Audio Engineering Society

Convention Paper

Presented at the 148th Convention
2020 May 25 – 28, Vienna, Austria

A flexible software tool for perceptual evaluation of audio material and VR environments

Stefan Gorzynski [1,2], Neofytos Kaplanis[1], and Søren Bech[1]

[1]*Bang & Olufsen, R&D Innovation Labs, Bang & Olufsen Allé 1, 7600, Struer, Denmark*
[2]*Applied Psychoacoustics Lab (APL), University of Huddersfield, Huddersfield, HD1 3DH, United Kingdom*

Correspondence should be addressed to Neofytos Kaplanis (neo@bang-olufsen.dk)

ABSTRACT

Audio evaluation makes use of listening tests that allow the presentation of audio stimuli and gather data on their influence on human perception. Due to the rapid technological advancement and application-specific protocols, such work typically requires the experimenter to develop custom software that suits the experimental design and/or incorporates technological advancements, i.e., spatial audio. This article describes a publicly available listening test suite (BO-LTS) that allows standardised and sensory analysis evaluation protocols to be built with a high degree of flexibility, whilst supporting multichannel audio processing and the use of Virtual Reality (VR) environments.

1 Introduction

Perceptual assessment in the form of listening tests forms an essential part in quantifying the auditory sensations and qualities of audio signals and systems. Currently, there are standardised listening test methods [1] [2] [3] that have been created to allow repeatable, reliable, and comparable experiments. In recent years, Sensory Analysis (SA) methods developed by, and traditionally used in, the food industry [4] have been implemented in perceptual audio as an alternative way to gather information from human assessors [5, 6]. VR has also become a widely explored topic in audio research, especially in spatial audio and audio-visual interactions, with experiments often utilising the aforementioned test methods modified for VR-based experimentation [7].

Few tools exist that allow the creation of such types of listening tests. In majority, one could find toolboxes that focus on standard listening test generation [8] [9] [10] or protocol-specific procedures [11]. There is still an apparent lack of tools that allow quick, dynamic, and flexible generation of experimental designs, and they could incorporate more advanced procedures to support new technologies, such as next generation audio. For example, Zacharov et al. [12] used Ideal Sensory Profiling to assess object-based audio renderings. With the recent development of VR tools, the potential of combining those methods with VR can enable equally flexible test generation for VR related experiments. A recent example has been presented in [13].

This manuscript presents a software solution that aims to provide a listening test suite, capable of conducting both standardised, and sensory analysis evaluation protocols for audio material, namely the *B&O Listening Test Suite* (BO-LTS). BO-LTS combines the fundamental blocks of both families of test protocols and allows assessment of complex systems, such as spatial audio and virtual reality.

In Section 2, the most commonly standard listening tests are briefly discussed, followed by the recently adopted Sensory Analysis methodologies. Section 3 contains an overview of the system, followed by a description of the implementation in Section 4. Finally, the limitations and future work are described in Section 5, and concluding remarks are given in Section 6.

2 Listening Tests

Several perceptual evaluation methodologies exists for audio signals and systems. Following the idiosyncrasy of the system under test, and the research question posed, one can conduct an assessment using human assessors in several ways. In this section, a brief introduction of those methods will be presented. This includes both internationally standardised methods and SA methods.

In this manuscript an *excerpt* refers to a signal sample (program) e.g., music, speech, or noise. Stimulus refers to the resulting sound event of combining the excerpt with the regarded processing/condition i.e., the system under test. Trial refers to a single presentation ‘page’ of a test protocol on which evaluation is carried out.

2.1 ITU-R BS.1116-3

BS.1116 is the foundation that several other common ITU-R Recommendations are based on. It is used to assess systems that *introduce impairments so small as to be undetectable without rigorous control of experimental conditions [...]* [1]. As it is a method for detecting subtle impairments, the underlying requirements are one of the most stringent and comprehensive in the literature. Expert listeners are required, as they need to be highly trained and sensitive to such impairments. The test method used is double-blind triple-stimulus with hidden reference, as it is especially sensitive and permits accurate detection of small impairments. Assessors are presented with three stimuli: A, B, and C. ‘A’ is always the known reference, whilst the hidden reference and the impaired audio are randomly assigned to B and C in each trial. The assessor is asked to assess the impairments on stimuli B and C compared to A using a 5-point ITU-R impairment scale. The hidden reference is used to evaluate assessor’s reliability and repeatability against the known reference.

