Variability in perceptual evaluation of HRTFs

David Schönstein\(^1\) and Brian F.G. Katz\(^2\)

\(^1\) ARKAMYS, 31 rue Pouchet, 75017, Paris, France
dschenstein@arkamys.com

\(^2\) LIMSI-CNRS, BP 133, Université Paris XI, F-91403, Orsay Cedex, France
brian.katz@limsi.fr

ABSTRACT

The implementation of the *head-related transfer function* (HRTF) is key to binaural rendering applications. An HRTF evaluation and selection is often required when individual HRTFs are not available. This study examines the variability in perceptual evaluations of HRTFs using a listening test. A set of six different HRTFs was selected and was then used in a listening test based on the standardised MUSHRA method for evaluating audio quality. A total of six subjects participated, each having their own recorded HRTFs available. Subjects performed five repetitions of the listening test. While conclusive HRTF judgments were evident, a significantly large degree of variance was found. The effect of listener expertise on variability in perceptual judgments was also analysed.

1. INTRODUCTION

Humans perceive qualities of a sound source in space such as direction and distance using cues that arise at each ear embedded in the acoustic filtering of the source by the auditory periphery (see [1] for a review). Knowledge of the acoustic filtering, known as the *head-related transfer function* (HRTF), for a particular listener can allow for a simulation of sound sources in space, known as a *binaural synthesis*. In order to achieve high fidelity renderings in *virtual auditory space* (VAS) many studies have shown that HRTFs need to be individualised for the listener, particularly in terms of accuracy of localization [2-5]. Some studies however, using listening tests, have shown that a listener’s own HRTFs are not always judged as optimal [6].

Applications of binaural synthesis such as in the domains of teleconferencing, hearing aids, video gaming and general immersion listening are becoming more prevalent and as a consequence there is an increased need and interest to find appropriate means to evaluate the quality of a rendering in VAS. A method for specifically evaluating different HRTFs, which is more geared towards applications for the general public than localization tasks, is evaluation with use of a listening test. Listening tests have been used in the literature to evaluate HRTFs [6-8] with little investigation as to how reliable or conclusive listener judgments might be, i.e. how repeatable the results are.
In this study a concise methodology for conducting listening tests using different HRTFs is proposed, along with an analysis of repeatability. Techniques for a pre- and post-selection, based on listener expertise, of subjects is also assessed.

2. Method

For the listening test in this study subjects were asked to judge a total of six different HRTFs, termed profiles, based on three different attributes. This is in contrast to more traditionally direct localisation tests. The listening test was repeated five times for each subject. The order in which the six profiles were presented was randomly varied for each trial. Subjects were explicitly told that the six profiles could be different for each trial.

2.1. Subjects

A total of six subjects, including the first author, participated in the listening test. Subjects were categorized in terms of their level of expertise as an assessor. The categorization was based on the subjects' responses to an HTML web-based questionnaire resembling that of [9]. The questionnaire assessed whether the subject:

- Was a musician and the level of experience.
- Has had experience listening to binaural syntheses.
- Has had experience with listening tests using binaural syntheses.

Based on subjects' responses to these three criterions they were classified according to definitions taken from the ISO standard 8586-2 [10], applied by the food industry and recommended for adoption in the field of audio by [11]. The mentioned definitions are displayed in Table 1. Standardized definitions were employed due to the inconsistency in the audio literature with respect to the terms untrained, naïve, experienced and expert [12]. A subject was categorized as an initiated assessor if they had no experience with binaural syntheses, were not musicians, and had only performed the training for the current listening test. Subjects that had had exposure to binaural syntheses and listening tests using binaural syntheses, had performed the training but were not musicians were categorized as a selected assessor. Subjects with the same level of expertise as the initiated assessor that were musicians were categorized as experts. This distinction between musicians and non-musicians was made due to the knowledge that musicians have highly specialized auditory skills as shown in [13], and notably enhanced auditory cortical representations for specific aspects of audition such as timbre [14]. The first author was assigned the category expert assessor due to the extensive experience in making judgments specific to this listening test, which was comparable to no other subject.

