OPEN ACCESS Freely available online

Journal of the Audio Engineering Society
Vol. 65, No. 3, March 2017
DOI: https://doi.org/10.17743/jaes.2016.0070

Evaluation of Spatial Audio Reproduction Methods (Part 1): Elicitation of Perceptual Differences

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

Institute of Sound Recording, University of Surrey, Guildford, UK

There are a wide variety of spatial audio reproduction systems available, from a single loudspeaker to many spatially distributed loudspeakers. An important factor in the selection, development, or optimization of such systems is listener preference and the important perceptual characteristics that contribute to this. An experiment was performed to determine the attributes that contribute to listener preference for a range of spatial audio reproduction methods. Experienced and inexperienced listeners made preference ratings for combinations of seven program items replayed over eight reproduction systems and reported the reasons for their judgments. Automatic text clustering reduced redundancy in the responses by approximately 90%, facilitating subsequent group discussions that produced clear attribute labels, descriptions, and scale end-points. Twenty-seven and twenty-four attributes contributed to preference for the experienced and inexperienced listeners respectively. The two sets of attributes contain a degree of overlap (ten attributes from the two sets were closely related); however, the experienced listeners used more technical terms while the inexperienced listeners used more broad descriptive categories.

0 INTRODUCTION

The growth of spatial audio reproduction, particularly in the domestic listening environment, can be attributed in part to a range of technological advancements. These include the development of surround sound recording techniques, ease of manufacture of loudspeakers, and increased bandwidth for transmission of multichannel audio files. However, there is not one particular reproduction system that has gained widespread uptake as stereo sound did in the latter half of the 20th century; rather, there are various systems competing for market share.

Perhaps the most common is 5.1 surround sound, which has been standardized by the International Telecommunication Union [1]. Other similar channel-based systems have also been used [2]: 7.1 systems reduce the inter-loudspeaker angle between the rear channels or add front channels [1]; 9.1 or 11.1 layouts add height loudspeakers; and 22.2 has been introduced alongside ultra-high-definition television and includes loudspeakers surrounding the listener in four layers of elevation. In domestic environments, it can be prohibitively difficult and expensive to install a large number of loudspeakers; therefore, small array-based systems—such as sound bars that sit under a television—are often used. In addition, headphone listening is also used for consuming audio at home, in work, and while travelling. It is clear that the current state of spatial audio reproduction is complex, and it is unlikely that a single system will become ubiquitous in the near future.

When selecting, developing, or optimizing spatial audio reproduction systems, it is important to consider their perceptual characteristics; knowing why listeners prefer some systems to others would (i) help to determine the best system to use in any particular situation, and (ii) help to steer future development of systems in such a way as to produce a perceptually optimal listening experience. The aim of the work documented in this paper was therefore to determine the attributes that differentiate a wide range of spatial audio systems in terms of listener preference.

1 EXPERIMENT BACKGROUND

There has been considerable research into determining perceptual differences between audio reproduction systems (e.g., see the recent literature review by Francombe et al. [3]). This started with studies that primarily focused on timbral differences between loudspeakers [4], although it has long been recognized that spatial characteristics also constitute an important aspect of sound quality [5]. While Rumsey et al. [6] found that timbral fidelity was more than twice as important as spatial fidelity to listener preference, there have been a large number of studies that consider the spatial attributes of reproduced sound [7–11].

As discussed by Francombe et al. [3], the numerous studies produce a complex picture and do not leave clear guidelines as to which attributes should be investigated or optimized. For example, in a review of academic and popular sources, Pedersen and Zacharov [12] found 200 descriptive terms, even after removing redundant words. The relative importances of such terms are unclear, and their relationship to listener preference is not known. This problem is compounded when different studies define concepts with different labels, or use the same label for apparently distinct attributes [13]; for example, Berg [14] highlighted contrasting definitions of the term "envelopment."

One factor that contributes to the complexity is that previous studies have elicited all terms that can be used to describe differences between the stimuli under test, rather than only those that are important or relevant to listener preference. It is likely that in real listening situations, some of the attributes may make little or no contribution to listener preference and are, therefore, less important. Where studies that try to map the attributes to preference judgments have been performed (e.g., [7]), the results have been inconclusive.

Another significant challenge is presented by the rapid development of new and competing technologies. It is likely that new loudspeaker technologies or signal processing methods introduce distortions, artifacts, and new capabilities that have not previously been considered. This is a problem that is difficult to overcome without periodically reinvestigating the important perceptual characteristics, performing a wide-ranging and future-proof experiment, or determining key attributes that will always be desirable regardless of the system under test. This means that additional elicitation experiments are required when paradigm shifts occur. For example, many spatial audio elicitation studies have been limited to horizontal-only loudspeaker configurations or, where periphony has been considered, have used ambisonics rather than emerging channel-based setups such as 9.1 and 22.2; elicitation of terms related to elevated loudspeakers in these emerging systems is now required. Studies have tended to focus on a small range of techniques rather than comparing across different "families" of reproduction methods, and more recent developments (such as channel-based reproduction with height loudspeakers) have not been considered. For example, Lorho [15] investigated the perceptual characteristics of stereo audio replay in mobile devices, which might result in a different set of attributes. While it is not possible to include every potential reproduction method in an experiment, nor to completely future-proof against new technologies or the adaptation of listeners to such developments, it is still important to cover a wide range of potential reproduction methods when attempting to elicit an attribute set that is widely applicable. Similar considerations must also be made about the program material content used in experiments.