A key feature of BS.1116-3 tests is that the results must empirically and statistically show that any failure to find differences is not due to experimental insensitivity, but because the systems are actually transparent. This makes it as the ideal tool for evaluation of signal codecs.

2.2 ITU-R BS.1534-3 ’MUSHRA’

BS.1534-3 is used for *the assessment of systems that introduce medium to large impairments that would normally fall into the lower half of the BS.1116 impairment scale [...]* [2]. As BS.1534 allows assessors to compare differently degraded items directly to one another, they can more easily discern the differences between them. The test method used is *’MUltiple Stimulus test with Hidden Reference and Anchor’* (MUSHRA), and has been shown to give accurate and reliable results for evaluation of intermediate audio quality. In each trial the assessor is presented with a known reference, a hidden reference, hidden anchors, and the stimuli from the systems under test. The assessor may switch between the stimuli at will and grade them relative to the reference and each other using the 100 point MUSHRA scale (bad, poor, fair, good, excellent). The use of the hidden anchor allows the experimenter to exclude any results from assessors who either rated the hidden reference too low, or the hidden anchors too high, increasing the reliability of the results.

2.3 Other Standard Methods

Although BS.1116-3 and BS.1534-3 are the most commonly used methods by audio manufacturers, the scales and test paradigms from other ITU recommendations, such as ITU-T P.800 [3], are sometimes used. This is typically due to the limited applicability of fully compliant protocols, and the needs for specific experimental designs. In addition, variations of BS.1116-3 or BS.1534-3 that do not strictly follow the ITU guidelines are occasionally used, which requires flexibility in the design of testing interfaces.

2.4 Sensory Analysis Protocols

Sensory Analysis (SA), is a family of evaluation protocols for assessment of perceptually complex stimuli. Originally applied in food sciences, the ability of those methods to decompose the perceptual characteristics of stimuli have naturally seen their way in audio evaluation in room acoustics, reproduced sound, and spatial audio; see Zacharov [6].

SA methods can be categorized into two groups: (1) discriminant methods and (2) descriptive methods. The former attempts to quantify the perceived differences between stimuli, i.e., 'are the stimuli different in any way?', whilst the latter goes beyond that and attempts to identify, describe, and quantify the perceived differences.

There are many protocols within the SA analysis framework. In this manuscript, the most relevant protocols will be briefly described, based on their relevance in audio evaluation.

2.4.1 Descriptive SA and Attribute Elicitation

The most commonly used SA techniques in audio evaluation are found within descriptive techniques. These protocols allow the assessment of stimuli based on the assessors' elicited vocabulary; a set of words, known as perceptual *attributes*, that describe the sensory characteristics of a given stimuli set, as identified by the assessors. Those attributes are later used to quantify the magnitude of the difference between the stimuli, typically in a grading task.

The elicitation of these characteristics can be achieved with verbalization methods, i.e., the assessors labeling the characteristics of the stimuli, as well as non-verbal/indirect methods such as sorting tasks or graphical positioning tasks.

Verbal elicitation has been instrumental in audio evaluation, as it allows new audio algorithms and complex sound fields to be quantified. The development of the BO-LTS has therefore focused on developing tools for direct verbal-based elicitation tasks [14, p. 19].

In audio evaluation literature, two verbal elicitation approaches are apparent: *consensus* vocabulary, and *individual* vocabulary. In individual vocabulary methods, each assessor identifies and epitomizes the sensory characteristics, building their own attribute set. This means that each assessor grades the stimuli using their own attributes, only. In contrast, in consensus methods, the individual attribute sets of each assessor are merged into a common vocabulary through discussions or regulation, and the same vocabulary is used by all assessors in the grading tasks that follow.

BO-LTS is primarily focused on individual vocabulary elicitation methods due to their popularity in audio. These methods follow a two step process: (1) the elicitation of perceptual attributes, and (2) the grading stage which can consist of ranking, scoring, or a comparison task.

With this in mind, the interface elements in BO-LTS were designed to be generic rather than limited to a specific way of attribute elicitation, as many of the methods have similar interface requirements, as shown in Table 1.

In this way, the software would also allow consensus vocabularies to be used in the grading tasks. This allows the experimenter to conduct the individual elicitation tasks first, and introduce the merged consensus vocabulary in BO-LTS framework at a later stage.