<table>
<thead>
<tr>
<th>Assessor category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessor</td>
<td>Any person taking part in a sensory test</td>
</tr>
<tr>
<td>Naïve assessor</td>
<td>A person who does not meet any particular criterion</td>
</tr>
<tr>
<td>Initiated assessor</td>
<td>A person who has already participated in a sensory test</td>
</tr>
<tr>
<td>Selected assessor</td>
<td>Assessor chosen for his/her ability to carry out a sensory test</td>
</tr>
<tr>
<td>Expert</td>
<td>In the general sense, a person who through knowledge or experience has competence to give an opinion in the field about which he/she is consulted (Please note that the term expert does not provide any indication regarding the qualification or suitability of the individual to perform listening tests)</td>
</tr>
<tr>
<td>Expert assessor</td>
<td>Selected assessor with a high degree of sensory sensitivity and experience in sensory methodology, who is able to make consistent and repeatable sensory assessments of various products</td>
</tr>
<tr>
<td>Specialised expert assessor</td>
<td>Expert assessor who has additional experience as a specialist in the product and/or process and/or marketing, and who is able to perform sensory analysis of the product and evaluate or predict effects of variations relating to raw materials, recipes, processing, storage, ageing, and so on</td>
</tr>
</tbody>
</table>

All subjects presented positive results for a pure tone audiometry test (ISO [15]).

2.2. Test Stimuli

The test stimuli were created from a randomly generated white Gaussian noise. Each noise burst used in the
listening test was freshly generated in order to avoid any learning effects of the signal. The noise signal had a duration of 200 ms and was ramped by applying a raised cosine to the first and last 10 ms. A stimulus duration of 200 ms was chosen with the knowledge that localization performance peaks and plateaus at approximately 100 ms for a noise stimulus [16].

A set of five HRTF measurements were used in this study, taken from the public database LISTEN [17] of 46 HRTFs. The five HRTFs were selected based on the results of a previous listening test in which 45 of the 46 subjects in the database judged all the HRTFs including their own as either excellent, average or bad [18]. All possible combinations of five HRTFs were evaluated and the subset of HRTFs that satisfied the largest number of subjects (42 out of 45) in terms of excellent judgements was selected. In order to ensure that the HRTFs selected in the subset were not poorly rated globally, possibly relating to an error in the measurements, the subset with the largest number of excellent ratings for individual HRTFs was chosen. The minimum number of subjects that rated a particular HRTF as excellent in the subset was nine. In addition to the five mentioned HRTFs, each subject's own HRTF was used in the listening test as explained in section 2.3.

All HRTFs were individually converted to directional transfer functions (DTFs) [19] by dividing out the spatially weighted average of all 187 measurement positions. The DTFs were normalised in root-mean-square across all positions for the left and right ear for all subjects. The DTFs were decomposed into the minimum phase and all-pass components. The all-pass component was replaced with a pure delay that represented the inter-aural time delay (ITD) of each subject. The ITD was calculated using the maximum inter-aural cross-correlation of the energy envelopes. For each subject, identical ITD values were used for each position, removing ITD as a test variable.

The two stimuli presented in the listening test were generated by convolving the noise signal with the DTF for two specified trajectory sequences (Figure 1):

- (A) 3 positions of elevation -15°, 0° and 15° and azimuth 135° (hoop coordinates [20]) starting at the position with elevation -15° and ending at elevation 15°.
- (B) 3 positions of azimuth 0°, 15° and 30° and elevation 0° starting at the position with azimuth 0° and ending at azimuth 30°.

Figure 1 Trajectories used for the two test stimuli. White circles represent rendered positions. Black circles represent the position directly in front of the listener, included in the figure as a reference.

Only 3 positions were used in each trajectory out of a possible 187 measurements in space for this listening test. The choice of such a small number of positions to be judged by the listener is due to the fact that a large number of positions would have resulted in stimuli of long duration. Listeners are known to not make reliable judgments when performing a perceptual average of quality for stimuli with a long duration, as explained by [21]. Listener’s judgments are biased toward the quality of that part of the stimulus that is auditioned last. This psychological effect related to the dominance of short-term memory over long-term memory, termed the recency effect, has been shown in studies such as [22] for the evaluation of telephone speech quality. The effect of reducing the number of positions being judged, along with use of attribute judgments as opposed to a global localization judgment, were informally tested and found to reduce variance across a number of trials.

Each of the two stimuli, for each trajectory, were created for the common set of five HRTFs and for the subject's own HRTFs. The stimuli were generated
separately for each subject taking into account the subject's ITDs. The minimum phase and ITD generation procedure was also performed on the subject's own HRTFs so there would be no difference in the manner in which the binaural syntheses were created for all the HRTFs being judged.

