Audio Engineering Society

Convention e-Brief 151

Presented at the 136th Convention
2014 April 26–29 Berlin

*This Engineering Brief was selected on the basis of a submitted synopsis. The author is solely responsible for its presentation, and the AES takes no responsibility for the contents. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the* Audio Engineering Society.

# APE: Audio Perceptual Evaluation toolbox for MATLAB

Brecht De Man, Joshua D. Reiss

*Centre for Digital Music, Queen Mary University of London, UK*
*b.deman@qmul.ac.uk*

**ABSTRACT**

We present a toolbox for multi-stimulus perceptual evaluation of audio samples. Different from MUSHRA (typical for evaluating audio codecs), the audio samples under test are represented by sliders on a single axis, encouraging careful rating, relative to adjacent samples, where both reference and anchor are optional. Intended as a more flexible, versatile test design environment, subjects can rate the same samples on different scales simultaneously, with separate comment boxes for each sample, an arbitrary scale, and various randomisation options. Other tools include a pairwise evaluation tool and a loudness equalisation stage. We discuss some notable experiences and considerations based on various studies where these tools were used. We have found this test design to be highly effective when perceptually evaluating qualities pertaining to music and audio production.

## 1. INTRODUCTION

The subjective assessment of audio, where multiple samples are compared against each other, requires a well thought-out procedure and a suitable interface. To allow for double-blind testing (where biases carried by both the investigator and the test participants are eliminated), a digital interface is by far the most effective as it supports easy randomisation of the order of stimuli and the order of sets of stimuli.

Various listening test interface designs for different purposes are available, such as MUSHRA, which was designed for the assessment of audio codecs and is therefore suited for investigating the perceptual effects of undesired sonic artefacts [1]. The Audio Perceptual Evaluation (APE) toolbox described herein has been developed for and shaped by research into music mixing practices, audio effects and more generally audio processing of identical or similar, high quality source material for musical purposes, see Section 2.

MUSHRA interfaces have a separate slider per audio sample under test, for instance, as implemented in MUSHRAM [2]. In contrast, APE features just one axis (for every quality to be assessed) where markers, corresponding with different audio samples, can be placed according to the subject's assessment of a certain sonic quality. In the experience of the authors, this encourages careful relative placement of each sample, as opposed to a rating of every individual sample against a single reference in the case of a MUSHRA test or similar. Beyond that, we support all recommendations regarding interface design in the MUSHRA specification [1], but also allow for many other options, see Section 3.

Furthermore, we include a set of other tools, such as a pairwise comparison interface (A/B testing) used in [3], a batch loudness equalisation and error checking tool, among others, see Section 4.

Implementations of other types of listening test interfaces for MATLAB exist, including semantic differential, single stimulus rating, triple stimulus, n alternative forced choice (n-AFC), repertory grid technique (RGT) and many more [2, 4–6].

## 2. STUDIES

Five perceptual evaluation studies where this tool was used are considered: a perceptual evaluation study of different microphones using both pairwise and multi-stimulus test designs [3], a validation of a knowledge-engineered autonomous mixing system [7], an evaluation of an adaptive distortion algorithm [8], and two more studies where different mixes performed by different mixing engineers are compared. These studies helped inform and shape the design of the toolbox. Data from these studies is also shared to serve as examples.

In [7], where different mixes of the same multitrack content where compared, the 'hidden anchor' provided was an unprocessed, monophonic sum of normalised audio. Without the requirement to rate any of the samples below a certain value, it seemed that the supposedly low quality anchor was not at the bottom of the ratings of some subjects, for some sets of samples. This confirmed that for perceptual evaluation of this kind, it is ineffective to require one of the samples to be rated very low, although it can most definitely be interesting to include a purposely low quality sample. Similarly, in [8], 3 listening test participants' results were omitted as they did not understand the assignment well enough, rating the 'hidden reference' (still not automatically checked during the test) very low for one of the scales. Had these participants been automatically notified of the need to rate one of these stimuli at maximum value, it could have been impossible to discriminate these participants from others. To account for these situations as well as a test more closely following the MUSHRA standard or similar, we choose to provide the option, that can easily be included or omitted, to have both hidden and visible references and anchors with or without the requirement to rate one sample at maximum and/or minimum value.

Between the two last tests (unpublished), participants were able and encouraged to comment using first a single comment field (with numbers '1:' through '10:' already present), and a separate comment field per stimulus (10 in total). In the latter case, the comment rate (percentage of filled out comment boxes) was 96.5% (99.8% when two participants weren't included), as opposed to 82.1% in the former case for the same participants ($N = 13$). , and comments were also 47% longer (88.3 rather than 60.0 characters per comment on average). The two tests were near identical otherwise, with different but similar samples.