2.4.2 Free Choice Profiling

Free Choice Profiling (FCP)[15] works on the assumption that assessors only differ in the way they describe different attributes, not in how they perceive them. Based on this, the assessors are presented with all available stimuli in the elicitation phase, and can record any attributes that they believe to be relevant. They are also required to develop their own endpoint descriptors of the attributes for each pole of the sensory spectrum. Following this, the grading is carried out by presenting the assessor with one stimulus at a time, and having them grade it for all attributes before moving on to the next stimulus.

2.4.3 Flash Profile

Flash Profile (FP) is a well known as 'rapid' sensory method [15], and uses the same elicitation method as FCP. FP assumes that a general expertise of the assessors in the experimental stimuli set allows a more robust and quick evaluation without the use of pre-test trials and stimuli-specific training. This allows a smaller assessor panel (4-5) to be sufficient for a statistically robust profiling exercise. The main difference from FCP

Method	Elicitation Presentation	Grading Presentation
Standardised Methods	N/A	Multiple, one attribute, structured scale
Free Choice Profiling	Single stimulus	Single, all attributes, unstructured line scale
Flash Profile	All Stimuli	Multiple, ranked, unstructured line scale
Check All That Apply	N/A	Multiple, all attributes, checkboxes
Polarised Sensory Profiling	N/A	Di/Triad, unstructured line scale
Sorting	N/A	Multiple, sorted, 2 (or 3) dimensional area to sort on
Ideal Profile	N/A	Single, all attributes, unstructured line scale
Repertory Grid Technique	Di/Triad	Multiple, ranked, unstructured line scale

Table 1: The presentation and grading types used in common sensory analysis methods. 'Single' means that only on stimulus is presented per trial, even if multiple attributes will be evaluated. 'Multiple' means that all stimuli for each excerpt are presented in each trial.

lies in the grading stage, where a comparative assessment is carried out, that is, for each attribute all stimuli for a given excerpt are ranked relative to one another in each trial (no known reference is given). This has been claimed to remove the need for a familiarization phase with the attributes, with the drawback being that the data obtained from these experiments is the relative sensory positions of the stimuli and are difficult to interpret, requiring rigorous and complex multi-factorial analysis. Due to the speed, flexibility and the use of experts, naturally being available in audio manufacturers, these methods have been successfully used for the assessment of loudspeakers sound quality [16], room acoustics [17], and car cabins [18].

2.4.4 Check-All-That-Apply

Check-All-That-Apply (CATA) is a rapid method that tends to be used with untrained assessors to quickly gain descriptive data regarding a stimuli set. Assessors are presented with a high number of possible attributes, and by listening to the stimuli available, the task is to confirm which attributes describe the stimulus presented to them. In these tests, the attribute list may be selected by the assessor beforehand, and a dimensionality reduction is followed after the CATA experiment is performed. A recent example of CATA in audio has been described by Hicks et al. [19].

3 B&O - Listening Test Suite Overview

This chapter details the interfaces that are available in the BO-LTS. Currently the software package is split into two separate versions, the desktop version and VR version, respectivly called BO-LTS and BO-LTS-VR respectively. BO-LTS-VR is based on the same MAX patch as the desktop version, with the interface elements implemented in VR environment using Unity. In future those will be integrated.

The listening tests types have been built using protocols that can be combined in each experiment. The following sections' naming follows the user interface and basic process in a trial, not the name of a specific procedure. For example, 'One Attribute Evaluation, Multiple Stimuli per trial' describes an interface that presents of all stimuli for a given excerpt, for single attribute evaluation in each trial, as in BS.1116 or BS.1534.

3.1 BO-LTS (Non-VR)

BO-LTS was developed first to lay the groundwork for the VR tests and as such has more test paradigms, as these were used to decide which would translate best to VR applications. The paradigms are split into different sub-patches as shown in Figure 1, and their use cases and purposes are described below.

3.1.1 One Attribute Evaluation, Multiple Stimuli - Standard

This software implementation was designed to run standardised listening grading tasks and it allows for only one attribute to be graded in each experiment; for example a standardised Basic Audio Quality listening test. It has nonetheless flexible scales, and it allows the use of anchors. The user can also select repetitions of the entire test when needed. Although it allows custom scale labels to be used, several presets are available to accommodate standardised methods such as BS.1116, BS.1534, and other common scales.In our

