Evaluation of spatial audio reproduction methods (part 2) : analysis of listener preference

Francombe, J, Brookes, T, Mason, R and Woodcock, J

10.17743/jaes.2016.0071

OPEN ACCESS Freely available online
Journal of the Audio Engineering Society
Vol. 65, No. 3, March 2017
DOI: https://doi.org/10.17743/jaes.2016.0071

Evaluation of Spatial Audio Reproduction Methods (Part 2): Analysis of Listener Preference

JON FRANCOMBE[1], TIM BROOKES,[1] *AES Member*, **RUSSELL MASON,[1]** *AES Member*, **AND**
(j.francombe@surrey.ac.uk)

JAMES WOODCOCK[2]

[1]*Institute of Sound Recording, University of Surrey, Guildford, UK*
[2]*Acoustics Research Centre, University of Salford, Salford, UK*

It is desirable to determine which of the many different spatial audio reproduction systems listeners prefer, and the perceptual attributes that are most important to listener experience, so that future systems can be perceptually optimized. A paired comparison preference rating experiment was performed alongside a free elicitation task for eight reproduction methods (consumer and professional systems with a wide range of expected quality) and seven program items (representative of potential broadcast material). The experiment was performed by groups of experienced and inexperienced listeners. Thurstone Case V modeling was used to produce preference scales. Both listener groups preferred systems with increased spatial content; nine- and five-channel systems were most preferred. The use of elicited attributes was analyzed alongside the preference ratings, resulting in an approximate hierarchy of attribute importance: three attributes (*amount of distortion*, *output quality*, and *bandwidth*) were found to be important for differentiating systems where there was a large preference difference; sixteen were always important (most notably *enveloping* and *horizontal width*); and seven were used alongside small preference differences.

0 INTRODUCTION

There is a wide range of spatial audio reproduction methods used for domestic or professional audio replay. Systems include mono and two-channel stereo, channel-based surround sound methods (5.1, 7.1, 9.1, 11.1, and 22.2 [1] are all used domestically or within the audio industry), "one box" solutions such as sound bars, and reproduction over headphones. With such a wide range of methods in use, it is important to discover what aspects of spatial audio reproduction particularly enhance the listener experience.

A previous study by the authors [2] identified the perceptual attributes that contribute to preference judgments made between alternative reproduction systems. The research described in the current paper aimed to determine which of these attributes are most important to listeners—that is, which of them have a strong relationship with listener preference. With this knowledge, these attributes can then be targeted in the design of new reproduction systems and the improvement of existing ones, and perceptual models of the important attributes can be developed and used to meter and optimize spatial audio reproduction. The problem was approached from two directions: determining listener preference for certain systems and ascertaining the perceptual characteristics of the differences between the systems. By collecting both quantitative and qualitative data, the relationship between the attributes and preference scores could be investigated. This facilitated analysis of the perceptual attributes that most contribute to creating a positive listener experience.

1 EXPERIMENT BACKGROUND

There have been previous attempts to understand the relationship between audio attributes and listener preference. Choisel and Wickelmaier [3] performed a spatial audio elicitation experiment and determined a set of eight attributes: *width*, *elevation*, *spaciousness*, *envelopment*, *distance*, *brightness*, *clarity*, and *naturalness*. They then collected ratings of these attributes, as well as listener preference, for eight reproduction methods (from mono to variants of 5.1 surround sound) in a paired comparison test [4]. They used multiple regression on a principal component reduction of the attribute data to determine the relationships between the attributes and listener preference, and showed that preference was correlated with two components: the first was related to spatial characteristics of the sound and the second to spectral characteristics. However, the principal component reduction reduced transparency in the model; that is, it is no longer easy to interpret the model

coefficients and the percepts to which they relate. As each feature in the model relates to a complex combination of multiple perceptual attributes that have been combined in the dimension-reduction stage, it is not possible to assess the coefficients and thereby determine the contribution of a single attribute to preference.

Zacharov and Koivuniemi [5] also elicited a set of attributes producing twelve scales: *sense of direction*, *sense of depth*, *sense of space*, *sense of movement*, *penetration*, *distance to events*, *broadness*, *naturalness*, *richness*, *hardness*, *emphasis*, and *tone color*. They performed a preference mapping experiment, collecting attribute ratings for eight reproduction systems (including mono, stereo, five-channel, eight-channel, and crosstalk-cancelled binaural systems) on scales with a resolution of 100 points using a single stimulus presentation. They analyzed the correlation between the scales and found that some attributes were significantly correlated. A partial least squares regression (PLS-R) was performed to map the attribute ratings to preference scores; a model with four components explaining 71% of the variance in the scores was found to be most suitable, with the following attributes contributing most significantly to the prediction: *sense of movement*, *sense of depth*, *direction* × *distance*, *broadness* × *distance*, *broadness* × *tone color*, and *depth* × *naturalness*. In this case, the PLS-R analysis combined more than one attribute to give an interaction term that is both relatively difficult to interpret and likely to be less generalizable (e.g., what does the combination of *depth* × *naturalness* mean and under what circumstances is it important?).

In the above studies, the use of statistical dimension-reduction techniques has been problematic. While the results may have provided some information on the relationship between attributes and preference, they are not sufficient to provide clear guidance for future researchers on what attribute should be investigated or optimized.

In the current study an alternative approach was taken that avoided the need to collect ratings on a large number of scales and also avoided the use of data reduction techniques such as PCA. An attribute elicitation experiment was performed alongside a preference rating task in an attempt to produce only those attributes that made a significant contribution to listener preference. Participants were asked to give preference ratings and, on the same screen, to type their reasons for making a particular judgment. The attribute elicitation aspects of this work are reported by Francombe et al. [2]. The attribute sets produced (by experienced and inexperienced listeners) were large (a total of 51 attributes—27 and 24 by the experienced and inexperienced listeners respectively, with some overlap between the sets). The preference ratings (which are analyzed in this paper) and elicited data were used in conjunction to determine particularly important attributes.

1.1 Research Questions

The research questions addressed in this study were as follows.

1. What spatial audio reproduction methods do listeners prefer?
2. Which attributes are most important to listener experience and should therefore be used in further evaluations?

It was also considered desirable to investigate the effect of listener experience on the answers to these questions. Descriptive analysis experiments often rely solely on experienced listeners for scale development and on inexperienced listeners for preference ratings [6, p. 346]; for example, Rumsey et al. [7] determined the relationship between the preference of inexperienced listeners and quality judgments made by experienced listeners. However, in this case it was considered beneficial to discover how inexperienced listeners perceive spatial audio stimuli, as such listeners account for the majority of domestic spatial audio consumption.

The methodology for the paired comparison preference rating experiment is described in Sec. 2. The preference results are detailed in Sec. 3, and the relationship between the elicited attributes and listener preference is discussed in Sec. 4. Finally, the outcomes from this work are discussed and concluded in Sec. 5.

2 PREFERENCE RATING EXPERIMENT METHODOLOGY

The methodology employed for the preference rating experiment is described in the following sections.

2.1 Stimuli and Participants

One limitation of previous studies is the restricted range of reproduction systems tested. It is not possible to test every conceivable reproduction method, especially as new systems are developed; however, it is desirable to future-proof experiment results to as great a degree as possible by including a wide range of systems. It is also necessary to select program material items that are representative of broadcast content and also able to reveal the differences between reproduction systems.

In this study eight reproduction methods were used: headphones, low-quality mono (a small computer loudspeaker), mono, stereo, 5-channel, 9-channel, 22-channel, and ambisonic cuboid. These methods were selected to cover a range of loudspeaker counts, positions, and expected quality. More details about the reproduction methods are given by Francombe et al. [2].

Seven program material items were used. The items were selected to be representative of a range of potential broadcast content and to meet a range of criteria including: various genres and musical elements, different numbers of sources, different types of source, different recording environments, different source positions and movement, and a variety of recording and production methods. The selected items were brass quintet, jazz quintet, pop track, big band, sport, experimental music, and film excerpt. Each excerpt was 20 seconds long. Again, more detail is given by Francombe et al. [2]. The excerpts were produced for each

reproduction system using a production technique that is typical for that system (including spatial microphone techniques, amplitude panning, ambisonic decoding, and binaural recording [8]). It is therefore not possible to completely separate the reproduction method from the production technique used; however, as a range of techniques were used for each reproduction system, the results averaged across stimuli retain some independence between the production technique and the reproduction method. Hence, the results can still be used to show the magnitudes of preference for different systems (as discussed in Sec. 3.4), and verbal elicitation alongside preference judgments allows analysis of why particular systems were preferred.

