

WEB AUDIO EVALUATION TOOL: A BROWSER-BASED LISTENING TEST ENVIRONMENT

Nicholas Jillings Brecht De Man David Moffat Joshua D. Reiss
n.g.r.jillings@se14.qmul.ac.uk, {b.deman, d.j.moffat, joshua.reiss}@qmul.ac.uk
Centre for Digital Music, Queen Mary University of London

ABSTRACT

Perceptual evaluation tests where subjects assess certain qualities of different audio fragments are an integral part of audio and music research. These require specialised software, usually custom-made, to collect large amounts of data using meticulously designed interfaces with carefully formulated questions, and play back audio with rapid switching between different samples. New functionality in HTML5 included in the Web Audio API allows for increasingly powerful media applications in a platform independent environment. The advantage of a web application is easy deployment on any platform, without requiring any other application, enabling multiple tests to be easily conducted across locations. In this paper we propose a tool supporting a wide variety of easily configurable, multi-stimulus perceptual audio evaluation tests over the web with multiple test interfaces, pre- and post-test surveys, custom configuration, collection of test metrics and other features. Test design and setup doesn't require programming background, and results are gathered automatically using web friendly formats for easy storing of results on a server.

1. INTRODUCTION

Perceptual evaluation of audio plays an important role in a wide range of research on audio quality [1, 2], sound synthesis [3, 4], audio effect design [5], source separation [6, 7], music and emotion analysis [8, 9], and many others [10].

Table 1. Available audio perceptual evaluation tools

Name	Language	Ref.
APE	MATLAB	[11]
BeagleJS	HTML5/JS	[12]
HULTI-GEN	Max	[13]
MUSHRAM	MATLAB	[6]
Scale	MATLAB	[14]
WhisPER	MATLAB	[15]

Various listening test design tools are already available, see Table 1. A few other listening test tools, such as OPA-QUE [16] and GuineaPig [17], are described but not available to the public at the time of writing.

Many are MATLAB-based, useful for easily processing and visualising the data produced by the listening tests, but requiring MATLAB to be installed to run or - in the case of an executable created with MATLAB - at least create the test. Furthermore, compatibility is usually limited across different versions of MATLAB. Similarly, Max requires little or no programming background but it is proprietary software as well, which is especially undesirable when tests need to be deployed at different sites. More recently, BeagleJS [12] makes use of the HTML5 audio capabilities and comes with a number of predefined, established test interfaces such as ABX and MUSHRA [18]. BeagleJS provides a number of similar features including saving of test data to a web server. The main difference is that with BeagleJS, the configuration is done through writing a JavaScript file holding a JavaScript Object of the notation. Instead our presented system uses the XML document standard, which allows configuration outside of a web-centric editor. The results are also presented in XML again allowing 3rd party editors and programs to easily access. Finally, the presented system does not require web access to run, instead being deployed with a Python server script. This is particularly useful in studios where machines may not, by design, be web connected, or use in locations where web access is limited.

A browser-based perceptual evaluation tool for audio has a number of advantages. First of all, it doesn't need any other software than a browser, meaning deployment is very easy and cheap. As such, it can also run on a variety of devices and platforms. The test can be hosted on a central server with subjects all over the world, who can simply go to a webpage. This means that multiple participants can take the test simultaneously, potentially in their usual listening environment if this is beneficial for the test. Naturally, the constraints on the listening environment and other variables still need to be controlled if they are important to the experiment. Depending on the requirements a survey or a variety of tests preceding the experiment could establish whether remote participants and their environments are adequate for the experiment at hand.

The Web Audio API is a high-level JavaScript Application Programming Interface (API) designed for real-time processing of audio inside the browser through various pro-

cessing nodes¹. Various web sites have used the Web Audio API for creative purposes, such as drum machines and score creation tools², others from the list show real-time captured audio processing such as room reverberation tools and a phase vocoder from the system microphone. The BBC Radiophonic Workshop shows effects used on famous TV shows such as Doctor Who, being simulated inside the browser³. Another example is the BBC R&D personalised compressor which applies a dynamic range compressor on a radio station that dynamically adjusts the compressor settings to match the listener’s environment [19].

