

# REVERB Subjective Evaluation Guidelines

The REVERB subjective evaluation is based on a modified version of MUSHRA, defined by ITU-R recommendation BS.1534-1, where the modifications are intended for allowing the method of MUSHRA to be reasonably applied to the "crowdsourced" evaluation of the audio quality of the reverberant and dereverberated sound excerpts provided by the REVERB Challenge. The whole evaluation consists of a number of evaluation sessions, each performed by different evaluators. By clicking the "Start" link, you will be redirected to a web page showing one randomly chosen session of the evaluation. Each session consists of two phases: one for training and one for grading (Fig. 1 illustrates the concept of the sessions). The following sections describe the objectives, procedures, and instructions for the training and grading phases. *As an evaluator, you are requested to carefully read these guidelines before moving forward to the evaluation sessions to understand how the sessions will proceed and what you will be asked to do.*

## 1. Basics

*In the same way to MUSHRA, the evaluation framework adopted in the REVERB subjective evaluation is based on comparisons of a reference sound excerpt (i.e., a clean or headset recording) and a number of test sound excerpts that are distorted (i.e., reverberant or dereverberated, possibly mixed with additive noise) versions of the reference.* The following two attributes are used for subjective evaluation of the audio quality.

- Perceived amount of reverberation: This attribute is related to perceptual impressions about the degrees to which the reference and test sound excerpts are reverberant.
- Overall quality: This attribute is used to judge any and all detected differences (in terms of characteristics of reverberation, additive noise, and processing distortion, timbral characteristics, naturalness, and so on) between the reference and test sound excerpts.

During the evaluation, you are requested to follow the following instructions as an evaluator.

- During the evaluation, you should not discuss your personal interpretation of the audio quality with any other evaluators at any time.
- For the evaluation, you are required to use a sensibly high quality headphone to listen to the provided sound excerpts. Earphones and loudspeakers should not be used.
- During the evaluation, you should not use any buttons of your browser, such as "Back" and "Refresh".

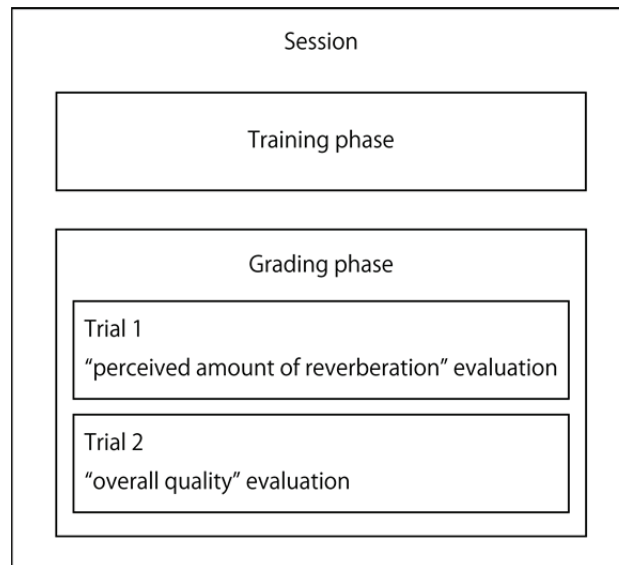


Figure 1: Structure of session

## 2. Training Phase

The training phase is intended to allow you to grasp the whole range of the possible qualities and distortions that you will experience in the grading phase. In this phase, you are asked to listen to all sound excerpts that have been selected for the session assigned to you. This is important to allow subjective ratings from different evaluators to be reasonably used for statistical analysis. Figure 2 shows the user interface of the training system. You may click on different buttons to listen to different sound excerpts including the reference excerpts. In this way, you can appreciate the full range of quality levels and distortion types.

## Training phase

Please listen to all sound excerpts presented in this page until you will feel you are familiarized with the qualities of those sound excerpts.

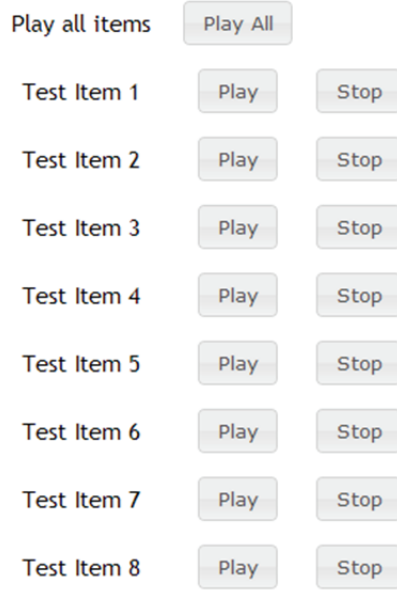


Figure 2: User interface for training phase (snip).

### 3. Grading Phase

The purpose of the grading phase is to invite you, as an evaluator, to assign your subjective ratings to the sound excerpts that have been selected for the evaluation. Your ratings should reflect your subjective judgment of the quality level for each of the sound excerpts presented to you.

