Gribben, Christopher and Lee, Hyunkook

Towards the development of a universal listening test interface generator in Max

Towards the Development of a Universal Listening Test Interface Generator in Max

Christopher Gribben and Hyunkook Lee

Applied Psychoacoustics Lab (APL), University of Huddersfield, Huddersfield, HD1 3DH, UK

christopher.gribben@hud.ac.uk; h.lee@hud.ac.uk

ABSTRACT

This engineering brief describes HULTI-GEN (Huddersfield Universal Listening Test Interface Generator), a Cycling '74 Max-based tool. HULTI-GEN is a user-customisable environment, which takes user-defined parameters (e.g. the number of trials, stimuli and scale settings) and automatically constructs an interface for comparing auditory stimuli, whilst also randomising the stimuli and trial order. To assist the user, templates based on ITU-R recommended methods have been included. As the recommended methods are often adjusted for different test requirements, HULTI-GEN also supports flexible editing of these presets. Furthermore, some existing techniques have been summarised within this brief, including their restrictions and how they might be altered through using HULTI-GEN.

1. INTRODUCTION

Within the audio industry, efficient and reliable means of assessing auditory attributes are vital to helping us understand our perception of sound. Subjective listening tests on a computer are often used to carry out such assessments, where a listener is presented with a graphical user interface and asked to grade auditory stimuli one way or another. There is no 'one method fits all' approach to listening tests, and there are cases where the formats of existing recommendations and methods (as discussed briefly in Section 2 below) require adjustment to test for novel attributes. When designing a listening test for these attributes, a robust and repeatable testing method is often thought to be the most important consideration. A few key features that require thought during test design are: the scale on which the user is grading, whether there are audible reference or anchor points on the scale, and also how a large number of stimuli might be split into separate trials, to make testing more manageable for the listener.

The content of this brief describes HULTI-GEN (Huddersfield Universal Listening Test Interface Generator), a Cycling '74 Max-based tool that generates a listening test interface from user-defined parameters. The main aim of HULTI-GEN is to address the key features of test design mentioned above, by delivering a step-by-step process to customise and build a listening test interface; this has resulted in an adaptable tool that can be used for a broad range of listening test scenarios, as well as being a platform for novel test development.

Initially developed for final year students at the University of Huddersfield, HULTI-GEN can be useful to both experienced and inexperienced audio researchers alike. In particular, those who want to pilot various approaches for comparing auditory stimuli, as well as develop new techniques, can benefit from the user-friendly flexibility of the software. Templates based on commonly used listening test methods have also been included, along with the ability to alter these presets.

2. LISTENING TEST METHODS

This section provides a brief summary and discussion on common listening test methods.

As defined in ITU-R Recommendation BS.1116-3 [1], the double-blind triple-stimulus test method features a set of three auditory stimuli per trial. Two of the stimuli are graded for impairments against a third reference signal; one of those being graded is a hidden case of the reference signal. These judgments are made on a continuous five-grade scale, with descriptive anchors ranging from 'Imperceptible' (5.0) to 'Very annoying' (1.0). The method is commonly used to compare and detect small impairments of audio quality between high quality audio samples. It is recommended that 'Basic audio quality' be set as the single, global attribute during testing, although an experimenter may also choose to define and evaluate other attributes. Potential attributes in the document, include, 'Stereophonic image quality', 'Timbral quality', and 'Localisation quality'. Although these attributes also relate to audio quality, the grading scale might benefit from alternative labelling, which is something to consider during test design. The five-grade scale implemented here is also similar to the Mean Opinion Score (MOS) standard, which has been used for assessing the transmission quality of audio [2].

In contrast to the double-blind triple-stimulus, a format named the "MUltiple Stimulus test with Hidden Reference and Anchor (MUSHRA)" concerns medium to large impairments of intermediate audio quality, as described in ITU-R Recommendation BS.1534-2 [3]. While MUSHRA also assesses perceived audio quality, the test features a multi-comparison layout instead, and was designed to test for the impairments of audio codec processing. These judgments are made on a continuous scale of 0-100, with five grading regions from Bad (0- 20) through to Excellent (80-100). For each trial, multiple stimuli are compared against a high quality, unprocessed reference. Amongst the stimuli, three anchor samples derived from the reference are included – two with low-pass filters at 3.5 and 7 kHz (low and intermediate anchors), and the third is the unprocessed reference (hidden reference i.e. high anchor at 100).

Variations of both the main test methods described above are often used for assessing auditory attributes other than audio quality e.g. spatial characteristics, such as, apparent source width and listener envelopment. For example, if there were large spatial differences between stimuli, a multi-comparison method based on MUSHRA might be used. In this instance, the choice of reference is subjective and requires rational consideration; it would be invalid for an investigator to use the stimulus they perceive to be most spacious as the high anchor reference (i.e. 100 on a scale of 0-100), as a test subject may perceive another stimulus to be more so. Therefore, movement of the reference from 100 would reduce bias and give room for the listener to grade higher. One solution for this has been to use a continuous bi-polar scale (e.g. -50 to 50) where the reference is at 0, similar to the seven-grade comparison scale in [4]. This type of scale could also feature a semantic differential grading system, where two opposing adjectives are at either end of the scale (i.e. louder and quieter).