In contrast with the tools listed above, we aim to provide an environment in which a variety of multi-stimulus tests can be designed, with a wide range of configurability, while keeping setup and collecting results as straightforward as possible. For instance, the option to provide free-text comment fields allows for tests with individual vocabulary methods, as opposed to only allowing quantitative scales associated to a fixed set of descriptors. To make the tool accessible to a wide range of researchers, we aim to offer maximum functionality even to those with little or no programming background. The tool we present can set up a listening test without reading or adjusting any code, provided no new types of interfaces need to be created.

Specifically, we present a browser-based perceptual evaluation tool from which any kind of multiple stimulus audio evaluation tool where subjects need to rank, rate, select, or comment on different audio samples can be built. We also include an example of the multiple stimulus user interface included with the APE tool [11], which presents the subject with a number of axes on which a number of markers, corresponding to audio samples, can be moved to reflect any subjective quality, as well as corresponding comment boxes. However, other graphical user interfaces can be put on top of the engine that we provide with minimal or no modifications. Examples of this are the MUSHRA test [18], single or multiple stimulus evaluation with a two-dimensional interface (such as valence and arousal dimensions), or simple annotation (using free-form text, check boxes, radio buttons or drop-down menus) of one or more audio samples at a time. In some cases, such as method of adjustment, where the audio is processed by the user, or AB test, where the interface does not show all audio samples to be evaluated at once [20], the back end of the tool needs to be modified as well.

In the following sections, we describe the included interface in more detail, discuss the implementation, and cover considerations that were made in the design process of this tool.

2. INTERFACE

At this point, we have implemented the interface of the MATLAB-based APE (Audio Perceptual Evaluation) toolbox [11]. This shows one marker for each simultaneously evaluated audio fragment on one or more horizontal axes, that can be moved to rate or rank the respective fragments

in terms of any subjective property, as well as a comment box for every marker, and any extra text boxes for extra comments. The reason for such an interface, where all stimuli are presented on a single rating axis (or multiple axes if multiple subjective qualities need to be evaluated), is that it urges the subject to consider the rating and/or ranking of the stimuli relative to one another, as opposed to comparing each individual stimulus to a given reference, as is the case with e.g. a MUSHRA test [18]. As such, it is ideal for any type of test where the goal is to carefully compare samples against each other, like perceptual evaluation of different mixes of music recordings [21] or sound synthesis models [4], as opposed to comparing results of source separation algorithms [6] or audio with lower data rate [18] to a high quality reference signal.

The markers on the slider at the top of the page are positioned randomly, to minimise the bias that may be introduced when the initial positions are near the beginning, end or middle of the slider. Another approach is to place the markers outside of the slider bar at first and have the subject drag them in, but the authors believe this doesn’t encourage careful consideration and comparison of the different fragments as the implicit goal of the test becomes to audition and drag each fragment in just once, rather than to compare all fragments rigorously.

See Figure 1 for an example of the interface.

3. ARCHITECTURE

The tool uses entirely client side processing utilising the new HTML5 Web Audio API, supported by most major web browsers. The API allows for constructing audio processing elements and connecting them together to produce a high quality, real time signal process to manipulate audio streams. The API supports multichannel processing and has an accurate playback timer for precise, scheduled playback control. The API is controlled through the browser JavaScript engine and is therefore highly configurable. Processing is all performed in a low latency thread separate from the main JavaScript thread, so there is no blocking due to real time processing.

The web tool itself is split into several files to operate:

- `index.html`: The main index file to load the scripts, this is the file the browser must request to load.
- `core.js`: Contains global functions and object prototypes to define the audio playback engine, audio objects and loading media files
- `ape.js`: Parses setup files to create the interface as instructed, following the same style chain as the MATLAB APE Tool [11].