## 3. DESIGN REQUIREMENTS

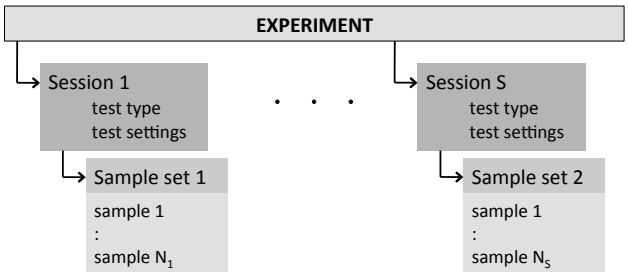



Fig. 1: Structure of a listening test.

Consider an experiment consisting of $S$ different listening sessions, where every session $s$ corresponds with a listening test interface with certain settings, and a set of $N_p$ audio samples (see Figure 1) [9]. We will only discuss the requirements pertaining to listening test design, and not consider those related to subject selection, content selection, or data processing and presentation.

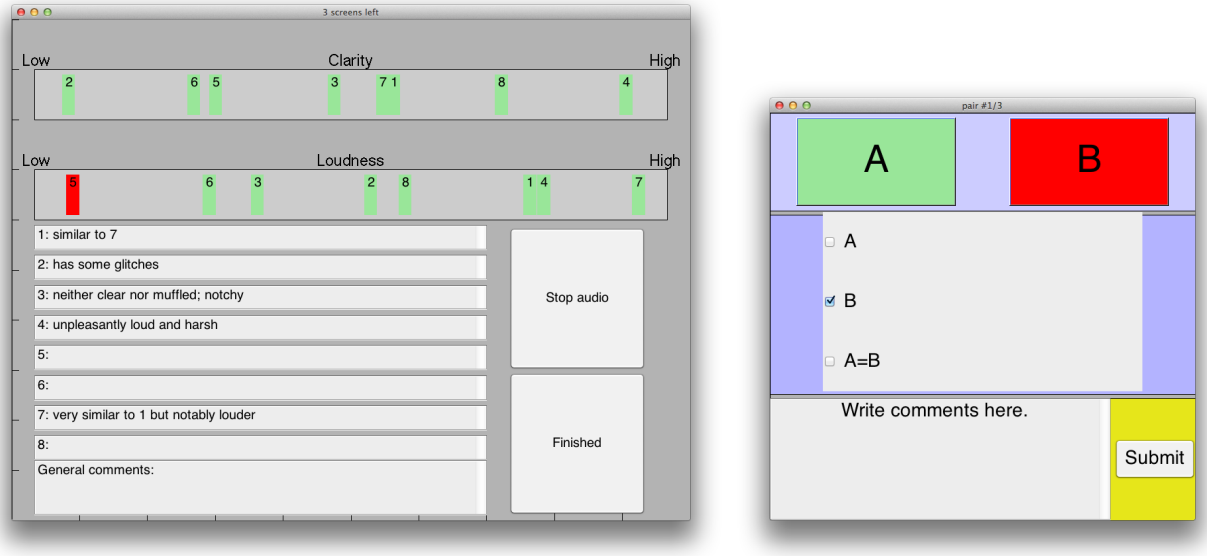
To accommodate existing recommendations, previous needs and foreseeable, similar experiments, we want a listening test interface that offers the following options:

**Randomisation** To minimise bias, the order of $P$ sample sets and order of samples within each sample set should be randomised [1, 6]. We make this optional to support more exotic test designs where e.g. the samples should be auditioned in a specific order.

**Reference and anchor** Provide the possibility of a fixed reference and/or anchor sample with fixed rating, and optional requirement to rate at least one (or more) samples above or below certain value ('hidden reference'/'hidden anchor' [1])

**Multiple scales** To rate and describe different characteristics (e.g. the parameters recommended in [10]), any number of scales can be provided.

**Rating scale layout** We support a quality scale with intervals *Bad*, *Poor*, *Fair*, *Good*, and *Excellent*, which like MUSHRA, corresponds with recommendation




(a) Multi-stimulus (b) Pairwise

Fig. 2: An example of what the multi-stimulus interface with 3 scales and 8 stimuli (no visible reference or anchor) and the pairwise interface with ternary response could look like.

ITU-R BT.500 [1, 11], but other scale layouts are possible as well.

**Log extra data** For further research, or to justify the exclusion of certain participants, we keep track of the time spent on each session and of the order in which the samples were played back.

**Comments** Participants can comment on each sample, and a general comment field is also provided for comments about the test design, the participant's experience or a shared characteristic of the samples.

**Autosave** Participants can quit test at arbitrary times, if permitted, and resume later without losing data. This also helps in the event of an unexpected error.

**Loudness equalisation** To minimise bias towards (or away from) louder samples, the loudness of all samples should be equal for most test designs. As an alternative to setting levels by an expert panel [1], gain can automatically be applied to the samples so that the loudness is close to equal.

**Detection of identical samples** As samples will often sound very similar, it will be hard to notice that the exact same sample is unintentionally included twice.

**Memory** We provide the choice between faster sample recall (load all samples in memory) and a smaller memory footprint (load each sample at playback).