The experiment was performed by two groups of participants: seven experienced listeners and eight inexperienced listeners. The experienced listeners were fourth year undergraduate students on the Music and Sound Recording course at the University of Surrey, Guildford, UK, who had all completed a module in technical ear training and had critical listening experience in recording studios. The inexperienced listeners were current students or recent graduates in a range of disciplines. None of the inexperienced listeners had specific technical ear training, although they may have had a musical background and/or have participated in listening tests before.

2.2 Methodology

The listening test was run using a software user interface with multiple pages created in Max/MSP. Each experiment session was preceded by a familiarization stage, again using a bespoke Max/MSP interface. When participants switched between stimuli (in either interface) the position in the audio excerpt was maintained to enable easier comparisons. It should be noted that comparisons featuring the headphone reproduction meant that participants had to put on and remove the headphones when switching between reproduction methods, introducing a longer delay than for comparisons not involving the headphones.

All judgments for a single program item were made in one test session; therefore, a total of seven test sessions (each including a familiarization stage and the main test) were performed by each participant. The program items were presented in a different random order for each participant.

2.2.1 Familiarization Task

Before each test session participants were asked to listen to all of the reproduction methods for the program item under test so that they could learn and understand the full range of stimuli that they would hear in the session and therefore use the rating scale appropriately. If participants were not aware of the full range of reproduction methods before undergoing the test, it is possible that they might give a high preference rating for one stimulus over another, before auditioning a stimulus pair with an even greater difference, leaving no room on the scale to indicate this. The wording for the familiarization task given in the instructions was as follows.

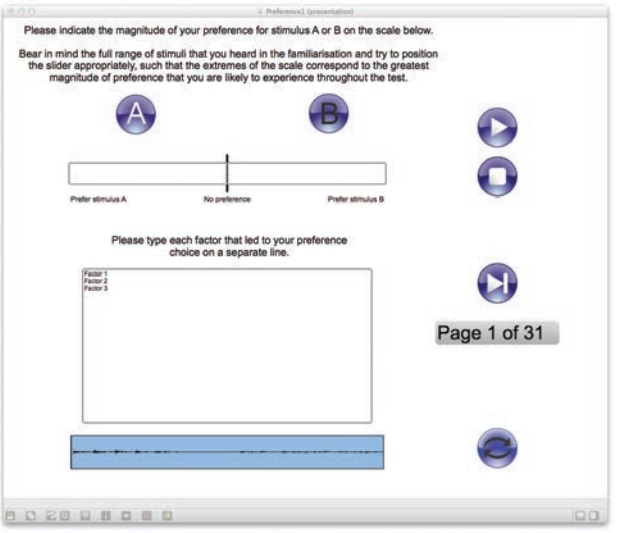


Fig. 1. Interface for the preference rating and free elicitation task.

"In the test stage, you will be asked to indicate the magnitude of your preference when comparing a pair of audio stimuli. There will be a total of 31 pairs in each session. Before the test, you will be asked to listen to all of the stimuli that you will hear during the test. Use this familiarization to get an idea of the range of stimuli, and of how much you prefer some to others, so that you can express the magnitude of your preference appropriately during the test. For example, in the test, within one pair you might have only a slight preference for one of the stimuli over the other; within another pair you might have a very large preference for one stimulus over the other. After listening to the stimuli on the familiarization page, you should be aware of the full range of magnitudes of preference that you are likely to experience during the test, and should therefore be able to rate these accordingly."

The user interface displayed eight buttons, and the reproduction methods were randomly assigned to these. Where participants were required to wear headphones, this was indicated by a headphone icon next to the appropriate button.

2.2.2 Main Test

The stimuli were presented to participants in a paired comparison paradigm with continuous ratings. The paired comparison methodology was selected as it allowed direct comparison between two stimuli, facilitating the differential elicitation aspect of the study. Paired comparisons were also considered to be easier for participants to make compared to, say, a multiple stimulus test. A continuous scale rating was used since, although this is more demanding to perform than a forced choice, it allows more fine-grained analysis of the difference between stimuli [9] and can be converted to binary judgments if necessary.

The preference ratings were made on a horizontal slider positioned below the stimulus replay buttons. The user interface is shown in Fig. 1. Participants were allowed to make a "no preference" judgment by positioning the slider in the middle of the scale; this position was clearly marked with a

vertical line. With eight reproduction methods there is a total of $\binom{8}{2} = 28$ comparisons. Three of these comparisons—selected at random for each of the program items but kept the same for all participants—were repeated to facilitate analysis of participant reliability; this resulted in a total of 31 judgments per program item for each participant. In each session, the stimulus combinations were presented in a random order and the two stimuli were assigned randomly to buttons A and B. A free elicitation was performed alongside the preference rating task: participants were asked to describe the reasons for giving a particular preference rating [2].

The wording of the preference rating task in the instructions given to participants was as follows.

"Each test session consists of 31 pages. On each page there is a pair of stimuli (labelled A and B). Please listen to the two stimuli and indicate which you prefer, and the magnitude of your preference, on the scale. Bear in mind the full range of stimuli that you heard in the familiarization and try to position the slider appropriately, where: far left indicates that you prefer stimulus A to stimulus B and that the magnitude of your preference is as great as you are likely to experience throughout the test session; far right indicates that you prefer stimulus B to stimulus A and that the magnitude of your preference is as great as you are likely to experience throughout the test session; immediately to the left or right of center indicates the smallest magnitude of preference, for stimulus A (left) or B (right); and center indicates no preference for either A (left) or B (right). Please note that you must move the slider at least once before you can continue to the next page. If you'd like to leave the slider in the middle of the scale, just move it away and then back."

The interface also showed the waveform of the audio; participants were able to select a short section of the excerpt to loop so that they could repeatedly listen to sections that highlighted differences that they felt to be important.

3 RESULTS

The data preprocessing, analysis of participant reliability, and analysis of preference results are presented in the following sections.

3.1 Data Preprocessing

To gain an overview of the range of the preference scale that participants were using, box plots of participant responses (pooled over all reproduction methods and program items) were plotted (Fig. 2). The box plots show the median, lower and upper quartiles, and range of the data, as well as any outliers. The notches can be used to determine significant differences in the median value; where notches do not overlap, the medians can be considered to be significantly different [10]. The extremes of the preference scale are indicated by −50 and 50, which indicate a preference for stimulus A and stimulus B respectively. In the experiment, the assignment of each stimulus in a pair to buttons A or B on the interface was randomized. Therefore, when

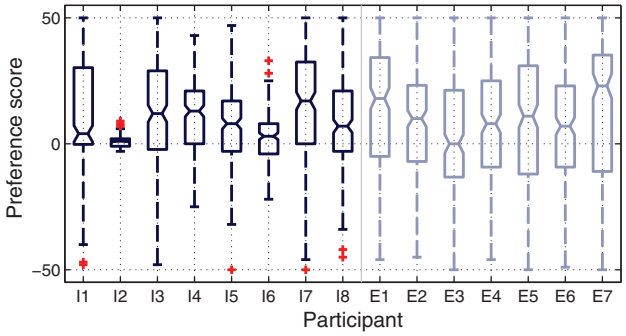


Fig. 2. Box plots of all responses by each participant. Participant prefix "E" indicates an experienced listener; "I" indicates an inexperienced listener.

generating this plot some data were inverted so that the A–B ordering for each pair was consistent across all tests. Thus, the slight positive offset in the scores seen in Fig. 2 does not indicate a bias towards stimulus B but, rather, a preference for the higher channel count systems, that tend (arbitrarily) to come second in the modified comparison order more often (79% of comparisons). It can be seen that the experienced listeners tended to use the full range of the scale, while the inexperienced listeners exhibit a high degree of inter-listener variance, with some participants (notably I2, I4, I5, and I6) using a reduced scale range. Consequently, the preference values were scaled for each participant by dividing each score by the standard deviation of the scores for that participant (a z-score transformation was not used as this includes mean-centering; in this case, centering the data was not appropriate as the center point of zero has a particular meaning, i.e., no preference). However, even with the application of the scaling, the very small range of results given by I2 is likely to mean that the results are noisy and unreliable (and in fact these results were discarded due to their unreliability—see Sec. 3.2.2).