The HTML file loads the `core.js` file along with a few other ancillary files (such as the jQuery JavaScript extensions⁴), at which point the browser JavaScript begins to execute the on-page instructions, which gives the URL of the test setup XML document (outlined in Section 5). `core.js` parses this document and executes the functions in `ape.js` to build the web page. The reason for separating these two files is to allow for further interface

¹ <http://webaudio.github.io/web-audio-api/>

² <http://webaudio.github.io/demo-list/>

³ <http://webaudio.prototype.bbc.co.uk/>

⁴ <http://jquery.com/>

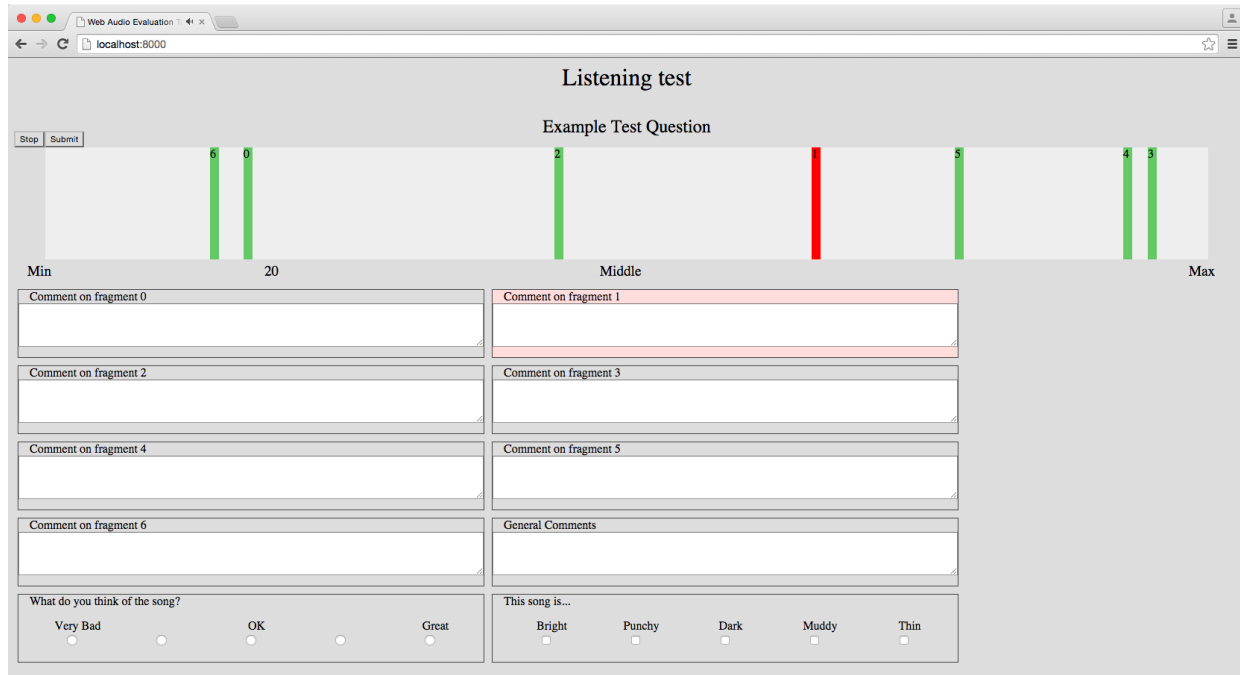


Figure 1. Example interface, with one axis, seven fragments, and text, radio button and check box style comments.

designs (such as MUSHRA [18] or 2D rating [20]) to be used, which would still require the same underlying core functions outlined in `core.js`.

The `ape.js` file has several main functions but the most important are documented here. `loadInterface(xmlDoc)` is called to decode the supplied project document in respect for the interface specified and define any global structures (such as the slider interface). It also identifies the number of pages in the test and randomises the order, if specified to do so. This is the only mandatory function in any of the interface files as this is called by `core.js` when the document is ready. `core.js` cannot ‘see’ any interface specific functions and therefore cannot assume any are available. Therefore `loadInterface(xmlDoc)` is essential to set up the entire test environment. Because the interface files are loaded by `core.js` and because the functions in `core.js` are global, the interface files can ‘see’ the `core.js` file and can therefore not only interact with it, but also modify it.

