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) [1] listening tests on the web, recruiting from Amazon’s Mechanical Turk, comparing results to MUSHRA listening tests conducted in a lab setting and also to the BSS EVAL signal measures of audio quality.

Task: source-separation quality evaluation
- 4 quality-scales:
  1. *Overall Quality*
  2. *Preservation of the Target Source*
  3. *Suppression of Other Sources*
  4. *Absence of Artificial Noises*
- 10 mixes
- 4 systems under test per mix

Test-type: MUSHRA
- ITU standard for the subjective assessment of intermediate quality
- 8 stimuli presented simultaneously
- Stimuli rated on 0 – 100 scale
- *Target* and *Mixture* as references
- *Target* as hidden reference
- 3 anchors

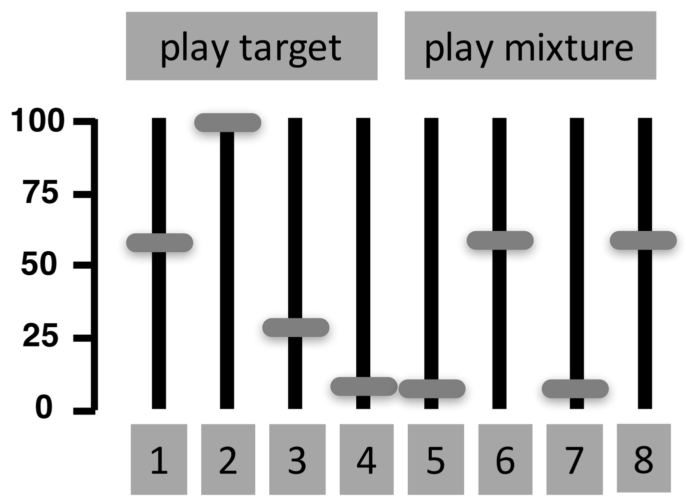


Figure 1. Example MUSHRA interface. Multiple Stimulus Hidden Reference and Anchor

Lab MUSHRA Listening Test	Web MUSHRA Listening Test
• Data from PEASS [2] • Expert participants • Controlled lab environment • Number of participants: 20 • Trials per participant: 40 (10 mixes x 4 qualities)	• Mostly novice participants recruited from Amazon’s Mechanical Turk • Varied listening environments • Number of participants: 530 • Trials per participant: mean=3.3 (min=1, max=10) • Data collected in 8.2 hours

Accounting for Hearing Abilities and Listening Environments in Web MUSHRA

1. **Hearing screening**
   - Screen to participants that hear 55 - 10000Hz by using a simple tone-counting task
   - 336 of 530 passed
2. **Weight importance of participant rating**
   - Roughly estimate participants’ in-situ hearing response using simple tone-counting task (see Figure 2)
   - Use hearing response to weight the importance of their rating. Weight is higher when the stimulus contains frequency content they hear well and lower when it contains frequency content they hear poorly.