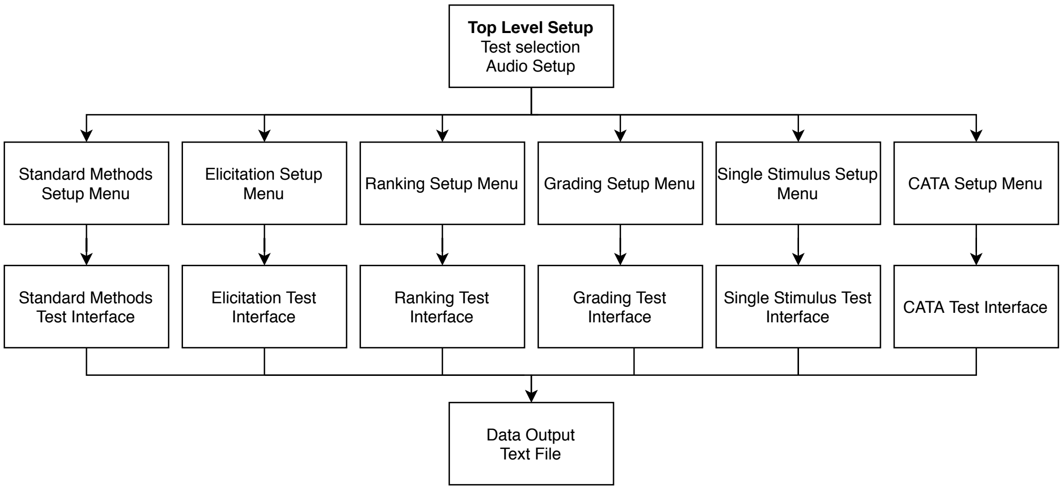

Fig. 1: Topology of the MAX based interface generation

experience such methods are rarely used, as grading multiple attributes are more often needed in the development cycles.

3.1.2 Multiple Stimulus Ranking, Single Attribute per trial

In this ranking task, the interface in each trial displays all stimuli for a given excerpt on a single scale/line, allowing the assessor to directly rank them against one another. This approach is recommended in ranking tasks such as FP. Combined with the elicitation interface, a full FP experiment can be carried out using this element. To offer flexibility in many experimental approaches, the order of presentation can be selected, with either trials grouped together based on an attribute or by each excerpt depending on what is required. Randomisation for all elements of the test can also be toggled individually.

Finally, options to include hidden repeats of a random stimulus in each trial, a reference, and hidden reference are available to aid the experimenter in gauging the assessors' ability. Those options were found useful for audio evaluation in our experience and they have therefore been implemented in this package.

3.1.3 Multiple Stimulus Grading, Single Attribute per trial

In this grading task, the interface is identical in functionality to the ranking task detailed above (3.1.2). The main difference between the two is that the stimuli in the grading interface have their own separate scale. This type of interface can facilitate MUSHRA-based experiments that follow the BS.1534 recommendation but it can be modified, for example, to allow for rating of multiple attributes in the experiment. For this reason, the recommended MUSHRA scale has been included as an option, which replaces the attributes' endpoint descriptors from elicitation. In future release, the scale will be completely customisable to allow many more standardised tests to be easily carried out with modifications for SA.

3.1.4 Single Stimulus Grading, Multiple Attributes per trial

Intended predominantly for FCP or experiments in which audio attributes are not being graded, assessors are presented with one stimulus at a time, and all attributes need to be graded for said stimulus before the assessor moves on to the next trial. This interface is not recommended for grading auditory attributes, as it has been found that assessors can be overloaded and confused when assessing multiple attributes at once. Instead, it has been found that this is useful for assessing other aspects of audio products, such as headphone comfort, fit, and ease of use, for example.

3.1.5 Check All That Apply (CATA)

This refers to implementation of the CATA method, as described in Sec. 2.4.4. A grid is presented whereby the rows describe the available attributes and columns include the stimuli to be evaluated. As this is intended to quickly gain data from the assessors, all attributes and all stimuli for a given excerpt are presented in each trial, and assessors are required to tick a box if they believe that an attribute is present for each stimulus. The presentation does not allow the evaluation of audio that is not selected.

3.1.6 Verbal Elicitation

The interface for verbal elicitation can host attribute generation tasks, for example the first part of FP or FCP protocol. The assessor is given all the stimuli (excerpts $\times$ conditions) on screen. The assessors are able to create and edit their own list of attributes by presenting them with all available stimuli and excerpts at once, in one trial, before recording their attributes and definitions/endpoint descriptors of the attributes. The experimenter can choose to pre-load a list of attributes into the interface as well if they wish to include them as a guide for less confident assessors, however this should be done with caution as it can introduce bias.

3.2 BO-LTS-VR

BO-LTS-VR is capable of creating two different types of interface, a grading interface that can be used for ITU or SA methods, and an elicitation interface that can be used in several different SA methods.