Each test page is loaded using `loadTest(id)` which performs two major tasks: to populate the interface with the slider elements and comment boxes; and secondly to instruct the `audioEngine` to load the audio fragments and construct the backend audio graph. `loadTest(id)` also instructs the audio engine in `core.js` to create the `audioObject`. These are custom audio nodes, one representing each audio element specified in each page. They consist of a `bufferSourceNode` (a node which holds a buffer of audio samples for playback) and a `gainNode`, both of which are Web Audio API Nodes. Various functions are applied, depending on which metrics are enabled, to record the interaction with the audio element. These nodes are then connected to the `audioEngine` (itself a custom web audio node) containing a `gainNode` (where the various `audioObjects` connect to) for summation before passing the output

to the `destinationNode`, a permanent node of the Web Audio API created as the master output. Here, the browser then passes the audio information to the system.

When an `audioObject` is created, it is given the URL of the audio sample to load. This is downloaded into the browser asynchronously using the `XMLHttpRequest` object, which downloads any file into the JavaScript environment for further processing. This is particularly useful for the Web Audio API because it supports downloading of files in their binary form for decoding. Once downloaded the file is decoded using the Web Audio API offline decoder. This uses the browser available decoding schemes to decode the audio files into raw float32 arrays, which are in turn passed to the relevant `audioObject` for playback.

Once each page of the test is completed, identified by pressing the Submit button, the `pageXMLSave(testId)` is called to store all of the collected data until all pages of the test are completed. After the final test and any post-test questions are completed, the `interfaceXMLSave()` function is called. This function generates the final XML file for submission as outlined in Section 5.

4. SUPPORT AND LIMITATIONS

Different browsers support a different set of audio file formats and are not consistent in any format. Currently the Web Audio API is best supported in Chrome, Firefox, Opera and Safari. All of these support the use of the uncompressed WAV format. Although not a compact, web friendly format, most transport systems are of a high enough bandwidth this should not be a problem. Ogg Vorbis is another well supported format across the four supported major desktop browsers, as well as MP3 (although Firefox may not support all MP3 types⁵). One issue of the Web

⁵ https://developer.mozilla.org/en-US/docs/Web/HTML/Supported_media_formats

Audio API is that the sample rate is assigned by the system sound device, rather than requested and does not have the ability to request a different one. As the sampling rate and the effect of resampling may be critical for some listening tests, the default operation when an audio file is loaded with a different sample rate to that of the system is to convert the sample rate. To provide a check for this, the desired sample rate can be supplied with the setup XML and checked against. If the sample rates do not match, a browser alert window is shown asking for the sample rate to be correctly adjusted. This happens before any loading or decoding of audio files so the browser will only be instructed to fetch files if the system sample rate meets the requirements, avoiding multiple requests for large files until they are actually needed.

5. INPUT AND RESULT FILES

The setup and result files both use the common XML document format to outline the various parameters. The setup file determines the interface to use, the location of audio files, the number of pages and other parameters to define the testing environment. Having one document to modify allows for quick manipulation in a ‘human readable’ form to create new tests, or adjust current ones, without needing to edit multiple web files. Furthermore, we also provide a simple web page to enter all these settings without needing to manipulate the raw XML. An example of such an XML document is presented below.

```
<?xml version="1.0" encoding="utf-8"?>
<BrowserEvalProjectDocument>
  <setup interface="APE" projectReturn="/save"
    randomiseOrder='false' collectMetrics='true'
  >
    <PreTest>
      <question id="location" mandatory="
        true">Please enter your location
      </question>
      <number id="age" min="0">Please enter
        your age</number>
    </PreTest>
    <PostTest>
      <statement>Thank you for taking this
        listening test!</statement>
    </PostTest>
    <Metric>
      <metricEnable>testTimer</metricEnable>
    >
      <metricEnable>elementTimer</
        metricEnable>
      <metricEnable>elementInitialPosition<
        /metricEnable>
      <metricEnable>elementTracker</
        metricEnable>
      <metricEnable>elementFlagListenedTo</
        metricEnable>
      <metricEnable>elementFlagMoved</
        metricEnable>
    </Metric>
    <interface>
      <anchor>20</anchor>
      <reference>80</reference>
    </interface>
  </setup>
  <audioHolder id="test-0" hostURL="example_eval/"
    randomiseOrder='true'>
    <interface>
      <title>Example Test Question</title>
      <scale position="0">Min</scale>
      <scale position="100">Max</scale>
      <commentBoxPrefix>Comment on fragment
        </commentBoxPrefix>
    </interface>
    <audioElements url="1.wav" id="elem1"/>
    <audioElements url="2.wav" id="elem2"/>
```

```
<audioElements url="3.wav" id="elem3"/>
<CommentQuestion id="generalExperience"
  type="text">General Comments</
  CommentQuestion>
<PreTest/>
<PostTest>
  <question id="songGenre" mandatory="
    true">Please enter the genre of
    the song.</question>
</PostTest>
</audioHolder>
</BrowserEvalProjectDocument>
```