3.2 Participant Reliability

3.2.1 Circular Error Percentage

One measure of the participants' ability to perform a paired comparison task is the circular error percentage. The responses for each possible triad of stimuli should exhibit transitivity; that is, if $A \rightarrow B$ and $B \rightarrow C$, then $A \rightarrow C$ should hold (where $\rightarrow$ is read as "is preferred to") [11]. The circular error percentage is the percentage of all possible triads of paired comparisons (given by $\binom{S}{3}$ where S is the number of stimuli) for which the ratings are intransitive. The mean circular error percentages across all program material items were 5.55% and 3.46% for the inexperienced and experienced listeners respectively. No individual participant or program material item exceeded a mean circular error percentage of 10%.

3.2.2 Repeat Judgments

Three repeat judgments were made for each program item (a total of 21 judgments per participant) so that an assessment of listener reliability could be made. Fig. 3 shows

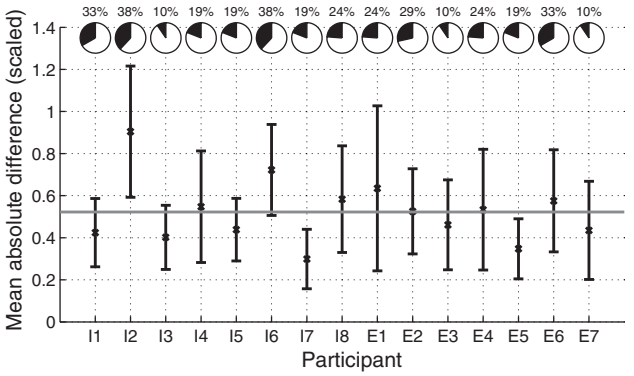


Fig. 3. Mean absolute difference between repeated judgments (scores scaled as described in Sec. 3.1). Error bars show 95% confidence intervals calculated using the *t*-distribution. The gray horizontal line shows the grand mean across all participants. The dark area of each pie chart shows the percentage of preference judgments that were changed between replicates. Participant prefix "E" indicates an experienced listener; "I" indicates an inexperienced listener.

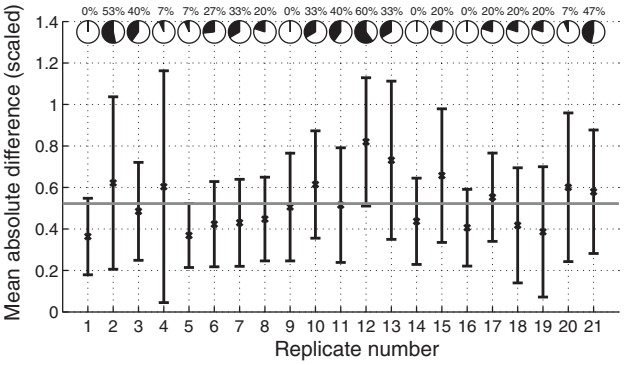


Fig. 4. Mean absolute difference between repeated judgments for each repeated stimulus (scores scaled as described in Sec. 3.1). Error bars show 95% confidence intervals calculated using the *t*-distribution. The gray horizontal line shows the grand mean across all replicates. The dark area of each pie chart shows the percentage of preference judgments that were changed between replicates.

the mean absolute difference between the first and second judgments, averaged across all 21 items for each participant. The horizontal line shows the mean across participants. One participant (I2) has a difference significantly greater than the mean; this is unsurprising given the limited range of responses mentioned above. Fig. 3 also shows the percentage of judgments in which a different preference was specified (i.e., one reproduction method was preferred in the first judgment and the other method preferred in the second judgment, or a "no preference" judgment was changed to a preference in either direction). Three inexperienced participants (including I2) have values above 25%. The mean percentage is slightly lower for the experienced listeners (21% and 25% for experienced and inexperienced listeners respectively); two experienced listeners gave reverse judgments in over 25% of cases.

As the replicates were selected at random, it is possible that those selected were particularly difficult (or easy) comparisons for which to make a preference judgment; for example, the program items selected may just have been very similar, making the judgment difficult. Unless every test item is repeated, this cannot be completely avoided; however, it was felt that this was a necessary trade-off to ensure that the experiment could be performed in a suitable time frame while allowing for analysis of listener reliability. To determine whether any of the selected replicates were found to be particularly difficult to perform, the mean absolute difference for each of the 21 replicates was plotted (Fig. 4), again alongside the percentage of judgments that were changed between the first and second replicate. The plot shows that there are no significant differences between the performance on each of the replicate stimuli. However, it is interesting to interpret these results alongside the percentage of changed judgments. When the mean absolute difference is low but the percentage of judgments changed is high (e.g., replicate 3: brass quintet, headphones versus 5-channel), this suggests that it was difficult to determine a preference for either stimulus. Therefore, small preference ratings were given in either direction and a small difference in rating could lead to a change in preference. In other cases, a relatively large mean absolute difference was accompanied by a high percentage of changed judgments (e.g., replicate 12: jazz quintet, stereo versus 5-channel), indicating that participants found it difficult to make judgments for this stimulus. Finally, a relatively large mean absolute difference accompanied by a small percentage of changed judgments (e.g., replicate 20: film excerpt, low-quality mono versus ambisonic cuboid) indicates that participants were confident of the method that was preferred but unsure of the magnitude of their preference judgments. The results on the whole indicate that the replicates cover a wide range of difficulty of judgment and therefore serve as a useful indicator of subject reliability.

Based on the evidence presented above (i.e., a relatively large mean absolute difference, high percentage of changed judgments, and small range of scale use), it was determined that the results from listener I2 were unreliable; they were therefore discarded for all further analysis of preference.

The repeat judgments were included in the experiment design solely to enable analysis of listener reliability; they were therefore removed from the data before any further analysis to maintain a balanced dataset. In every case, the first judgment was used.

3.3 Preference for Reproduction Methods

Thurstone Case V modeling [12] was used to convert the paired comparison data into a preference scale for the different reproduction methods, using the code presented by Tsukida and Gupta [13]. The model assumes that the relative magnitudes of preferences for the stimuli can be determined from the percentage of times that one stimulus is selected over another in a paired comparison task.

In order to generate confidence intervals around the Thurstone scale values, the data were bootstrapped; that is, a random sample of the data was drawn multiple times and the scale calculated for each subset, with the final values taken

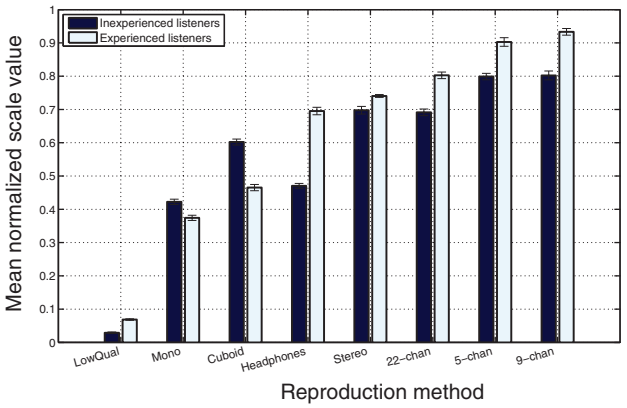
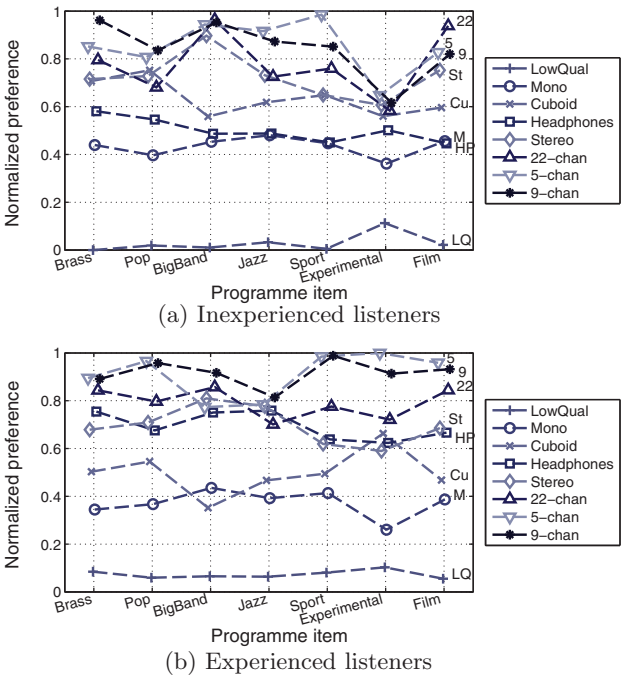


Fig. 5. Preference scale created using Thurstone Case V model with raw scores for experienced and inexperienced listeners. Confidence intervals show bootstrapped 95% confidence intervals calculated over 50 iterations.