The stimulus level was adjusted using what is termed sensation level [23]. Sensation level refers to the decibel gain above the audible threshold for a continuous binaural synthesis of the noise signal at azimuth 0° and elevation 0° (i.e. directly in front of the listener) played to the subject over headphones. The sensation level threshold was calculated for each subject by beginning at a level that was inaudible and incrementally increasing the level 1 dB at a time. Once the threshold was found, a sensation level of 45 dB was chosen based on findings that show a negative level effect for accuracy of judgements of sound-source elevation above this level. The negative level effect has been confirmed by [16] for levels above ~60 dB SPL. Limiting the gain of the stimulus was also important in order to avoid level adaptation and the acoustical reflex (Stapedius reflex) usually observed at levels of about 70 dB SPL.

2.3. Test Method

The listening test method used in this study was based on the "multistimulus test with hidden reference and anchor" (MUSHRA) test [24], standardized in ITU-R BS.1534-1 [25]. This listening test method was developed in order to evaluate audio systems exhibiting intermediate levels of audio quality. It is recommended that no more than 15 items be included in any one trial. For these reasons the MUSHRA test method was seen as being well adapted for the perceptual evaluation of HRTFs. As the name suggests, there is a hidden reference as one of the stimuli being assessed, along with a signified reference and anchor. The reference was in this case a binaural synthesis using the subject's own HRTFs. The anchor was a degraded version of the reference; 3.5 kHz low-pass filtered using a 200-point FIR filter in compliance with the MUSHRA standard.

Listeners were asked to record their judgments on a continuous scale, with the end point descriptors not definable and well definable, for the following 3 attributes:

- **Sense of direction.** This attribute described how well the direction of the sound source could be defined.

This attribute did not relate to the correlation of the position of the sound source relative to the reference, but rather whether the directions of the sounds in trajectory A were well defined and distinct. This attribute in essence related to whether the position of the source was diffuse (not definable) or clearly at a specific position in space (well definable).

- **Sense of distance.** This attribute described how well the distance between the sound source and the listener could be defined for trajectory A. This attribute did not relate to the perceived distance of the sound source, but rather how well the listener could perceive it as coming from a certain distance.

- **Front image quality.** This attribute described how well the percept of the sound coming from in front of the listener could be defined for trajectory B. This attribute did not relate to the correlation of the position of the sound source relative to the reference, but rather whether the sound source was perceived as coming from in front of the listener. It was specified that for a sound source that was diffuse a judgement be made at the not definable end point of the scale. For a sound source that was perceived clearly in front of the listener it was specified that a judgement be made at the well defined end point of the scale, with any sound source that had a position that was well defined but not in front of the listener be judged somewhere between these end points depending on the perceived position.

The use of three attributes was seen as the upper limit for a listening test of this type given that each trial was not to take longer than 30 minutes. The first two attributes (sense of distance and sense of direction) were taken from the results of a descriptive analysis experiment designed to explore the perceptual characteristics of sound reproduced over headphones [26]. These two attributes were taken from a set of four attributes relating to localization in the mentioned analysis, seen as being the most significant in terms of applications for binaural synthesis for the general public. The third attribute (front image quality) was based on the recommendation in ITU-R BS.1116 and the attribute used in [27] for subjective assessment of surround sound processing algorithms. [27] showed that conclusive results could be obtained with this attribute but not for a more generalised spatial attribute termed spatial impression. It was also evident from the same study that conclusive results could not be obtained for hedonic judgments. This finding is in line with
arguments by [21] that highlight the lack of reliability for hedonic judgments. It is for this reason that no hedonic judgments were included in this listening test since the goal was to reduce variability across repeated trials. Despite the fact that hedonic evaluations were not used in this study, the attributes were chosen with commercial applications in mind focusing on quality, rather than precision in localisation. The use of these three attributes was seen as a suitable trade-off between experimental efficiency, potential user confusion, and the need for relevant assessment information.

2.4. Test Procedure

The listening test was performed in an acoustically damped listening booth. The test stimuli were presented over tubephones (ER-2, Etymotic Research) specifically designed for psychoacoustic tests. The tubephones are designed to produce an approximately flat frequency response, within 3 dB, at the human eardrum over the frequency range of 200 Hz to 16 kHz. No ear-canal resonance synthesis was included in this study. This model of tubephones has been validated in terms of accuracy in localisation tests in a number of studies [28-30].