In both VR methods, when the test is started, the assessor is presented with a familiarisation environment. This contains an identical interface to the main experiment setup that can be used to play and switch between a random set of stimuli to allow the assessors to get comfortable using the VR controls before the actual experiment begins, reducing any bias that may be introduced due to assessors being unfamiliar with VR.

3.2.1 Verbal Elicitation - VR

The principles of this are identical to those in Section 3.1.6, implemented in VR.

Text entry in VR can be challenging, and after testing several different methods, including laser pointers, 'virtual touchscreens', and gaze based text entry, a more novel approach was implemented. The assessor has two 'drum sticks' attached to their hands, which they can then use to push down keys on a qwerty keyboard laid out in front of them as shown in Figure 2. Although this may sound counter-intuitive and slow, assessors who tried it in informal testing much preferred it and could type at relatively fast speeds after only a few minutes of familiarisation.

3.2.2 Grading - VR

The VR grading interface consists of multiple selectable stimuli and sliders to give them a score, with completely customisable scales, labels, reference and anchors (for MUSHRA), hidden repeat and randomisation options as listed in the desktop version. Those settings allow the experiment to set the appropriate parameters to meet certain specifications of multiple standardised methods, whilst keeping flexibility for novel approaches. For example, ITU-R BS.1534 (MUSHRA) requires that there be a hidden repeat of the reference in the stimulus set being graded, as well as hidden anchors and repeats (the same stimulus twice in one trial).

The test interface is shown in Figure 3. This interface can be used for grading tests, but can also be used as a pseudo-ranking interface for SA experiments by instructing assessors to rank the stimuli. Although the stimuli are being graded on separate sliders, the assessors will still visually see their positions relative to one another and can think of them as being 'ranked'.

3.3 Data Output

The data is given in the format of a CSV text file with scores shown for each attribute, along with the track name for each stimulus to ensure that the correct audio files were loaded. Presentation order is also stored in the the data file to allow the experimenter to check that correct randomisation settings have been used.

The data is auto-saved into a temporary file every time the page is changed on the test. This enables some data to be recovered even in the event of the test being cut short by a crash or power-cut.

4 Software Development

The design of the software package is based on independent modules, instead of fixed protocol wizards. The user can choose a combination of experimental settings to fulfil the requirements for a specific protocol. For example, if the experimental protocol requires a ranking task, the user is able to choose the type of ranking

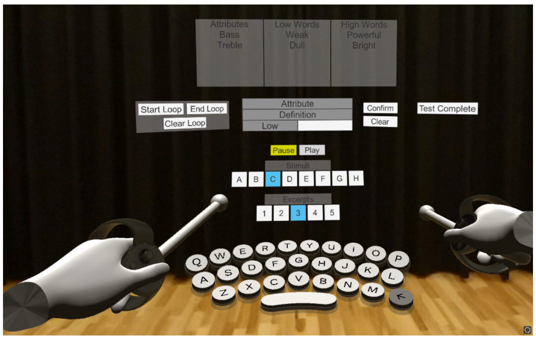

Fig. 2: VR elicitation interface

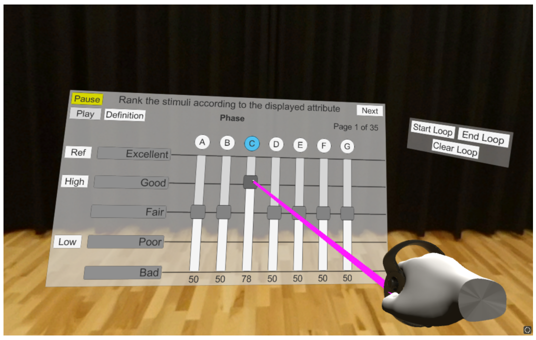

Fig. 3: VR grading interface

protocol. Combined with an attribute elicitation protocol, a full Flash Profile methodology can be followed, for example, without limiting the customization of the experimental design to specific tasks.

The audio processing and data handling engine was developed in MAX 8 [20] along with the desktop Graphical User Interface (GUI), whilst VR interfaces have been developed in the Unity games engine [21] to provide a VR interactive environment and control the MAX engine. This allows the experimenter to conduct the same experimental procedures both on a PC and in VR, as a complete audio-visual VR experience. This can deliver a fully flexible listening test generation software that, combined, offers state of the art evaluation protocols without compromising ease of use, with enough flexibility to allow for a number of listening test methods to be implemented.