## 4. TOOLBOX CONTENT

### 4.1. Multiple stimulus evaluation

The multiple stimulus evaluation tool, shown in Figure 2a, is the most important tool of the toolbox and has been shaped by feedback from the aforementioned studies. It follows the design requirements listed in Section 3.

Experience from previous listening tests also showed that to adequately compare sonic qualities of fragments longer than a few seconds, it is advisable that audio does not start from the beginning when switching between samples, but rather skips to the corresponding position. This obviously only applies when the samples are different versions of the same source audio.

### 4.2. Pairwise evaluation

Depending on the number of samples to be tested, the similarity of the samples, and the goal of the experiment, one might choose to use pairwise evaluation, i.e. presenting just two samples a time and offering a choice between different options (e.g. $A > B$, $A < B$, $A = B$, and possibly more gradations), see Figure 2b. This inherently leads to less information, as the multiple stimulus evaluation includes a near-continuous rating of the samples and thus allows to analyse e.g. the magnitude of differences in subjects' liking, rather than just the order. In [3] a comparative study between both approaches is conducted.

### 4.3. Batch sample processing tools

- Fragment selection with fade in/fade out: select begin and end if only a fragment of an audio file should be auditioned
- Automatic loudness equalisation of all samples in a sample set
- Check if every fragment is different (avoid including the same sample twice)
- Check if all sample rates are the same
- Join files ending in *.L* and *.R*, or *.M* and *.S* in the case of mid/side stereophony, as one stereo file.

## 5. CONCLUSION

By sharing this toolbox, we hope to offer an easy to use, flexible and powerful alternative to existing open source listening test interfaces. The source code of this toolbox can be found on `code.soundsoftware.ac.uk/projects/ape`. Please refer to the documentation for more information on the toolbox and how to use it. This continually updated repository also includes a dummy test set, with recordings of spoken numbers for easy debugging, and example data from the listening test discussed in this paper. Anyone can use, alter and redistribute the code, for whatever purpose, so long as this work is referenced.

## 6. FUTURE WORK

The toolbox will continue to be improved and expanded, for example to include other common test types such as mean opinion score (MOS), ABX, MUSHRA, two-dimensional rating (e.g. in a valence-arousal space), as well as automatic data analysis and result presentation. We welcome all contributions and feedback.

## 7. ACKNOWLEDGEMENT

The structure of this toolbox and the user interface design are inspired by the MATLAB routines from Vincent Rioux's PhD thesis [12], also used in [13, 14] (pairwise comparison).

## 8. REFERENCES

[1] *Method for the subjective assessment of intermediate quality level of coding systems*. Recommendation ITU-R BS.1534-1, 2003.

[2] E. Vincent, M. G. Jafari, and M. D. Plumbley, "Preliminary guidelines for subjective evalutation of audio source separation algorithms," in *UK ICA Research Network Workshop*, 2006.

[3] B. De Man and J. D. Reiss, "A pairwise and multiple stimuli approach to perceptual evaluation of microphone types," in *134th Convention of the Audio Engineering Society*, May 2013.

[4] S. Ciba, A. Wlodarski, and H.-J. Maempel, "Whisper – a new tool for performing listening tests," in *126th Convention of the Audio Engineering Society*, May 7-10 2009.

[5] A. V. Giner, "Scale - a software tool for listening experiments," in *AIA/DAGA Conference on Acoustics, Merano (Italy)*, 2013.

[6] S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application*. John Wiley & Sons, 2007.

[7] B. De Man and J. D. Reiss, "A knowledge-engineered autonomous mixing system," in *135th Convention of the Audio Engineering Society*, October 2013.

[8] B. De Man and J. D. Reiss, "Adaptive control of amplitude distortion effects," in *53rd Conference of the Audio Engineering Society*, January 2014.

[9] S. Bech, "Listening tests on loudspeakers: a discussion of experimental procedures and evaluation of the response data," in *Proceedings from the 8th International Conference of the Audio Engineering Society, Washington, DC*, May 1990.

[10] W. Hoeg, L. Christensen, and R. Walker, "Subjective assessment of audio quality-the means and methods in the EBU," *EBU Technical Review*, pp. 40–50, 1997.

[11] *Methodology for the subjective assessment of the quality of television pictures*. Recommendation ITU-R BT.500-13, January 2012.

[12] V. Rioux, *Sound Quality of Flue Organ Pipes - An Interdisciplinary Study on the Art of Voicing*. PhD thesis, Department of Applied Acoustics, Chalmers University of Technology, Sweden, 2001.

[13] F. Martellotta, E. Cirillo, M. Mannacio, and C. Skaug, "Subjective assessment of church acoustics," in *13th International Congress on Sound and Vibration*, July 2-6 2006.

[14] F. Martellotta, "A preliminary investigation on the subjective evaluation of church acoustics using listening tests," in *118th Convention of the Audio Engineering Society*, May 28-31 2005.