The trajectory A was presented for judgements of the first two attributes (sense of direction and sense of distance), while trajectory B was presented for judgements of the third attribute (front image quality). The graphical user interface (GUI), shown in Figure 2 was designed so that judgements for all six profiles would be made on the same sliding scale for each attribute. Subjects selected a profile to judge from a drop-down menu and once it had been judged for all three attributes the responses were registered by pressing a button. After registering the judgements for a particular profile, its corresponding number (a value between 1 and 6) was displayed on the sliding scale. Playback of any of the six profiles, presented as buttons numbered from 1 to 6, was permitted at any time during the listening test. Subjects could only play one trajectory for one of the profiles at a time. The numbered buttons on the left of the GUI were for trajectory A and the buttons on the right were for trajectory B. Along with the 6 different profiles, there was a button for the specified reference and anchor. Each profile corresponded to a continuous loop of the signal noise convolved with the specified DTF for the given trajectory, as was the case for the reference and anchor (except that the anchor was a degraded version of the reference). Subjects could make judgements of the profiles in any order and return to a previously judged profile to change a response.

Each subject was individually familiarized with the equipment and the test procedure. The trajectories were displayed as they are presented in Figure 1. Particular attention was paid to the definitions of the attributes to be used. It was highlighted that judgments were not to be made with respect to the position of the sound source for the three attributes, but rather with respect to whether the attribute was definable as specified in section 2.3.

Given the known importance of listener training, due to its ability to increase listeners’ sensitivity to auditory characteristics and ability to reliably judge their perceptions [31], all subjects participated in a training session before starting the five repetitions. Subjects were given as long as they needed to familiarize themselves with the interface. Subjects were told to continue listening to the six profiles until they were able to discern the differences between them for each attribute. No objective evaluation of the training, in
terms of whether the subjects could discern the differences among the profiles, was performed.

3. RESULTS

3.1. Subjective reports

Subjects responded unanimously via a questionnaire following the listening test that they found the task difficult. Subjects reported that the attribute sense of distance was the most challenging to evaluate due to difficulty in judging the distance of the sound source, with some subjects admitting that judgments were almost random. This finding is in line with studies that have shown a fairly large degree of variability in judgements for a given sound source distance in VAS [32]. Two of the subjects, including the author of this paper, reported to having been aware that one of the HRTFs being judged was a hidden version of the reference. None of the subjects, except the first author of this paper, reported to have been aware that the HRTFs used in each trial were the same. The first author of this paper was included in the analysis of the results despite the biases stated due to the interest of having an expert assessor. Most of the HRTFs judged were not perceived by the subjects as coming from in front for trajectory B (Figure 1).

3.2. Repeatability

ISO 5497 [33] defines repeatability as the closeness in agreement between mutually independent test results obtained under conditions where mutually independent test results are obtained with the same method on identical test material in the same laboratory by the same operator using the same equipment within short intervals of time. This definition of repeatability has been strictly adhered to except with respect to the laboratory used, as the listening tests were conducted at two different locations, however the setting of both were for all intents and purposes identical; they were both acoustically damped and isolated chambers. The measure of repeatability was understood to be at the level of the listener for this study.