Randomisation, use of references, and different block designs (i.e., grouping trials by attribute or by excerpt) can all be easily implemented from the setup menu. Data is auto-saved every time a trial is completed as an extra precaution for data preservation. Presets can also be saved and loaded for all methods, meaning that once the interface has been set up for a particular experiment, it can be quickly recalled between sessions.

4.1 MAX Implementation

Non-VR BO-LTS experiments are based entirely in MAX, with the use of spat5 [22] for audio playback. The architecture is based on having a 'top level' menu that is used to set audio routing, fade times, speaker compensation, etc, and to select the interface type to be used in a specific experiment. Once the interface type is selected, a separate patch is loaded with settings specific to that interface, an example of which is shown in Figure 4.

4.2 VR Implementation

The BO-LTS-VR test mimics the way the MAX presentation works. However, the current version has implementation differences between the elicitation and grading processes meaning that it is a separate program. These will be merged in future releases.

The interface generation works in a similar way to BO-LTS, in that all setup is done in MAX, then all the parameters are sent to Unity, which then generates the interface and begins the test. Unity is used to present the assessor with visuals and handle user input, while MAX is still used for all test setup, audio playback, and data management. Communication between Unity and MAX is handled using UDP, which is natively supported in MAX, and implemented via the UnityOSC asset package [23] in Unity. This allows two way communication, with MAX sending the test parameters to Unity for interface generation, and Unity sending any user input back to MAX. The data flow between the two is shown in Figure 5. The VR interaction was developed for Oculus Rift and as such, all user interaction within Unity is handled by the Oculus Integration asset package [24]. The VR presentation is a 360 picture being used as a skybox, while everything that the assessor interacts with are UI Canvas objects.

4.3 Audio Playback

As some of the experiments carried out in perceptual audio evaluation involve playback over large speaker arrays, the MAX patch was built to accommodate switching between tracks containing up to 64 channels of

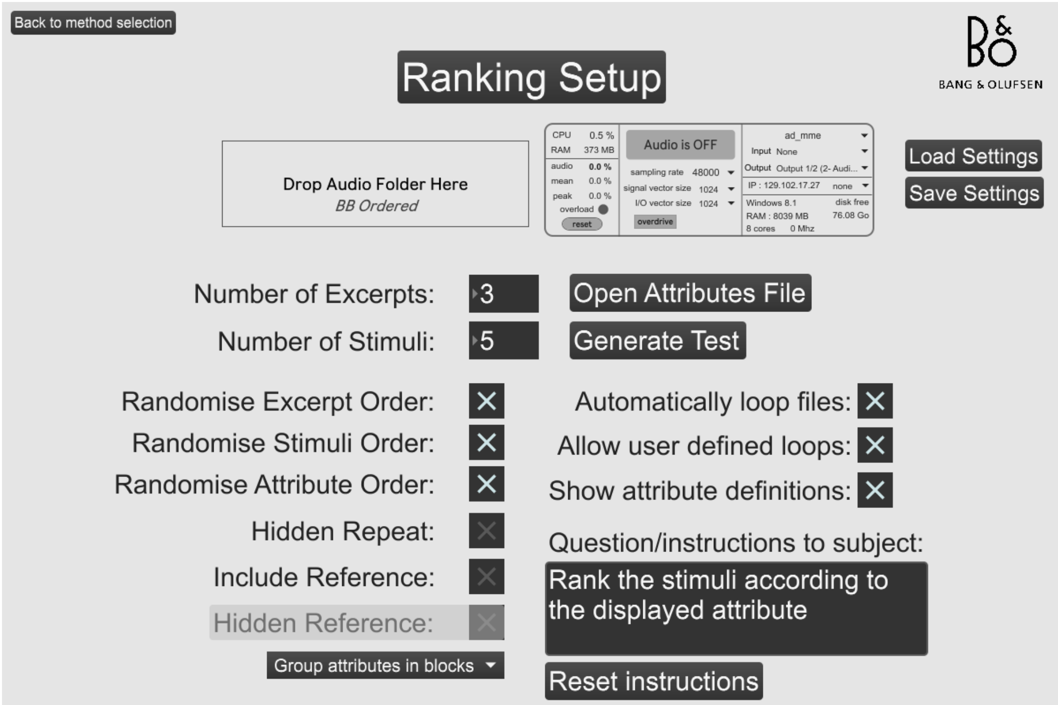

Fig. 4: Ranking setup menu where randomisation, repeats, reference etc. options can be selected.