All of the examples discussed so far use a continuous scale. A potential issue with this scale-type is a lack of control over the way a subject grades it, in terms of the different spread of scores between listeners – this is usually addressed by normalisation of the results [3]. Alternatively, a practical method to guide the use of a scale has been to introduce additional audible anchors, helping to audibly define the scale limits and support the labelling [5]; however, this technique may bring unwanted bias and affect the scale's continuous nature, removing the option to normalise the data. An ABX test is a simple alternative for detecting slight perceptual difference between two samples, without the need for a scale. It is the same triple-stimulus format as [1], but instead of grading, the listener is forced to identify which of the two stimuli is the hidden reference. Likewise, a pairwise comparison test is also used for small impairments, but does not feature a reference [6].

As there are many factors contributing to the design of a listening test, no single test method is correct for all situations, and new formats may need to be developed as novel auditory attributes are proposed. When contemplating test features, for instance, the inclusion of audible reference anchors or suitable labelling during test design, a listening test interface tool that allows the user to easily alter them would be of great use.

3. SOFTWARE

3.1. Existing Software

Considering the thoughts in Section 2 above, a number of adaptable listening test interfaces are already in existence [7-10], however, they all have their limitations in terms of customisability, with some only available commercially. A large proportion of the open-source listening test programs are based in MATLAB, due to its existing functions for handling audio and creating graphical user interfaces. Despite this, it has been found that programming knowledge can be advantageous when preparing and designing a test in this software, which can limit the adaptation and flexibility for a less experienced user. There have also been occasional instances of incompatibility with different versions of MATLAB, as well as some software being limited to certain operating systems.

3.2. Development Software: Cycling '74 Max

Given these restrictions, the tool described in this brief (HULTI-GEN) is a patch that has been developed using the software Max (also referred to as Max/MSP, of which MSP is the Max Signal Processing module). Max is a cross-platform visual programming language from Cycling '74, specifically designed for developing music and multimedia applications. Some useful features of Max, which have contributed in part to the initial development of HULTI-GEN, include: real-time manipulation of digital audio signals including multi- channel playback, an object/modular-based environment that encourages rapid prototyping, the ability to generate new objects (i.e. sliders) for a user without background editing, the capacity to export developed software as a standalone application or collective file, and the availability of graphical objects that allow the user to easily input and store data.

Many listening tests have already been conducted using reliable interfaces developed in Max, however, these patches are often restricted to a particular testing method – this makes them an inflexible tool to the layman who has no prior knowledge of programming in Max. As far as the authors are aware, no such universal listening test generator, similar to those mentioned in Section 3.1, exists in the Max environment at the time of writing. Therefore, HULTI-GEN is considered to be a useful tool for both professionals and students conducting auditory research, no matter their level of experience in listening test design.

4. HULTI-GEN OVERVIEW

HULTI-GEN is a flexible listening test tool that generates a graphical user interface (GUI) for audio comparison tests. It has a focus on simplifying the user- experience for both the test designer and test subject, and provides a customisable foundation on which to build novel testing formats. At present, HULTI-GEN has templates based on the ITU-R Recommendations 1534-2 (MUSHRA) and 1116-2, as described above.

No previous knowledge of the running software (Max) is needed to use HULTI-GEN, as all end-users interact solely with the GUI. The tool features a simple drag- and-drop system to import and store the audio filenames of the stimuli for testing, as well as a step-by-step guide for constructing and editing a listening test. This allows the test designer to efficiently prototype different test methods during pilot experimentation, through the swift adjustment of various test parameters and instantaneous generation of a new interface.

In an attempt to address some of the listening test limitations discussed in Section 2, key features of HULTI-GEN include the following:

• Full randomisation of the trial order and of stimuli within trials for each test conducted.

• Listening test templates of established methods are included, which can also be altered.

• A drag-and-drop function to quickly import the stimuli filenames when preparing a test.

• Definition of the scale limits and resolution, as well as the starting position of the slider.

• Flexible customisation of the scale labelling.

• The option to include an audible reference/anchor at varying positions on the scale.

A basic flow-chart detailing the listening test design process of HULTI-GEN's GUI can be seen in Fig. 1. It demonstrates the ability to easily navigate from the Main Menu and edit many parameters, for instance, the distribution of stimuli, use and position of a reference, alterations to the grading scale, and importing new stimuli. There are also options to load the saved settings or to create a completely new interface, which guides the user through the process and considerations to make.