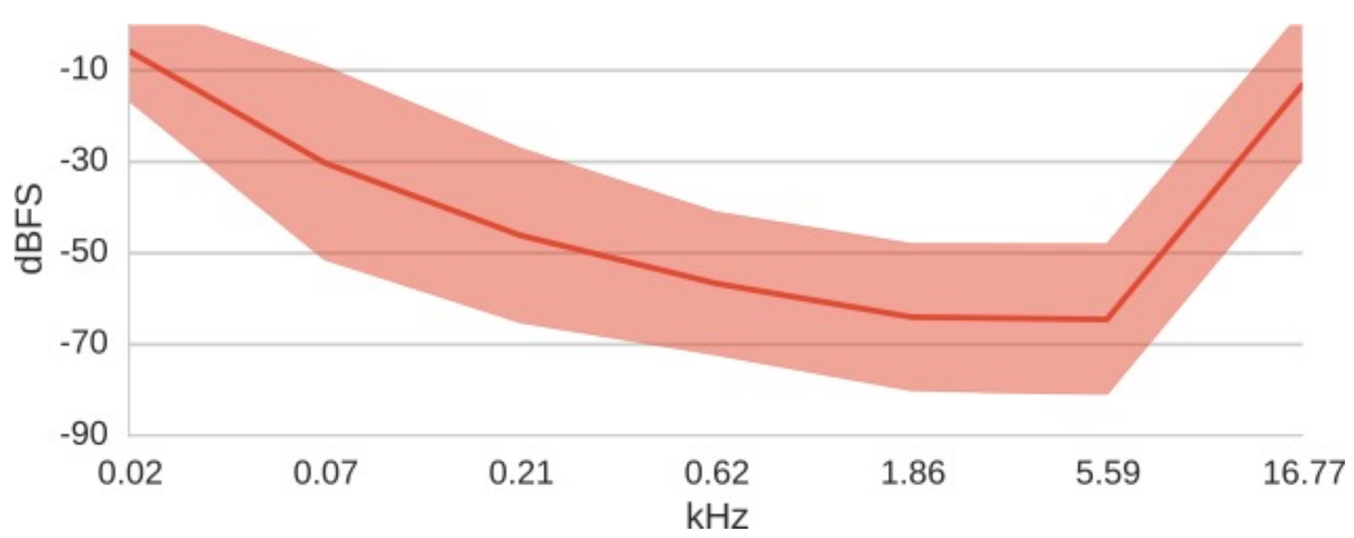


Figure 2. The mean in-situ hearing response of the participants. The lighter band is +/- SD.