5.1 Setup and configurability

The setup document has several defined nodes and structure which are documented with the source code. For example, there is a section for general setup options where any pre-test and post-test questions and statements can be defined. Pre- and post-test dialogue boxes allow for comments or questions to be presented before or after the test, to convey listening test instructions, and gather information about the subject, listening environment, and overall experience of the test. In the example set up document above, a question box with the id ‘location’ is added, which is set to be mandatory to answer. The question is in the PreTest node meaning it will appear before any testing will begin. When the result for the entire test is shown, the response will appear in the PreTest node with the id ‘location’ allowing it to be found easily, provided the id values are meaningful.

We try to cater to a diverse audience with this toolbox, while ensuring it is simple, elegant and straightforward. To that end, we currently include the following options that can be easily switched on and off, by setting the value in the input XML file.

- **Snap to corresponding position:** When enabled and a fragment is playing, the playhead skips to the same position in the next fragment that is clicked. Otherwise, each fragment is played from the start.
- **Loop fragments:** Repeat current fragment when end is reached, until the ‘Stop’ or ‘Submit’ button is clicked.
- **Comments:** Displays a separate comment box for each fragment in the page.
- **General comment:** Create additional comment boxes to the fragment comment boxes, with a custom question and various input formats such as checkbox or radio.
- **Resampling:** When this is enabled, fragments are re-sampled to match the subject’s system’s sample rate (a default feature of the Web Audio API). When it is not, an error is shown when the system does not match the requested sample rate.
- **Randomise page order:** Randomises the order in which different ‘pages’ are presented.
- **Randomise fragment order:** Randomises the order and numbering of the markers and comment boxes corresponding to the fragments. Fragments are referenced to their given ID so referencing is possible (such as ‘this is much brighter than fragment 4’).
- **Require (full) playback:** Require that each fragment has been played at least once, partly or fully.
- **Require moving:** Require that each marker is moved (dragged) at least once.

- **Require comments:** Require the subject to write a comment for each fragment.
- **Repeat test:** Number of times each page in the test should be repeated (none by default), to allow familiarisation with the content and experiment, and to investigate consistency of user and variability due to familiarity. These are all gathered before shuffling the order so repeated tests are not back-to-back if possible.
- **Returning to previous pages:** Indicates whether it is possible to go back to a previous ‘page’ in the test.
- **Lowest rating below [value]:** To enforce a certain use of the rating scale, it can be required to rate at least one sample below a specified value.
- **Highest rating above [value]:** To enforce a certain use of the rating scale, it can be required to rate at least one sample above a specified value.
- **Reference:** Allows for a separate sample (outside of the axis) to be the ‘reference’, which the subject can play back during the test to help with the task at hand [18].
- **Hidden reference/anchor:** Whether or not an explicit ‘reference’ is provided, the ‘hidden reference’ should be rated above a certain value [18] - this can be enforced. Similarly, a ‘hidden anchor’ should be rated lower than a certain value [18].
- **Show scrub bar:** Display a playhead on a scrub bar to show the position in the current fragment.