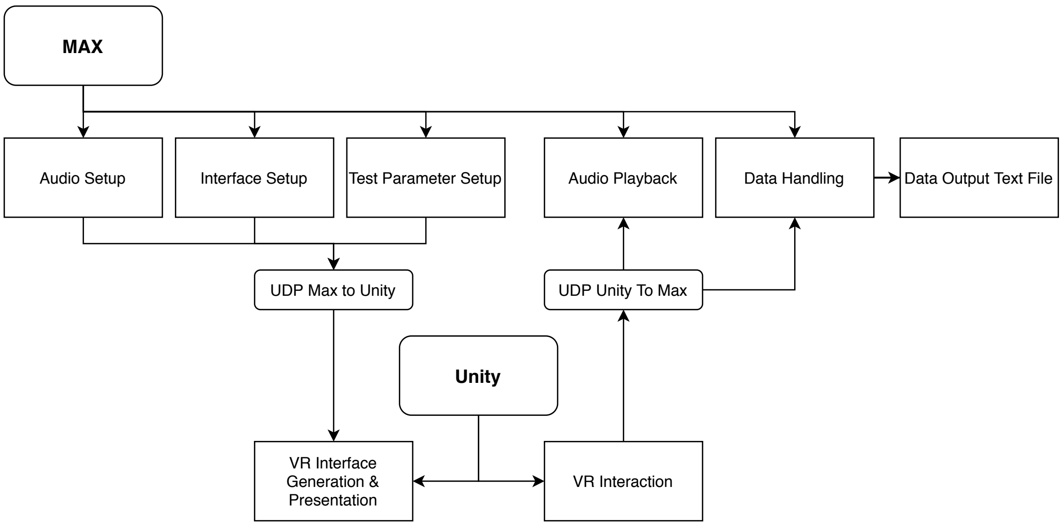

Fig. 5: Data flow for BO-LTS VR implementation

uncompressed audio with near zero latency. Due to the nature of listening tests having multiple stimuli in each trial, a specialized playback engine was developed to handle this.

The BO-LTS soundplayer utilizes the external SPAT package [22]. Two *spat5.sfplay* soundplayers are used, with one being active at a time. When a new stimulus is selected, the patch will first load the new audio file into the inactive player, retrieve the current playback position from the active soundplayer, then begin playing the new file from this position so that the two files are in sync. The lower limit of switching time is dependant on the computer processor, hard drive speed, and number of channels of audio being played back, and it has been observed that solid state drives may be required when multi-channel PCM files are used in this setting. This process can take as little as 50ms, depending on the computer hardware being used. The switching between tracks is synchronized.

4.3.1 Advanced Audio Features

A set of advanced audio control features have been added to further improve flexibility and ease of use.

Some protocols require specific timing requirements for switching stimuli during comparative tests. BO-LTS advanced audio settings allow the length of the fade time, as well as an additional length of silence between track switching to be specified by the experimenter from the main setup menu of the patch. This may seem like a trivial setting, but in stringent test methods such as BS.1116-3, specific fading is required in order to meet the requirements. It is also possible to loop small sections of an audio file, as recommended by [1] which can allow assessors to select smaller sections to focus on. A built in routing matrix is also implemented and it allows an audio file's tracks to be freely routed to any of up to 64 DAC outputs, a feature that has proved useful in large speaker array setups. Speaker gain and delay compensation can also be applied by the soundplayer using two column text files containing channel numbers and gain or delay values.

5 Discussion

The work done on this project has resulted in a solid foundation for flexible and accessible desktop and VR based listening test development. Several considerations relating to text-input in VR, as well as modification of standard 2D interfaces recommended in standard protocols in 3D and possibilities for new experimental paradigms seem to require further research.

There are nevertheless several opportunities to develop this package with currently used methods and create a more comprehensive listening test suite. This could include the addition of RGT triplet elicitation methods, familiarisation designs with alternative audio playback, as well as ability to introduce new elements for evaluation of advanced or prototype spatial audio schemes.

For further VR development, additions could be made such as allowing to loading of custom 360 photos through MAX, a feature that currently requires the scenes to be added in the Unity editor. Along with this, cross-platform compatibility could be implemented to allow the test to be run on both Oculus and OpenVR headsets.

In the next years, BO-LTS and BO-LTS-VR will be continuously updated. The first step would be to merge the VR and non-VR versions, with each other so that only one program needs to be opened to run any kind of test, in VR or in desktop mode. Moreover, options for new scenes and ability to have multiple playback is planned for next releases.