Results

1. How do web MUSHRA scores correlate with lab MUSHRA scores?

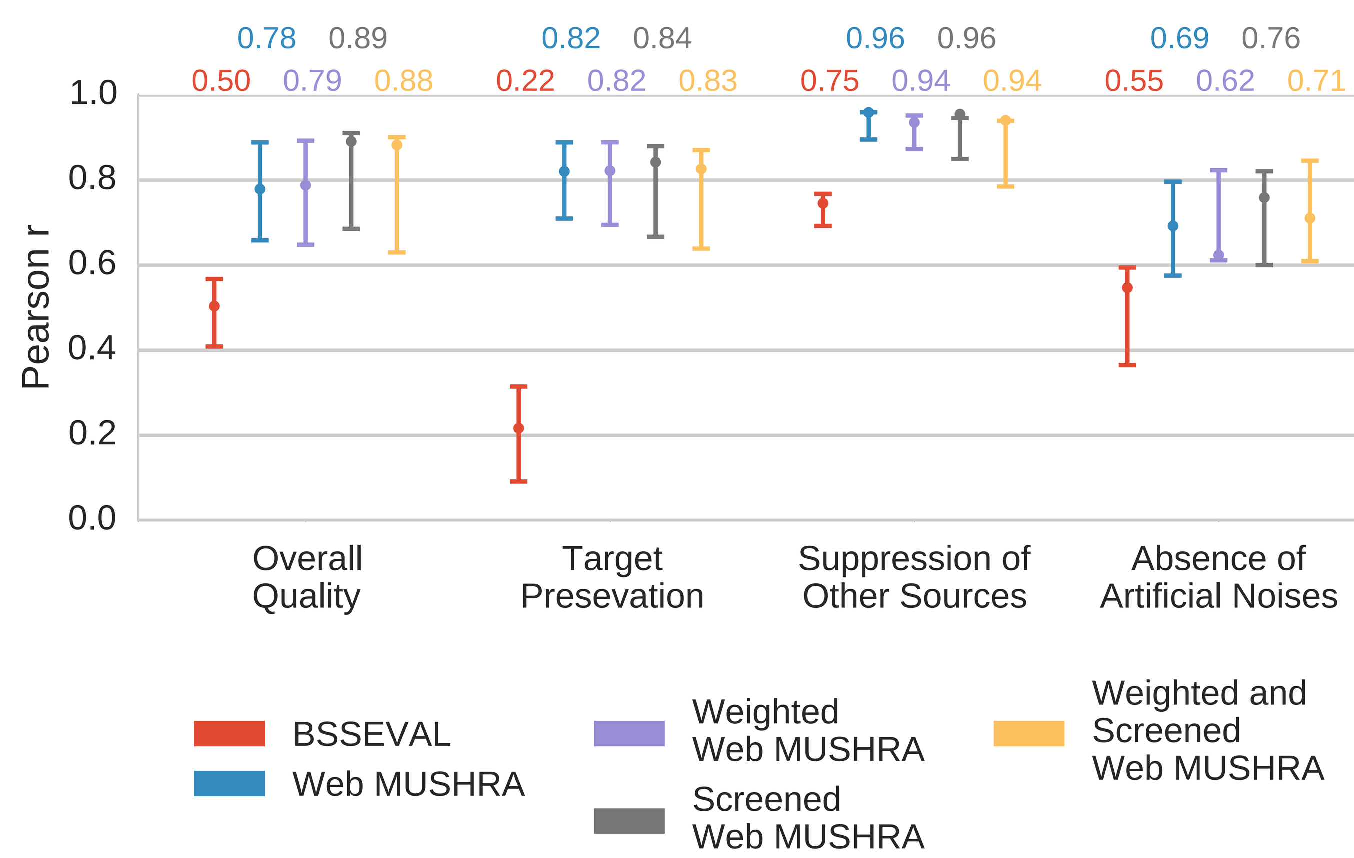


Figure 3. Correlation of web MUSHRA and BSS-Eval [2] scores with the lab MUSHRA scores for the 4 source separation quality scales. Scores were limited to the systems under test (i.e. excluding the reference and anchors) and estimated using the median of ratings from a sample size of 20 participants per mixture. Bars represent 95% CIs.

2. Are web MUSHRA scores “noisier” than lab MUSHRA scores?

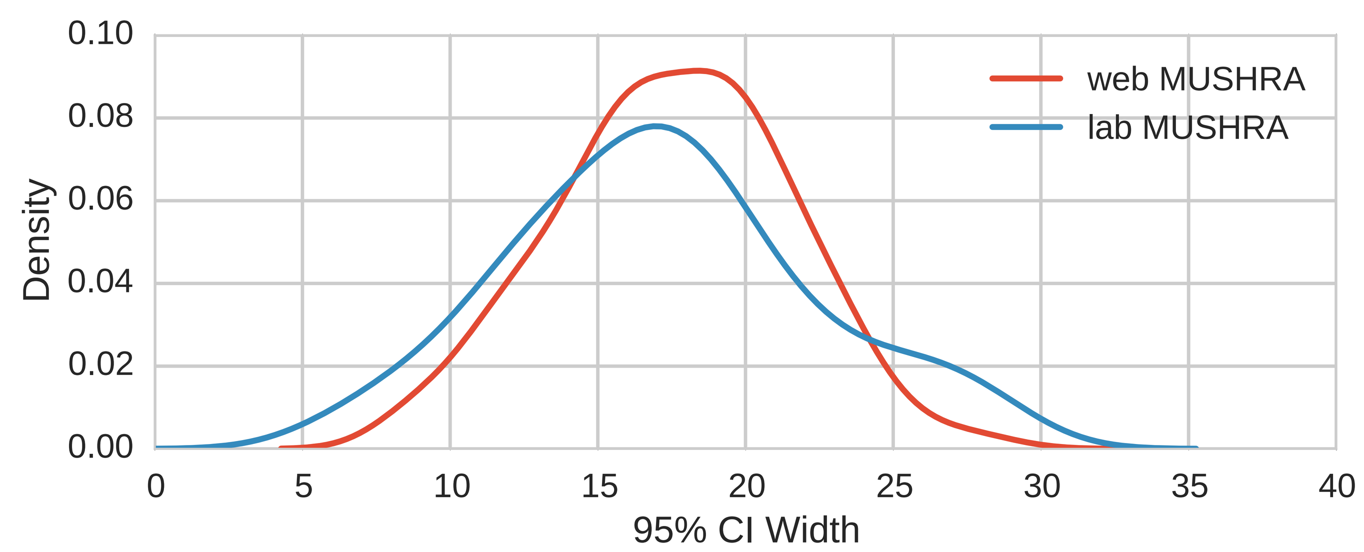


Figure 4. Distribution of 95% CI widths of scores.
Web MUSHRA: mean=17.5, SD=3.9
Lab MUSHRA: mean 17.1, SD=5.2

Conclusions

We compared MUSHRA listening tests performed in a controlled lab environment to MUSHRA 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.

References

[1] ITU, "Recommendation ITU-R BS.1534-2: Method for the subjective assessment of intermediate quality level of audio systems," ed, 2014.
[2] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and Objective Quality Assessment of Audio Source Separation," *IEEE TASLP*, vol. 19, pp. 2046-2057, 2011.

Acknowledgements
This work was supported by NSF Grant 1420971 and Adobe Research.