It is commonly held in the attribute elicitation literature that determining the differences between products, or making ratings on sensory attributes, should be performed by experts, while preference ratings should be made by consumers [16, p. 343]. This approach has been taken in spatial audio attribute elicitation; for example, in Zacharov and Koivuniemi's preference mapping experiment [7]. However, Francombe et al. [17] found that non-trained listeners were able to perceive and report important differences between auditory situations, and Francombe et al. [18] found that groups of experienced and inexperienced listeners perceived similar attributes when comparing real and reproduced audio—albeit with less detail from the inexperienced listeners. In order to determine attributes that can be used for improving audio quality for the general public, it is important to consider the characteristics of spatial audio replay that non-trained listeners deem to be important.

1.1 Experiment Aims

The brief literature review above suggests that there is a need to investigate expert and inexpert listener perceptions of the spatial and timbral characteristics of current spatial audio reproduction systems, including headphones and extended channel-based systems (with height loudspeakers), in order to derive a clearly-defined list of the attributes that are likely to be important to listener preference. Consequently, an experiment was designed to determine the important perceptual differences between spatial audio reproduction methods that contribute to listener preference.

The experiment methodology used in this paper was designed to compensate for the problems identified in Sec. 1 and, therefore, to produce an attribute set that: (i) contains only the most important attributes (i.e., those that directly contribute to listener preference) by basing the experimentation on attributes elicited alongside preference ratings (Sec. 3); (ii) is relevant to a representative range of loudspeaker arrangements and reproduction methods, including channel-based systems with height loudspeakers (Sec. 2.1); (iii) is derived from a variety of program material selected to encompass a wide range of both timbral and spatial characteristics (Sec. 2.2); (iv) takes the similarities and differences between the experience of trained and non-trained listeners into account (Sec. 5.2); and (v) provides clear, concise attribute labels, descriptions, and end-point scales developed by listeners through group discussions (Sec. 5).

This methodology was intended to produce an attribute set that closely reflects listener preference and could ultimately contribute to the development of a perceptual model of listener preference based on spatial audio quality.

1.2 Experiment Design

The experiment had three stages: a preference rating and free elicitation task, an automatic text clustering procedure, and a group discussion. These stages are discussed in more detail below. The methodology was based upon that used by Francombe et al. [17], which drew on techniques used in the Audio Descriptive Analysis and Mapping (ADAM) method [19]. The reproduction methods and program material used for the experiment, and the participants, are described in Sec. 2.

In the preference rating and free elicitation stage, participants were asked to make preference ratings for audio program material items replayed over various spatial reproduction methods in a paired comparison paradigm and to give reasons for their judgments. Stage one is discussed further in Sec. 3.

A reduction stage was then used to remove redundancy from the text data and facilitate the group discussion phase; without removing such redundancy, the task would have been lengthy and repetitive, which risks annoyance and boredom for the participants, resulting in lower quality results. An automatic text clustering algorithm was designed based upon the similarity of word use between the elicited phrases. Stage two is discussed further in Sec. 4.

Finally, the resulting clusters were presented to groups of listeners who were tasked with turning the individual responses into a set of attributes covering the perceptual differences that were experienced during stage one. The group discussion involved putting clusters that described essentially the same perceptual attribute into sets and then labelling, describing, and creating scale endpoints for the sets. Finally, relationships between the attributes produced by the experienced and inexperienced listeners were explored. Stage three is discussed further in Sec. 5.

The overall results are discussed in Sec. 6 alongside suggestions for further work.

2 EXPERIMENT SETUP

A limitation of existing spatial audio attribute elicitation experiments highlighted in the introduction was that the results are to some extent specific to the reproduction methods selected. The systems used in this experiment were selected to make the test as externally valid and future-proof as possible within reasonable constraints. It was also necessary to select program material items that covered a range of timbral and spatial properties. Details pertaining to the selection of reproduction methods and program materials are given in the following sections, along with a description of the experiment participants.

2.1 Reproduction Methods

The systems tested were selected from a longlist of potential reproduction methods and were intended to: cover a wide range of different types of technique (mono, channel-based, ambisonic, binaural); include commercially available methods that are commonly used in domestic listening as well as advanced reproduction methods that are used in professional audio or research; and cover a large range of expected spatial and timbral quality. The following methods were used (where a setup is specified by a letter, this refers to loudspeaker positions described in Table 1 of ITU-R recommendation BS.2051-0 [2]).