(a) Inexperienced listeners

(b) Experienced listeners

Fig. 6. Preference scale, broken down by stimuli, created using Thurstone Case V model with raw scores, for experienced and inexperienced listeners.

as the mean of the scale points across all of the samples, and confidence intervals calculated using the normal distribution. In this case, approximately 50% of the data were used in each iteration (25 out of 49 judgments for each reproduction method), and the procedure was repeated for 50 random samples. For comparisons in which one stimulus was always preferred over the other, an offset the equivalent size of a no preference judgment was added to the stimulus that was never preferred, and taken off the stimulus that was always preferred, in order to allow computation of the scale (the inverse normal distribution function is used to determine scale values, returning an infinite value for a probability of 1 or 0; the correction avoided this problem). The Thurstone scale values were normalized to the range 0–1 (over all iterations and both subject types) before taking the mean across bootstrap iterations. The resulting scales for the experienced and inexperienced listeners are shown in Fig. 5, ordered according to the experienced listeners' preference.

The confidence intervals are small, suggesting that the scale values are reliable and that the reproduction methods can clearly be differentiated; the only exception to this are small differences in the inexperienced listeners' results between the 5-channel and 9-channel methods and the 22-channel and stereo methods. The results show that the low-quality mono reproduction method is clearly the least preferred by both sets of listeners, followed by the monophonic reproduction. This suggests that both types of listener prefer the presence of increased spatial information. However, it is interesting to note that preference does not increase monotonically with channel count. The experienced listeners display only a marginal preference for the cuboid reproduction over mono, while the inexperienced listeners display only a small preference for headphones. For both groups of listeners, the 22-channel reproduction was preferred less than the 5- and 9-channel methods. The inexperienced listeners distinguished less between the different surround sound methods, although the pattern of preference is similar to that of the experienced listeners.

3.3.1 Preference by Program Item

In order to determine the effect of program item on listener preference for the reproduction methods, the Thurstone scaling was repeated with the data broken down by program item. Due to the reduced number of cases, the bootstrapping procedure was not performed; however, the small confidence intervals seen in the results presented above suggest that the preference scaling is consistent.

Fig. 6 shows the preference scale values for each reproduction method (denoted by line shade and marker shape) for each program material item. Lines that intersect indicate different rank orders of reproduction methods depending on the program item. There is some variance in preference for the different program items; however, the reproduction methods themselves have a greater effect on the results. This is confirmed by analyzing the rank correlation using Kendall's coefficient of concordance (W), which measures the agreement between multiple sets of ranks [14]. In this case, $W = 0.91$ and $W = 0.90$ for the inexperienced and experienced listeners respectively, indicating high agreement between the rank orders for each program item.

For the inexperienced listeners, notable differences include: a rise in preference for the cuboid reproduction of the pop and brass program items; an increase in preference for the 22-channel reproduction of the big band program item; and a general reduction in the range of preference scores for the experimental music program (including a large drop in preference for the 5-channel, 9-channel, and 22-channel reproduction methods, and an increase in preference for the low-quality mono reproduction). The range reduction for the experimental content could be attributed

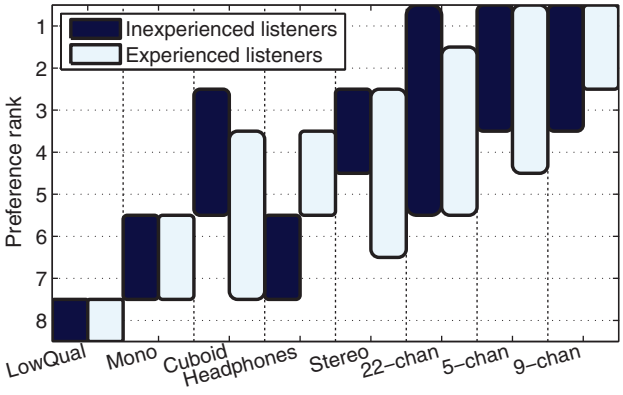


Fig. 7. Range of preference ranks for each reproduction method (experienced and inexperienced listeners).

to a lack of familiarity with the type of program material resulting in generally lower preference.

For the experienced listeners, the cuboid exhibited a relatively low score for the big band reproduction and a high score for the experimental music (which was specifically designed for ambisonic replay). The 5-channel reproduction of the big band program item received a relatively low preference rating. The mono reproduction of the experimental music received a low preference score, indicating that spatial content was important to the experienced listeners for this program item—this could be influenced by the pronounced movement within the sound sources in this item.

Some reproduction methods exhibited more variable preference ratings than others with different program material. The range of rank positions that each reproduction method achieved is displayed in Fig. 7. For both groups of listeners, the low-quality mono reproduction was always ranked eighth out of eight. For the inexperienced listeners, the 22-channel method shows the highest variance, ranking between fifth and first depending on the program material. For the experienced listeners, the 5-channel, 22-channel, stereo, and cuboid methods all have a high variance (four different rank positions over the seven program items). The 9-channel method consistently received high preference scores—it was always ranked in the top three for both listener groups (and first or second for the experienced listeners).

3.4 Discussion

The results presented above show a clear benefit of audio systems that provide spatial information, with stereo preferred to mono, and surround sound preferred to stereo. However, the trends in preference for the surround sound systems are less clear. The results show little difference between the 5-channel and 9-channel reproduction methods, suggesting that height content is not always beneficial, or results in little increase in preference—although there are specific program items where the preference score is increased with the addition of height channels. The 22-channel system has the highest loudspeaker count but for most program items scores lower than the 5- and 9-channel methods.

The experienced and inexperienced listeners tended to be in reasonably strong agreement, although the experienced listeners could discriminate between the most preferred systems slightly more. The main differences were in the headphone reproduction (preferred considerably by the experienced listeners) and the ambisonic reproduction (preferred by the inexperienced listeners).

It should be noted that the production of the program items was inextricably linked to the reproduction methods in this experiment (albeit using representative production techniques); an argument could be made that 22-channel was less preferred as the content was less suitable. There is generally less experience in producing 22-channel content, which may mean that the full capabilities of this system have not been achieved. Therefore, it is not possible to conclude that there will never be a benefit of 22-channel reproduction; however, based on these results, any preference shown over five or nine loudspeakers appears to be relatively small. This finding is supported by the results presented by Silzle et al. [15], who found only small differences in ratings of quality for 5-channel, 9-channel, and 22-channel reproductions in an experiment with no reference. Where a 22-channel reference was included, the differences between reproduction methods were more pronounced, with higher quality ratings for the higher channel count reproductions. Further analysis in this paper will focus on the reasons that listeners gave for making preference judgments, enabling the results to be used to determine the perceptual advantages and disadvantages of the particular methods.

4 DETERMINING THE IMPORTANT ATTRIBUTES

The second research question presented in Sec. 1.1 addressed the need for determining the attributes that are most important to listener experience. It is desirable to know which attributes are important so that they can be used for evaluation in listening tests, in meters for use by content producers, and in algorithms for the optimization of spatial audio systems.

One way to determine which of the attributes are of particular importance would be to collect and statistically analyze ratings of stimuli on every attribute scale. Such analyses might include determination of the correlation between attributes, assessment of listener agreement on each attribute, or observation of the relationship between the attributes and preference. However, due to redundancy in the attribute set, it is inefficient to rate every attribute; even when attempting to elicit unique attributes, the produced terms often still overlap. The results of dimension-reduction analysis can also be difficult to interpret (as discussed in Sec. 1). For example, multiple attributes or attribute interactions may be represented on a single dimension.