The grading phase of each session consists of two trials: one for perceived amount of reverberation evaluation and one for overall quality evaluation. Figures 3 and 4 show the user interfaces of the grading systems used for the perceived amount of reverberation evaluation and the overall quality evaluation, respectively. Both trials include the same test sound excerpts. Each session contain 13 or fewer signals to be graded. The number of signals presented to you depends on the session assigned to you. Each of the sound excerpts is approximately 10 s long. *You are required to listen to the reference and all test sound excerpts by clicking on the respective buttons. You may listen to the test materials in any order, any number of times. Your rating of the audio quality of each test*

*sound excerpt should be indicated with the respective sliders. When you are satisfied with your ratings of all the test materials, click on the “Next Test” button at the bottom of the screen.*

The grading scale is continuous from "Very large" to "Very small" in the perceived amount of reverberation evaluation and from "Bad" to "Excellent" in the overall quality evaluation. ***In evaluating the sound excerpts, one or more excerpts must be given a rating of the high end of "Very small" (or "Excellent") because the clean or headset reference signal is included as one of the excerpts to be graded.*** This is important for post-screening the results since each of the evaluators, being collected from the crowd, may have very different knowledge and expertise in listening evaluation. The results from the evaluators misidentified the reference signal will not be used for analysis.

In the evaluation in terms of perceived amount of reverberation, you do not need to give a rating in the bottom category to the sound excerpt with the lowest quality, although the excerpts to be graded include the un-processed reverberant signal. In the evaluation in terms of overall quality, a 3.5 kHz low-pass filtered version, called an anchor, of the un-processed reverberant signal is also included in the test excerpts. In MUSHRA, the test excerpt judged to be the anchor is commonly graded around the boundary between "Bad" and "Poor". Note that you can rate the other excerpts below or above the anchor.

## Evaluation in terms of "perceived amount of reverberation"

### [Important notes]

- Please give the maximum rating (i.e., right edge of "Very small") to the item that you judge to have the same quality as the reference.
- Please rate the other items based on your own impression.
- Note that the unprocessed microphone signal is included. However, you do NOT need to give the lowest rating (i.e., left most) to this item.

| Reference   | <input type="button" value="Play"/> | <input type="button" value="Stop"/> | Very large            | Large | Mid. | Small | Very small |
|-------------|-------------------------------------|-------------------------------------|-----------------------|-------|------|-------|------------|
| Test Item 1 | <input type="button" value="Play"/> | <input type="button" value="Stop"/> | <input type="range"/> |       |      |       |            |
| Test Item 2 | <input type="button" value="Play"/> | <input type="button" value="Stop"/> | <input type="range"/> |       |      |       |            |
| Test Item 3 | <input type="button" value="Play"/> | <input type="button" value="Stop"/> | <input type="range"/> |       |      |       |            |
| Test Item 4 | <input type="button" value="Play"/> | <input type="button" value="Stop"/> | <input type="range"/> |       |      |       |            |
| Test Item 5 | <input type="button" value="Play"/> | <input type="button" value="Stop"/> | <input type="range"/> |       |      |       |            |
| Test Item 6 | <input type="button" value="Play"/> | <input type="button" value="Stop"/> | <input type="range"/> |       |      |       |            |
| Test Item 7 | <input type="button" value="Play"/> | <input type="button" value="Stop"/> | <input type="range"/> |       |      |       |            |
| Test Item 8 | <input type="button" value="Play"/> | <input type="button" value="Stop"/> | <input type="range"/> |       |      |       |            |

Figure 3: User interface for perceived amount of reverberation evaluation (snip).

## Evaluation in terms of "overall quality"

### [Important notes]

- Please give the maximum rating (i.e., right edge of "Excellent") to the item that you judge to have the same quality as the reference.
- Here, a 3.5kHz low-pass filtered version of the unprocessed microphone signal is included as one of the test materials (called an hidden anchor).
- The item you judge to be the hidden anchor is commonly graded around the border between "Bad" and "Poor".
- The other test items can be graded below or above the anchor based on your own impression.

| Reference   | Play | Stop | Bad                   | Poor | Fair | Good | Excellent |
|-------------|------|------|-----------------------|------|------|------|-----------|
| Test Item 1 | Play | Stop | <input type="range"/> |      |      |      |           |
| Test Item 2 | Play | Stop | <input type="range"/> |      |      |      |           |
| Test Item 3 | Play | Stop | <input type="range"/> |      |      |      |           |
| Test Item 4 | Play | Stop | <input type="range"/> |      |      |      |           |
| Test Item 5 | Play | Stop | <input type="range"/> |      |      |      |           |
| Test Item 6 | Play | Stop | <input type="range"/> |      |      |      |           |
| Test Item 7 | Play | Stop | <input type="range"/> |      |      |      |           |
| Test Item 8 | Play | Stop | <input type="range"/> |      |      |      |           |

Figure 4: User interface for overall quality evaluation (snip).