When one of these options is not included in the setup file, they assume a default value. As a result, the input file can be kept very compact if default values suffice for the test.

5.2 Results

The results file is dynamically generated by the interface upon clicking the ‘Submit’ button. This also executes checks, depending on the setup file, to ensure that all fragments have been played back, rated and commented on. The XML output returned contains a node per fragment and contains both the corresponding marker’s position and any comments written in the associated comment box. The rating returned is normalised to be a value between 0 and 1, normalising the pixel representation of different browser windows. The results also contain information collected by any defined pre/post questions. An excerpt of an output file is presented below detailing the data collected for a single audioElement.

```
<browserevaluationresult>
  <datetime>
    <date year="2015" month="5" day="28">
      2015/5/28</date>
    <time hour="13" minute="19" secs="17">
      13:19:17</time>
    </datetime>
  <pretest>
    <comment id="location">Control Room</comment>
  </pretest>
  <audioholder>
    <pretest></pretest>
    <posttest>
      <comment id="songGenre">Pop</comment>
    </posttest>
    <metric>
      <metricresult id="testTime">813.32</metricresult>
    </metric>
  </audioelement id="elem1">
```

```
<comment>
  <question>Comment on fragment 1
  </question>
  <response>Good, but vocals too
  quiet.</response>
</comment>
<value>0.639010989010989</value>
<metric>
  <metricresult id="elementTimer">
    >111.05</metricresult>
  <metricresult id="
    elementTrackerFull">
    <timepos id="0">
      <time>61.60</time>
      <position>0.6390</position>
    </timepos>
  </metricresult>
  <metricresult id="
    elementInitialPosition">
    0.6571</metricresult>
  <metricresult id="
    elementFlagListenedTo">
    true</metricresult>
</metric>
</audioelement>
</browserevaluationresult>
```

Each page of testing is returned with the results of the entire page included in the structure. One audioelement node is created per audio fragment per page, along with its ID. This includes several child nodes including the rating between 0 and 1, the comment, and any other collected metrics including how long the element was listened for, the initial position, and boolean flags showing if the element was listened to, moved and commented on. Furthermore, each user action (manipulation of any interface element, such as playback or moving a marker) can be logged along with a the corresponding time code. We also store session data such as the time the test took place and the duration of the test. We provide the option to store the results locally, and/or to have them sent to a server.

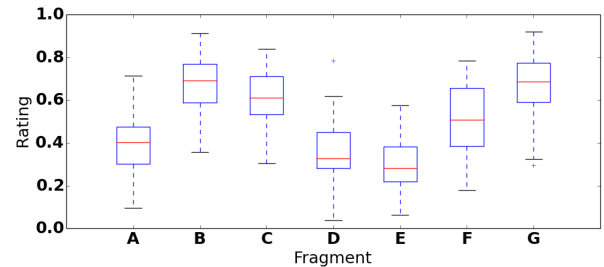


Figure 2. An example boxplot showing ratings by different subjects on fragments labeled ‘A’ through ‘G’.

Python scripts are included to easily store ratings and comments in a CSV file, and to display graphs of numerical ratings (see Figure 2) or visualise the test’s timeline. Visualisation of plots requires the free matplotlib library⁶.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented an approach to creating a browser-based listening test environment that can be used for a variety of types of perceptual evaluation of audio. Specifically, we discussed the use of the toolbox in the context of assessment of preference for different production practices, with identical source material. The purpose

⁶ <http://matplotlib.org>

of this paper is to outline the design of this tool, to describe our implementation using basic HTML5 functionality, and to discuss design challenges and limitations of our approach. This tool differentiates itself from other perceptual audio tools by enabling web technologies for multiple participants to perform the test without the need for proprietary software such as MATLAB. The tool also allows for any interface to be built using HTML5 elements to create a variety of dynamic, multiple-stimulus listening test interfaces. It enables quick setup of simple tests with the ability to manage complex tests through a single file. Finally it uses the XML document format to store the results allowing for processing and analysis of results in various third party software such as MATLAB or Python.