It is therefore desirable to select a subset of attributes that are of particular importance, for which more data can subsequently be collected. In an example of this type of redundancy reduction, Francombe et al. [16] performed an

experiment in which participants were requested to choose the most relevant attribute for each of a set of stimuli. Frequency of attribute use was employed to select four attributes from the original set of twelve; the selection was shown to be suitable in the subsequent rating experiment, as a principal component analysis showed that two dimensions accounted for 90% of the variance in the ratings, and two of the four attributes were highly correlated.

In this case, the collection of preference ratings alongside free elicitation responses facilitates analysis of the attribute use, allowing selection of a subset of important attributes. An attribute can be considered important [6, pp. 343–346] if:

- It is used and understood by all participants; and
- It allows differentiation between the stimuli under test.

The analyses described below were performed to determine which of the attributes meet these criteria and should therefore be investigated further. First, a mapping between the free elicitation data and resultant attributes was performed so that attribute use could be considered with relation to particular stimuli (Sec. 4.1). This was followed by analysis of the attributes against the two criteria. The percentage of participants that used each attribute is analyzed in Sec. 4.2. The discriminatory power of the attributes is assessed in Sec. 4.3 by looking at the frequency of use of the attributes alongside different preference judgments. Definitions of the attributes used in this study are given by Francombe et al. [2]

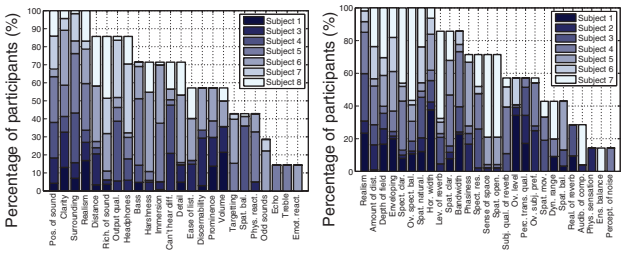
4.1 Free Elicitation Data to Attribute Mapping

The attributes considered below were determined by a free elicitation, an automatic text clustering process to reduce redundancy, and then group discussions in which the algorithmically-generated clusters were put into sets, which were subsequently labelled and defined [2].

To facilitate analysis of the attributes, each response from the free elicitation was mapped to the attribute to which it ultimately contributed by: (i) identifying the cluster in which the response was included (by the automatic clustering algorithm); and (ii) identifying the attribute to which that cluster ultimately contributed after the cluster-setting and attribute labelling procedure.

However, this process can fail to map some responses for the following reasons.

1. An incorrectly-clustered term may have remained in a cluster that eventually contributed to an attribute, resulting in an incorrect mapping between response and attribute.
2. Some of the automatically produced clusters were discarded. Subsequent listening to audio recordings of the elicitation group discussions identified two main reasons for discarding clusters: (i) the responses were not useful, i.e., they were meaningless or ambiguous; or (ii) items in a cluster were relevant to two or more attributes that already had clearly defined groups. For the experienced listeners, 58 out of 228 clusters were discarded (25.4%); the remaining 170 clusters accounted for 77.8% of the responses from the free elicitation experiment (1531 from 1967). For the inexperienced listeners, 95 out of 244 clusters were discarded (38.9%); the remaining 149 clusters accounted for 56.4% of the responses from the free elicitation experiment (1281 from 2270). It should be noted that these figures include the results given by inexperienced participant 2 (which were discarded from the preference analysis in this paper), but do not include identical responses given in the free elicitation (which were removed before the automatic clustering).



(a) Inexperienced listeners (b) Experienced listeners

Fig. 8. Percentage of participants that used terms contributing to each attribute. Each bar is proportioned according to the number of times each participant used a term contributing to the corresponding attribute.

Despite the discarding of some clusters, analysis of the remaining data can still highlight important attributes. There is a large proportion of the data available, and removal of meaningless responses, i.e., those that participants agreed did not describe the listening experience in a valuable manner, is not damaging—in fact, it is potentially beneficial to showing trends. Furthermore, removal of responses that would have fallen into categories that already exist would not change the essential nature of relationships (although it could reduce the strength of trends). Finally, allowances can be made in statistical analysis (as described below) to compensate for the missing data, and the attributes selected at this stage can be validated in future experiments.

4.2 Which Attributes Are Used by All Participants?

Lawless and Heymann [6] state that attributes should offer consensus on meaning and be unambiguous so that they are understood by all participants (the first of the importance criteria presented in Sec. 4). Using group discussions to produce the terms helps with this as participants are required to agree on a definition. However, statistical checks are also useful. Fig. 8 shows, for each attribute, the percentage of participants that gave a response in the free elicitation experiment that eventually contributed to that attribute. Each bar is proportioned according to how many of the responses by each participant contributed to the corresponding attribute, to assess whether a term was mainly produced by one participant.

All of the inexperienced listeners used terms that contributed to the following attributes: *position of sound*, *clarity*, *surrounding*, and *realism*. The proportion of responses from the different participants was split reasonably evenly. For the experienced listeners, all of the participants gave responses that contributed to the following attributes: *realism*, *amount of distortion*, *depth of field*, *enveloping*, *spectral clarity*, *overall spectral balance*, *spatial naturalness*, and *horizontal width*. The responses for *spectral clarity* and *overall spectral balance* are somewhat dominated by experienced participant 7.

4.3 Which Attributes Allow Differentiation between the Stimuli under Test?

Lawless and Heymann [6] state that attributes should be able to be used to discriminate between stimuli (the second of the importance criteria presented in Sec. 4). With the data available, it is not possible to say whether an attribute is directly responsible for a particular preference rating; however, it is possible to observe the mean preference ratings where particular attributes were used, in order to gain an understanding of whether attributes were generally used when there was a small or a large preference for one stimulus or the other. Attributes that were generally used when there was a large difference in preference between two reproduction methods are likely to be important as they highlight the percepts that engender a greatly improved listening experience; however, attributes that are consistently associated with a small preference difference may also be important, as they can be used to distinguish between similar systems.

A pre-selection was made by removing attributes that were used very infrequently (with the exception of those that were used by a high proportion of the participants), as these can also be considered unlikely to be particularly important compared with the more frequently-used attributes. This is discussed in Sec. 4.3.1. Following this, frequency of attribute use was plotted against mean absolute preference for each reproduction method combination in order to analyze the relationship between attribute use and preference. This is discussed in Sec. 4.3.2.

4.3.1 Frequency of Attribute Use

A chi-square test was used to assess the distribution of the frequency of attribute use [17, pp. 186–187]; the null hypothesis for the test was that the usage frequency was uniformly distributed. Rejection of the null hypothesis indicates that some attributes were used significantly more or less than others. Standardized residuals can be used to determine whether individual categories are used at greater or less than chance frequency [18, pp. 698–699]. The standardized residual for the ith attribute, R_{s_i}, is given by

$$R_{s_i} = \frac{(x_i - E)}{\sqrt{E}}, \tag{1}$$

where x is the frequency of use and E is the expected count (which, assuming a uniform distribution, is equal for each attribute and given by the number of observations divided by the number of attributes). If R_{s_i} lies outside of the range of the normal distribution for a given probability level (in this case, $\alpha = 0.05$, giving a range of ± 1.96 for a two-tailed test), the difference is considered to be significant.

The expected count E was modified to maintain the assumption of a uniform distribution if all of the discarded values had been present; E_{mod} was given by the number of observations plus the number of discarded observations, divided by the number of attributes. Eq. (1) was then rearranged to determine the frequency of attribute use (A) required to indicate a significant difference:

$$A_{\text{sig}} = \left(\pm 1.96 \cdot \sqrt{E_{\text{mod}}}\right) + E_{\text{mod}}. \tag{2}$$

A_{sig} gives a value for which an attribute is used at significantly greater or less than chance frequency, even if none of the discarded values would have been applied to that attribute were they present. However, it could also be the case that some of the discarded values would have been given to a certain attribute. One-sided confidence intervals were generated by proportionally distributing the discarded values to the attributes. If T is the total frequency of use (given by $T = \sum_{i=1}^{I} x_i$), and D is the number of discarded values, then the size of the confidence interval C_i is given by

$$C_i = \left[\frac{x_i}{T}\right] \cdot D. \tag{3}$$

If the attribute use frequency plus the confidence interval falls above A_{sig} (i.e., $x_i + C_i > A_{\text{sig}}$), it is likely that the attribute was used at greater than chance frequency; likewise, if the attribute use frequency plus the confidence interval falls below $-A_{\text{sig}}$ (i.e., $x_i + C_i < -A_{\text{sig}}$), it is likely that the attribute was used significantly infrequently.