6 Summary

The manuscript described the rationale and implementation implications of a software package that lays the foundation for desktop and VR based listening test development, without the need for any prior programming knowledge. BO-LTS functions as a standalone program and is able to generate several of the most common listening test types, but can also be used for tests that do not relate directly to audio analysis.

There are several areas that can be expanded, and will be in the future. Currently it offers a fully functional listening test suite and is now available with VR capability that enables fast, easy and reliable VR listening test generation.

The software package is freely available at https://github.com/APL-Huddersfield/BO-LTS

Acknowledgements

The authors would like to thank Hyunkook Lee, for his help during this project.

References

[1] *Recommendation BS.1116-3, Methods for the subjective assessment of small impairments in audio systems*, ITU-R, 2015.

[2] *Method for the subjective assessment of intermediate quality levels of coding systems*, ITU-R, 2015.

[3] *Recommendation P.800, Methods for objective and subjective assessment of quality*, ITU-T, 1996.

[4] Lawless, H. T. and Heymann, H., *Sensory evaluation of food: principles and practises*, International Thomson Pub., 1998.

[5] Bech, S. and Zacharov, N., *The Perceptual Audio Evaluation: Theory, Method and Application*, Wiley, 2006.

[6] Zacharov, N., *Sensory evaluation of sound*, CRC Press, Taylor Francis Group, 2019.

[7] Farina, A., Pinardi, D., Binelli, M., Ebri, M., and Ebri, L., "Virtual Reality for Subjective Assessment of Sound Quality in Cars," in *Audio Engineering Society Convention 144*, 2018.

[8] Jillings, N., De Man, B., Moffat, D., and Reiss, J., "Web Audio Evaluation Tool: A framework for subjective assessment of audio," 2016.

[9] Gribben, C. and Lee, H., "Toward the Development of a Universal Listening Test Interface Generator in Max," in *Audio Engineering Society Convention 138*, 2015.

[10] De Man, B. and Reiss, J. D., "APE: Audio Perceptual Evaluation Toolbox for MATLAB," in *Audio Engineering Society Convention 136*, 2014.

[11] Hummersone, C., "MUSHRA-MaxMSP," 2017.

[12] Zacharov, N., Pike, C., Melchior, F., and Worch, T., "Next generation audio system assessment using the multiple stimulus ideal profile method," *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2016.

[13] Johnston, D., Tsui, B., and Kearney, G., "SALTE Pt. 1: A Virtual Reality Tool for Streamlined and Standardized Spatial Audio Listening Tests," in *Audio Engineering Society Convention 147*, 2019.

[14] Lorho, G., *Perceived Quality Evaluation: An Application to Sound Reproduction over Headphones*, Ph.D. thesis, Aalto University, 2010.

[15] Delarue, J., Lawlor, J. B., and Rogeaux, M., *Rapid sensory profiling techniques and related methods: applications in new product development and consumer research*, Elsevier/Woodhead Publishing, 2015.

[16] Arrieta Sagredo, I., Moulin, S., Bech, S., and ren, "Sensory Profiling of High-End Loudspeakers Using Rapid Methods-Part 4: Flash Profile with Expert Assessors," in *Audio Engineering Society Convention 146*, 2019.

[17] Kaplanis, N., Bech, S., Lokki, T., van Waterschoot, T., and Holdt Jensen, S., "Perception and preference of reverberation in small listening rooms for multi-loudspeaker reproduction," *The Journal of the Acoustical Society of America*, 146(5), pp. 3562–3576, 2019, doi:10.1121/1.5135582.

[18] Kaplanis, N., Bech, S., Tervo, S., Pätynen, J., Lokki, T., Waterschoot, T., and Jensen, S., "A rapid sensory analysis method for perceptual assessment of automotive audio," *Journal of the Audio Engineering Society*, 65(1/2), pp. 130–146, 2017, ISSN 1549-4950, doi:10.17743/jaes.2016.0056.

[19] Hicks, L., Moulin, S., and Bech, S., "Sensory profiling of high-end loudspeakers using rapid methods - Part 3: Check-all-that-apply with naïve assessors," *Journal of the Audio Engineering Society*, 66(5), pp. 329–342, 2018, ISSN 1549-4950, doi:10.17743/jaes.2018.0015.

[20] Cycling 74, Accessed 2020.

[21] Unity, Accessed 2020.

[22] IRCAM, Accessed 2020.

[23] Fredericks, T., Accessed 2020.

[24] Oculus Integration, Accessed 2020.