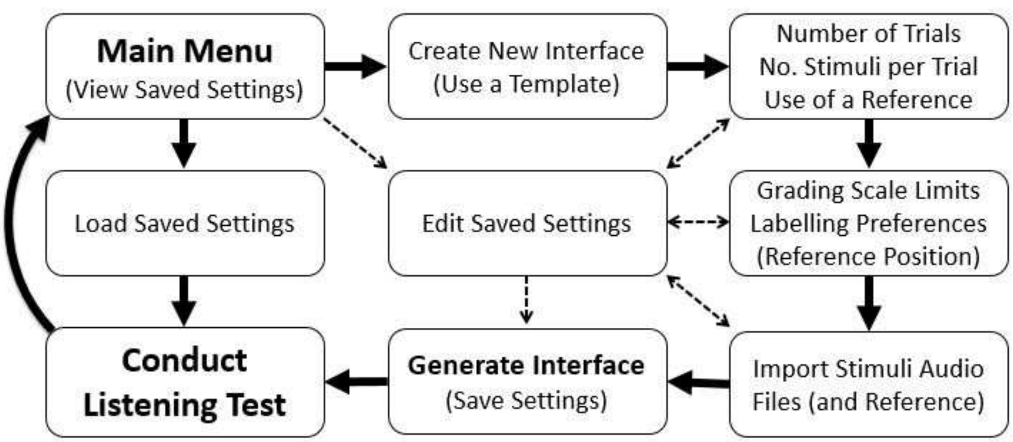

Fig. 1. Flow-chart of the process in HULTI-GEN

An example of a listening test interface that has been generated in a prototype of HULTI-GEN can be seen in Fig. 2 – it features a multi-comparison format similar to MUSHRA, but with a bi-polar scale, a reference signal at 0 and customised labelling.

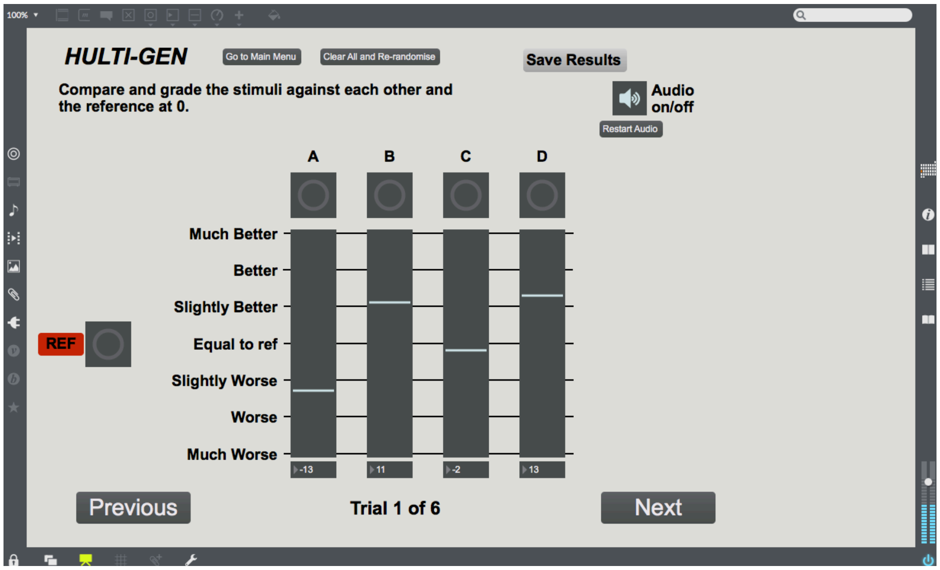

Fig. 2. An example HULTI-GEN interface in Max 7

5. CONCLUSION AND FUTURE WORK

HULTI-GEN is a Max-based tool for generating customisable listening test interfaces. It is thought the tool will be of interest (and use) to all audio researchers, particularly those interested in developing new testing methods. HULTI-GEN is freely available from:

http://www.hud.ac.uk/research/researchcentres/mtprg/projects/apl/

In future, the tool would benefit from the incorporation of additional features, such as, a control for the subject to adjust the loop size and the capability of multi- channel playback, giving an increased compliance with the recommended documents. There is also the potential to include other test methods in further development, for example, Pairwise Comparison, ABX, Mean Opinion Score and a way to help elicit novel auditory attributes.

6. ACKNOWLEDGEMENT

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), UK, Grant Ref. EP/L019906/1.

7. REFERENCES

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

[2] *Methods for subjective determination of transmiss- ion quality.* Recommendation ITU-T P.800, 1996.

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

[4] *General methods for the subjective assessment of sound quality* Rec. ITU-R BS.1284-1, 2003.

[5] S. George, S. Zielinski, F. Rumsey, P. Jackson, R. Conetta, M. Dewhirst, D. Meares and S. Bech, "Development and Validation of an Unintrusive Model for Predicting Sensation of Envelopment Arising from Surround Sound Recordings," in *Journal of the Audio Engineering Society*, December 2010.

[6] 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 4-7 2013.

[7] 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.

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

[9] B. De Man and J. D. Reiss, "APE: Audio Perceptual Evaluation toolbox for MATLAB" in *136th Convention of the Audio Engineering Society*, April 26-29 2014.

[10] Audio Research Labs, "ARL STEP: Essential Tool for Subjective Quality Evaluation", 2009. Available from: http://www.audioresearchlabs.com/step/