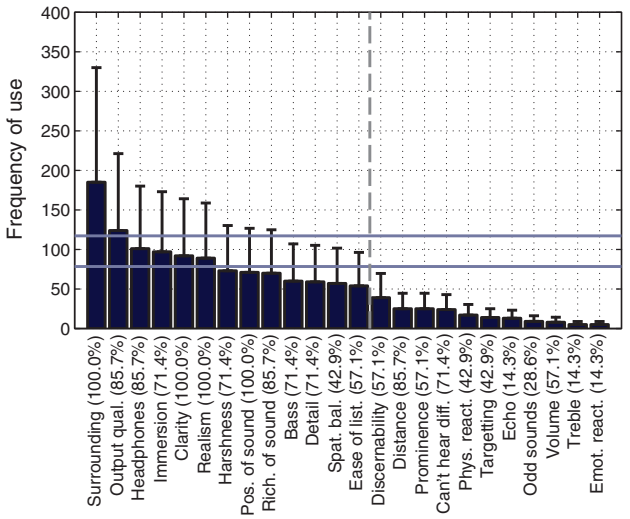
The results of this procedure are shown in Fig. 9, which shows the frequency of use and confidence intervals calculated as described above for each attribute and horizontal lines at the two values of A_{sig}. For the inexperienced and experienced listeners respectively, 11 and 15 attributes were used with less than chance frequency. These are the attributes to the right of the dashed line in each plot in Fig. 9.

For the following analysis, the attributes that were used with less than chance frequency are not shown, with the exception of those that were used by over 70% of participants (*distance*, *level of reverb*, *phasiness*, and *spectral resonances*). *Can't hear difference* was excluded as this was felt to be irrelevant.

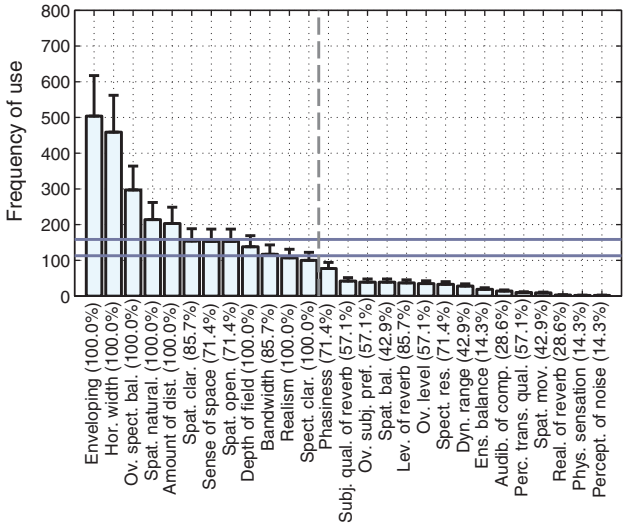
4.3.2 Attribute–Preference Relationship

Fig. 10 shows frequency of attribute use against mean absolute preference for each reproduction method combination, for those attributes that were used by over 70% of participants and/or not used at less than chance frequency (after accounting for discarded values). The points are differentiated by the two reproduction methods used in each comparison. The plots are ordered by overall frequency of attribute use.

The inexperienced listener plots show that the most frequently used attributes were mainly used when a particular



(a) Inexperienced listeners



(b) Experienced listeners

Fig. 9. Overall frequency of attribute use for each listener group. Upper and lower horizontal lines show the frequency required for an attribute to be used significantly more or less than chance frequency respectively. Error bars show a proportional distribution of discarded values (see text for full explanation). Dashed vertical lines show the cutoff point of attributes that were used at less than chance frequency after accounting for discarded values. Labels include the percentage of participants that used each attribute.

reproduction method was included in a comparison: *surrounding* with the mono reproduction, *output quality* with the low-quality mono (which also has the largest preference differences), and *headphones* for the headphone reproduction. The other attributes do not show pronounced relationships between frequency of use and preference, tending to be used with approximately equal frequency over the range of preference scores (although *spatial balance* was used mainly for small preference scores, and *bass* shows an increase in frequency for small preference scores). However, some individual comparisons stand out: *position of sound* was used most frequently for the mono–cuboid comparison; *richness of sound* was used most frequently for the mono–stereo comparison; *bass* was used most frequently for the 22-channel–cuboid comparison; and *spatial balance* was generally used more for comparisons involving the cuboid.

The experienced listener plots show stronger trends. Two attributes were used almost exclusively alongside preference ratings of a high magnitude (i.e., there is a cluster of points in the top-right corner of the plot), when the low-quality mono speaker was involved in a comparison: *amount of distortion* and *bandwidth*. Other terms were used more frequently when there was a small difference in preference and less so when there was a pronounced difference (most points fall to the left-hand side of the plot): *spatial naturalness* (particularly for the cuboid reproduction), *depth of field*, *level of reverb*, and *spectral resonances*. The remaining attributes were used regardless of the magnitude of preference scores. However, *enveloping* and *horizontal width* show a pronounced trend, being used more frequently as the preference scores increase. *Spatial naturalness*, *spatial clarity*, and *phasiness* were used more frequently in comparisons involving the cuboid reproduction method.

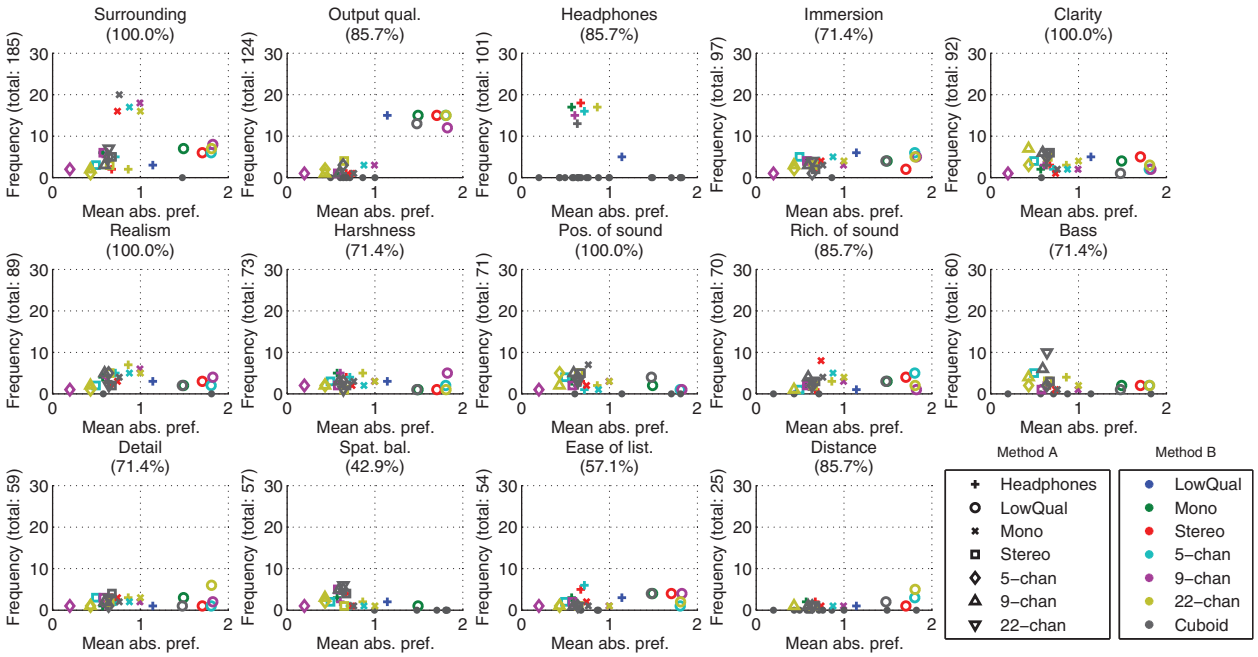
4.4 Important Attributes Summary

Table 1 gives a summary of the attributes analyzed above, i.e., those that were used at chance frequency or greater as well as those used at less than chance frequency but mentioned by over 70% of participants. The attributes are sorted by the magnitude of preference judgments for which they were used, then their frequency of use, then the number of participants who gave a response (in the free elicitation stage) that contributed to the attribute. Where an attribute was elicited by both groups of listeners, the inexperienced listeners' attribute has been indicated in parentheses; Francombe et al. [16] found that the experienced listener labels and definitions were preferred by both groups of listeners in cases where the attribute sets overlapped.