Further work may include the development of other common test designs, such as MUSHRA [18], 2D valence and arousal/activity [9], and others. We will add functionality to assist with setting up large-scale tests with remote subjects, so this becomes straightforward and intuitive. In addition, we will keep on improving and expanding the tool, and highly welcome feedback and contributions from the community.

The source code of this tool can be found on `code.soundsoftware.ac.uk/projects/webaudioevaluationtool`.

7. REFERENCES

- [1] M. Schoeffler and J. Herre, "About the impact of audio quality on overall listening experience," in *Proceedings of the 10th Sound and Music Computing Conference*, 2013, pp. 48–53.
- [2] R. Repp, "Recording quality ratings by music professionals," in *Proceedings of the 2006 International Computer Music Conference*, 2006, pp. 468–474.
- [3] A. de Götzen, E. Sikström, F. Grani, and S. Serafin, "Real, foley or synthetic? An evaluation of everyday walking sounds," in *Proceedings of SMC 2013 : 10th Sound and Music Computing Conference*, 2013.
- [4] G. Durr, L. Peixoto, M. Souza, R. Tanoue, and J. D. Reiss, "Implementation and evaluation of dynamic level of audio detail," in *Audio Engineering Society Conference: 56th International Conference: Audio for Games*, 2015.
- [5] B. De Man and J. D. Reiss, "Adaptive control of amplitude distortion effects," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, 2014.
- [6] E. Vincent, M. G. Jafari, and M. D. Plumbley, "Preliminary guidelines for subjective evaluation of audio source separation algorithms," in *UK ICA Research Network Workshop*, 2006.
- [7] J. D. Reiss and C. Uhle, "Determined source separation for microphone recordings using IIR filters," in *129th Convention of the Audio Engineering Society*, 2010.
- [8] Y. Song, S. Dixon, M. T. Pearce, and G. Fazekas, "Using tags to select stimuli in the study of music and emotion," *Proceedings of the 3rd International Conference on Music & Emotion (ICME)*, 2013.
- [9] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *Proceedings of the 10th International Society for Music Information Retrieval (ISMIR2009)*, 2009, pp. 621–626.
- [10] A. Friberg and A. Hedblad, "A comparison of perceptual ratings and computed audio features," in *Proceedings of the 8th Sound and Music Computing Conference*, 2011, pp. 122–127.
- [11] B. De Man and J. D. Reiss, "APE: Audio Perceptual Evaluation toolbox for MATLAB," in *136th Convention of the Audio Engineering Society*, 2014.
- [12] S. Kraft and U. Zölzer, "BeagleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality," in *Linux Audio Conference, Karlsruhe, DE*, 2014.
- [13] C. Gribben and H. Lee, "Toward the development of a universal listening test interface generator in Max," in *138th Convention of the Audio Engineering Society*, 2015.
- [14] A. V. Giner, "Scale - a software tool for listening experiments," in *AIA/DAGA Conference on Acoustics, Merano (Italy)*, 2013.
- [15] S. Ciba, A. Wlodarski, and H.-J. Maempel, "Whisper - A new tool for performing listening tests," in *126th Convention of the Audio Engineering Society*, 2009.
- [16] J. Berg, "OPAQUE - A tool for the elicitation and grading of audio quality attributes," in *118th Convention of the Audio Engineering Society*, 2005.
- [17] J. Hynninen and N. Zacharov, "GuineaPig - A generic subjective test system for multichannel audio," in *106th Convention of the Audio Engineering Society*, 1999.
- [18] *Method for the subjective assessment of intermediate quality level of coding systems*. Recommendation ITU-R BS.1534-1, 2003.
- [19] A. Mason, N. Jillings, Z. Ma, J. D. Reiss, and F. Melchior, "Adaptive audio reproduction using personalized compression," in *Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology - Cinema, Television and the Internet*, 2015.
- [20] S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application*. John Wiley & Sons, 2007.
- [21] B. De Man, M. Boerum, B. Leonard, G. Massenburg, R. King, and J. D. Reiss, "Perceptual evaluation of music mixing practices," in *138th Convention of the Audio Engineering Society*, 2015.