Globally the variation in perceptual judgments of the different HRTFs across the five randomised trials was significant. Scaling differences, which exist when listeners use different amounts of the scale for judgments, were observed between subjects. Methods for compensating for these differences among subjects have been assessed extensively in the field of descriptive sensory analysis in the food industry [34]. The simplest of these methods has been employed in this study in an effort to compare the repeatability of HRTF judgments. A scaling of the judgements based on the standard deviation for each listener and attribute was performed. More precisely, the judgment data for each subject was scaled to unit variance and zero mean. Figure 4 shows the HRTF judgments using the standardized data for each attribute. Figure 4(A), 4(B) and 4(C) show judgments for the attributes sense of direction, sense of distance and front image quality respectively, presented as box plots. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles and the whiskers extend to the most extreme data points. The subjects are represented on the horizontal axis and the judged HRTFs are marked by shade. The subject’s own HRTF is labelled Subject in the legend. Subjects are ordered from left to right from most variance to least, calculated across all attributes. The standardized judgment values are represented on the vertical axis with the labels Well and Not corresponding to the descriptor end points well definable and not definable.
Figure 2 Box plots representing the spread of judgments for HRTFs for each attribute as labelled on the vertical axis. Subjects are represented on the horizontal axis and the HRTF rated is determined by shade and displayed on the legend. Judgments are represented on the vertical axis between the labels Well and Not corresponding to the descriptor end points well defined and not definable respectively.
From the results it can be seen that the spread of judgments for a particular HRTF varies significantly between subjects (section 3.3). Also evident is the correlation between judgements for attributes *sense of direction* and *sense of distance*; the mean correlation coefficient across all subjects was 0.74. The correlation between the other pairs of attributes was less pronounced with a mean value of 0.46. The attribute that exhibited the least variance in judgments between trials for all subjects was that of *front image quality*. The attributes *sense of direction* and *sense of distance* displayed similar amounts of variance. These findings taken together with the results from the questionnaire would suggest that the attribute *sense of distance* is not a particularly reliable one and most probably redundant when also judging *sense of direction*. Also noteworthy is the fact that, for subjects whom the variance was low, the subject’s own HRTFs (the hidden reference) were not always judged as being the most *well defined*. This type of result has also been shown in a study by [6] where the subject’s own HRTFs were not judged as being perceived as the most natural sounding. This result has a direct impact on methods that aim to provide a listener with individualised HRTFs [35-37], since it is possible that an HRTF other than the subject’s own will be judged better. What characteristics of a particular HRTF lead to a favourable judgment is still not entirely clear in this context. Most studies that seek to single out the important aspects of the HRTF have assessed only accuracy in localization tasks [29, 38-40].

### 3.3. Subject expertise

The variance of HRTF judgments for a particular attribute across all five trials was calculated from the standardized data for each subject. An analysis of variance (ANOVA) was performed on the variance values for each attribute to test the effect of subject on repeatability. For all three attributes subject was a statistically significant factor (p < 0.01). Subject mean variance across all judged HRTFs is shown in Figure 5 along with the results from section 2.1 corresponding to listener’s level of expertise. The mean variance is represented on the horizontal axis and the assessor category on the vertical axis. The labels on the vertical axis correspond to the assessor categories from *initiated assessor* to *expert assessor* (IA and EA respectively) taken from Table 1. The shade of the points along with the legend, determines what attribute the mean variances correspond to. The points in the figure are also labelled in terms of the subject making the judgment. Each circle on the figure thus represents the variance across all trials for one attribute for a particular subject. Given the assignment of assessor category from Table 1, the correlation between level of expertise and variance was measured. The correlation coefficients for the attributes *sense of direction*, *sense of distance* and *front image quality* were 0.73, 0.69 and 0.77 respectively. These findings suggest experienced listeners are more consistent in the evaluation of HRTFs, particularly for the third attribute mentioned, a finding supported by studies that look at the benefits of subject training [31]. It is also noteworthy that some subjects showed the same amount of variance across all three attributes and other subjects had large differences. This most probably relates to how difficult subjects found judgments for a particular attribute and whether a consistent strategy was developed.

![Figure 5 Mean variance for each attribute judged as a function of assessor category (Table 1). See text explanation of labels on vertical axis. Attributes are represented by the shade of the circles and labelled in the legend. Each circle is labelled with a number that corresponds to the subject that made the judgments.](image)

### 4. CONCLUSION

The current study aimed at assessing the variability in perceptual evaluations of HRTFs using a listening test. The results showed that the variability across trials was significant for all subjects. Given the responses from subjective reports and an analysis of the repeatability of subjects’ judgments, the attributes *sense of direction* and *front image quality* are suggested as the most useful. These two attributes were not strongly correlated suggesting that they represent different aspects of the quality of a rendering in VAS. The effect of subject expertise on variability was also analysed and a correlation was found; the more experienced the listener, lower the variance in judgments across trials.
Therefore given the large variance in judgments for listeners with little expertise it would be difficult to determine an optimum HRTF in a listening test setting and it is recommended that only experts be used. Taken together this study offers a concise methodology for obtaining conclusive evaluations of HRTFs geared towards commercial applications.

5. ACKNOWLEDGEMENTS

We would like to acknowledge NASA for use of their ER-2 tubephones.

6. REFERENCES


[34] Romano R. et al., "Correcting for different use of the scale and the need for further analysis of individual differences in sensory analysis", Food Quality and Preference, 19, 197-209 (2008 Mar)