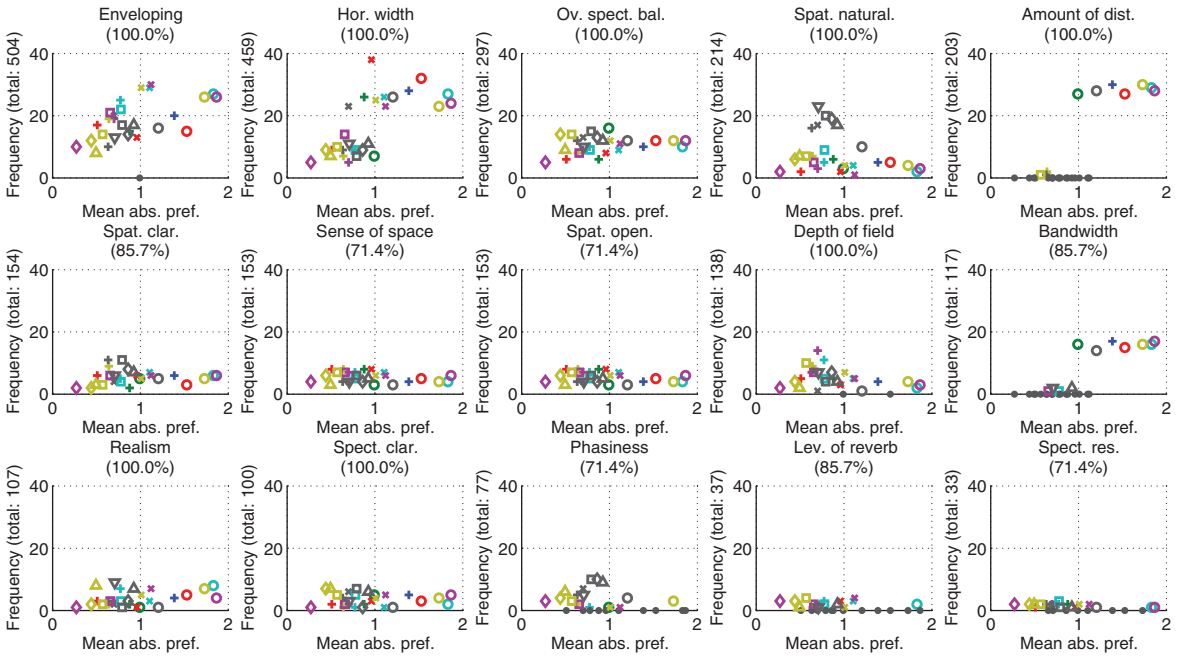
5 CONCLUSIONS AND DISCUSSION

A literature review into the attributes that are important when making preference judgments between spatial audio reproduction methods identified a number of limitations of the existing research. Previous studies have been designed to identify all of the differences between listening experiences, rather than just those that make a meaningful contribution to listener preference. Similar percepts are often labelled differently across studies, or different percepts are given similar labels. There is a lack of studies in which comparisons between different types of reproduction method are drawn, and recent developments in channel-based surround sound with height have not been investigated in detail. The relationship between individual attributes and preference has not been made clear. Finally, inexperienced listeners have not been used in attribute elicitation studies.

This study aimed to address these limitations and answer the following questions: (i) what spatial audio reproduction methods do listeners prefer; and (ii) which attributes are most important to listener experience and should therefore



(a) Inexperienced listeners



(b) Experienced listeners

Fig. 10. Frequency of attribute use against mean absolute preference for each reproduction method combination. The points are differentiated by the two reproduction methods used in each combination (shape and color). A light gray dot indicates that an attribute was not used for a particular stimulus. The percentage in the title is the percentage of participants that used the attribute.

be used in further evaluations? The conclusions to these questions are discussed below.

5.1 What Spatial Audio Reproduction Methods Do Listeners Prefer?

The Thurstone Case V model was used to produce preference scales from paired comparison preference ratings for eight reproduction methods and seven program items. The results suggested that providing two or three dimensions of stable spatial information is desirable; the 5- and 9-channel reproductions were preferred overall, while the 22-channel method was preferred by the inexperienced listeners for two program items (big band and film excerpt). The results were broadly similar between experienced and inexperienced listener groups, although experienced listeners

Table 1. Summary of important attributes. The letters "E" and "I" indicate that an attribute was produced by the experienced and inexperienced listeners respectively. Pref. indicates the approximate magnitude of the mean absolute preference scores most often associated with each attribute. Freq. is the overall frequency of use, and "used by" indicates the percentage of participants that gave a response that contributed to each attribute.

Attribute	Listener type	Pref.	Freq.	Used by (%)
Amount of distortion	E		203	100
Output quality	I	Large	124	86
Bandwidth	E		117	86
Enveloping (Immersion)	E (I)		504	100
Horizontal width	E		459	100
Overall spectral balance	E		297	100
Spatial naturalness	E		214	100
Surrounding	I		185	100
Spatial clarity	E		154	86
Sense of space	E		153	71
Spatial openness	E	All	153	71
Realism (Realism)	E (I)		107	100
Spectral clarity	E		100	100
Clarity	I		92	100
Harshness	I		73	71
Position of sound	I		71	100
Richness of sound	I		70	86
Detail	I		59	71
Ease of listening	I		54	57
Depth of field (Distance)	E (I)		138	100
Headphones	I		101	86
Phasiness	E		77	71
Bass	I	Small	60	71
Spatial balance	I		57	43
Level of reverb	E		37	86
Spectral resonances	E		33	71

preferred headphone reproduction more than inexperienced listeners, and inexperienced listeners showed a larger preference for the cuboid reproduction. Inexperienced listeners tended to discriminate less between the methods, with little difference between preference for mono and headphones, or between stereo and 22-channel.

It should be noted that this study cannot conclusively state that there may never be a benefit of having more than nine loudspeakers; the systems under test were inextricability linked with the program material items and the methods used to create them. It is likely that content creators are more practiced at producing program material for lower channel count systems—particularly stereo—and that, as expertise with other systems grows, there may be further benefits available.

5.2 Which Attributes Are Most Important to Listener Experience and Should therefore Be Used in Further Evaluations?

Important attributes were selected by observing which of the attributes were used by the majority of participants and were not used significantly infrequently, as well as by analyzing the relationship between the attribute use and preference scores.

The attributes that were used at chance frequency or greater, as well as those that were used at less than chance frequency but mentioned by over 70% of participants, are summarized in Table 1. The results suggest an approximate hierarchy of attributes. There is a small group of attributes that are consistently associated with large differences in preference between systems; in many regards, these could be considered to be the most important attributes and therefore should be considered first in evaluation of spatial audio systems. These attributes are *amount of distortion* and *bandwidth* from the experienced listeners and *output quality* from the inexperienced listeners (although this could be considered to be an umbrella term that includes both aspects identified by the experienced listeners). However, the pronounced differences in quality associated with these attributes are generally produced by the low-quality mono reproduction method; for distinguishing between higher quality methods, different attributes may be more relevant. Other attributes are important regardless of the magnitude of preference; of these attributes, *enveloping* and *horizontal width* were used very frequently and showed a strong relationship with preference scores, with an increase in frequency of use as preference increased. Further attributes are used mainly alongside small preference ratings. Such attributes are likely to be less related to overall preference, but could be used to differentiate between high-quality

systems. These attributes are *depth of field*, *phasiness*, *level of reverb* and *spectral resonances* for the experienced listeners and *headphones*, *bass*, and *spatial balance* by the inexperienced listeners.

Some attributes are most frequently used for particular reproduction methods. The experienced listeners mainly used *spatial naturalness* and *phasiness* when describing comparisons involving the ambisonic cuboid—this is unsurprising given the phase manipulation and small sweet spot involved in first-order ambisonic reproduction. The inexperienced listeners used *surrounding* most frequently in comparisons including the stereo reproduction method, *output quality* in comparisons involving the low-quality mono method, *headphones* in comparisons involving the headphones, and *spatial balance* in comparisons including the cuboid.

5.3 Summary

The research presented in this paper was designed to determine which reproduction methods listeners prefer and the important attributes that contribute to those preference judgments. It was shown that, broadly, the presence of more spatial content leads to an increase in preference, but that simply adding loudspeaker channels does not necessarily give a corresponding rise in preference. Consequently, the perceptual attributes that contribute to listener preference are of particular interest. Development of perceptually optimal spatial audio reproduction methods should focus on improving performance on the key attributes.

The results presented above (see Table 1) provide guidance to researchers who would like to select attributes for perceptual evaluation of spatial audio reproduction systems. Different attributes might be used depending on the nature of the analysis. For example, a coarse evaluation of low-quality or consumer devices might focus on the attributes that had a large effect on listener preference (*amount of distortion*, *bandwidth*, *output quality*, and also potentially *envelopment* and *horizontal width*, which were commonly used alongside large differences in preference). Evaluation of high-quality surround sound systems with small differences might focus on more subtle differences (e.g., *depth of field*, *phasiness*, *bass*, *spatial balance*, *level of reverb*, and *spectral resonances*). Finally, where a particular reproduction method or facet of the listening experience is under investigation, there might be certain attributes that are of particular interest (e.g., *spatial naturalness* and *phasiness* for evaluation of systems including ambisonics).

5.3.1 Future Work

The analysis presented above highlighted some attributes that are important because they contribute significantly to listener preference for spatial audio reproduction methods. The natural progression of this work would be to collect attribute ratings alongside preference ratings for a set of stimuli, which would enable quantitative analysis of each attribute's contribution to listener preference, as well as validation that the correct attributes were determined to be important.

Determination of the physical parameters that affect ratings of each of the important attributes then leads naturally to the development of predictive models. Such models can be used for quick evaluation of existing systems, metering in a spatial audio content production workflow, and development of new, perceptually optimized spatial audio reproduction methods.

6 ACKNOWLEDGMENTS

This work was supported by the EPSRC program Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership. Details about the data underlying this work, along with the terms for data access, are available from https://doi.org/10.15126/surreydata.00809533.

7 REFERENCES

[1] ITU-R rec. BS.2051, "Advanced Sound System for Programme Production," Tech. rep., ITU-R Broadcasing Service (Sound) Series (2014).

[2] J. Francombe, T. Brookes, and R. Mason, "Evaluation of Spatial Audio Reproduction Methods (Part 1): Elicitation of Perceptual Differences," *J. Audio. Eng. Soc.*, vol. 65, pp. 198–211 (2017 Jan./Feb.).

[3] S. Choisel and F. Wickelmaier, "Extraction of Auditory Features and Elicitation of Attributes for the Assessment of Multichannel Reproduced Sound," *J. Audio Eng. Soc.*, vol. 54, pp. 815–826 (2006 Sep.).

[4] S. Choisel and F. Wickelmaier, "Evaluation of Multichannel Reproduced Sound: Scaling Auditory Attributes Underlying Listener Preference," *J. Acoust. Soc. Am.*, vol. 121, pp. 388–400 (2007), https://doi.org/10.1121/1.2385043.

[5] N. Zacharov and K. Koivuniemi, "Audio Descriptive Analysis & Mapping of Spatial Sound Displays," *Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, 29 July–1 August* (2001).

[6] H. T. Lawless and H. Heymann, *Sensory Evaluation of Food: Principles and Practices* (Springer, New York, 1999), pp. 343–346.

[7] F. Rumsey, S. Zieliński, R. Kassier, and S. Bech, "On the Relative Importance of Spatial and Timbral Fidelities in Judgments of Degraded Multichannel Audio Quality," *J. Acoust. Soc. Am.*, vol. 118, pp. 968–976 (2005), https://doi.org/10.1121/1.1945368.

[8] J. Francombe, T. Brookes, R. Mason, R. Flindt, P. Coleman, Q. Liu, and P. Jackson, "Production and Reproduction of Programme Material for a Variety of Spatial Audio Formats," presented at the *138th Convention of the Audio Engineering Society* (2015 May), eBrief 199.

[9] E. Parizet, N. Hamzaoui, and G. Sabatie, "Comparison of Some Listening Test Methods: A Case Study," *Acta Acustica united with Acustica*, vol. 91, pp. 356–364 (2005).

[10] R. McGill, J. W. Turkey, and W. A. Larsen, "Variations of Box Plots," *Amer. Statistician*, vol. 32, pp. 12–16 (1978), https://doi.org/10.2307/2683468.

[11] E. Parizet, "Paired Comparison Listening Tests and Circular Error Rates," *Acta Acustica united with Acustica*, vol. 88, pp. 594–598 (2002).

[12] L. Thurstone, "A Law of Comparative Judgment," *Psych. Rev.*, vol. 34, pp. 273–286 (1927), https://doi.org/10.1037/h0070288.

[13] K. Tsukida, M. Gupta, "How to Analyze Paired Comparison Data," Tech. Rep. UWEETR-2011-0004, University of Washington (2011).

[14] M. G. Kendall and B. Babington Smith, "The Problem of *m* Rankings," *Annals of Mathematical Statistics*, vol. 10, pp. 275–287 (1939).

[15] A. Silzle, S. George, E. A. P. Habets, and T. Bachmann, "Investigation on the Quality of 3D Sound Reproduction," *Proceedings of the First International Conference on Spatial Audio, Detmold, Germany, 10–13 November* (2011), pp. 334–341.

[16] J. Francombe, R. Mason, M. Dewhurst, and S. Bech, "Elicitation of Attributes for the Evaluation of Audio-on-Audio Interference," *J. Acoust. Soc. Am.*, vol. 136, pp. 2630–2641 (2014).

[17] S. Bech and N. Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application* (Wiley, Chichester, 2006), pp. 186–187.

[18] A. P. Field, *Discovering Statistics Using SPSS* (Sage Publications, London, 2005), pp. 698–699.

THE AUTHORS


Jon Francombe


Tim Brookes


Russell Mason


James Woodcock

Jon Francombe graduated with a first-class honors degree in music and sound recording (Tonmeister) from the University of Surrey, Guildford, UK, in 2010, and received a Ph.D. in perceptual audio quality evaluation from the same institution in 2014. His Ph.D. research investigated the experience of a listener in an audio-on-audio interference situation. He is currently working as a research fellow on the EPSRC-funded "S3A: Future Spatial Audio" project investigating the perceptual attributes of spatial audio reproduction. He has also worked as a music technician, freelance musician, and sound engineer.

•

Tim Brookes received the B.Sc. degree in mathematics and the M.Sc. and D.Phil. degrees in music technology from the University of York, York, UK, in 1990, 1992, and 1997, respectively. He was employed as a software engineer, recording engineer, and research associate before joining, in 1997, the academic staff at the Institute of Sound Recording, University of Surrey, Guildford, UK, where he is now senior lecturer in audio and director of research. His teaching focuses on acoustics and psychoacoustics and his research is in psychoacoustic engineering: measuring, modeling, and exploiting the relationships between the physical characteristics of sound and its perception by human listeners.

•

Russell Mason graduated from the University of Surrey in 1998 with a B.Mus. in music and sound recording (Tonmeister). He was awarded a Ph.D. in audio engineering and psychoacoustics from the University of Surrey in 2002 and was subsequently employed as a research fellow. He is currently a senior lecturer in the Institute of Sound Recording, University of Surrey, and is program director of the undergraduate Tonmeister program. Russell's research interests are focused on psychoacoustic engineering including the development of methods for subjective evaluation and modelling aspects of auditory perception.

•

James Woodcock is a research fellow at the University of Salford. His primary area of research is the perception and cognition of complex sound and vibration. James holds a B.Sc. in audio technology, an M.Sc. by research in product sound quality, and a Ph.D. in the human response to whole body vibration, all from the University of Salford. James's work is currently focused on how object-based audio can improve the listener experience of spatial audio.