- Mono (single loudspeaker at 0 degrees azimuth and elevation)
- Low quality mono (single USB-powered computer loudspeaker with built-in DAC, positioned 16 cm behind a computer monitor at 0 degrees azimuth, −15 degrees elevation)
- Stereo (layout “A”)
- 5-channel surround (layout “B”)
- 9-channel surround (layout “D”)
- 22-channel surround (layout “H”)
- Cuboid (loudspeakers at ±45 degrees and ±135 degrees azimuth, ±30 degrees elevation)
- Headphones

The reproduction methods were set up on the Surrey Sound Sphere (a metal geodesic sphere of radius 1.9 m) in a room with dimensions of 7.85 m × 12.38 m (with a heavy curtain at 8.23 m) × 4.00 m, and an RT60 of 0.52 s to 0.26 s between 125 Hz and 4 kHz. All loudspeakers were Genelec 8020A, with the exception of the subwoofers (Mackie HRS120) and relatively low quality computer speaker (Logitech S150). The digital-to-analog converters (DACs) used were RME Fireface 800s with the exception of the computer speaker, which had a built-in DAC. The headphones used were Sennheiser HD600 open-back headphones, with a Focusrite Virtual Reference Monitoring (VRM) Box used as the DAC (only the DAC, and not the additional VRM engine, was used).

For all methods with the exception of headphones and low quality mono, a bass management algorithm was implemented in Max/MSP; the crossover frequency was set to 66 Hz (the lower limit of the Genelec free field frequency response [20]), and the subwoofer nearest to each speaker was used for the low frequency reproduction. Where the loudspeaker was equidistant from both subwoofers, i.e., on the median plane, a single subwoofer was selected arbitrarily.

The loudspeakers on the sphere were calibrated so that a −18 dBFS (RMS) pink noise signal replayed from a single loudspeaker measured 85 dBA (slow) at the listening position [21]. The subwoofers were calibrated so that the sound pressure levels (SPLs) of a single loudspeaker and each subwoofer were matched within a frequency band in which the devices each exhibited a flat frequency response [22].

2.2 Program Material

A list of desirable criteria for the program material used in the listening test was developed based on discussions between experts in various areas of spatial audio. The criteria included aspects such as: various genres and musical elements; different numbers of sources; different types of source (small/large, real/synthesized, etc.); different recording environments; different source positions and movement; and a variety of technical aspects (dynamic range, LFE, effects, etc.).

To meet these criteria, a stimulus set was created by making simultaneous recordings in different spatial audio formats as well as repurposing currently existing multitrack content. The following program material items (described in more detail by Francombe et al. [22]) were used. Each excerpt was 20 seconds long.

- Brass quintet (recorded live in the University of Surrey Studio 1 with different capture techniques for each reproduction method)
- Jazz quintet (as above)
- Pop recording (studio multitrack recording separately mixed for each reproduction method)
- Big band (multi-microphone recording in the Royal Albert Hall, separately mixed for each reproduction method)
- Sport (football match broadcast with commentary separately mixed for each reproduction method, including first-order ambisonic decode of Soundfield crowd microphone)
- Experimental music (first-order B-format rendering of *Rotating psychoacoustic tuning curves* by Florian Hecker, decoded to each reproduction method)
- Film excerpt (5.1 clip from *Skyfall* with dialogue, music, and sound effects; manually downmixed and upmixed to each different reproduction method, with added first-order B-format rain decoded to different loudspeaker arrangements)

It should be noted that it is not possible to fully separate the reproduction method from the production technique used, and that future capture, mixing, and encoding techniques may potentially change the listening experience for a particular reproduction method. However, a range of state-of-the-art methods (including mixing by experienced practitioners and direct microphone capture) was used in order to maximize the range of elicited attributes and to ensure generalizable results.

2.3 Participants

All stages of the experiment were performed by the same two groups of participants: seven experienced listeners and eight inexperienced listeners. The experienced participants were fourth year undergraduate students on the Music and Sound Recording course at the University of Surrey, Guildford, UK, all of whom had passed a technical ear training module and had critical listening experience in recording studios. The inexperienced participants 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.

3 STAGE ONE: FREE ELICITATION

The purpose of the free elicitation stage was to collect a pool of text data that contained listeners' reasons for giving particular preference judgments, so that these phrases could form the basis of subsequent group discussion sessions.

The stimuli were presented as paired comparisons between the reproduction methods, and listeners were asked to rate their degree of preference for stimulus A or B (or no preference). With eight methods, there are a total of $\binom{8}{2} = 28$ comparisons. In order to facilitate analysis of participant reliability, three comparisons were repeated for each program item. The repeated items were selected randomly from the possible 28 combinations. Different random selections were made for each program item, but all listeners repeated the same stimuli. This resulted in a total of 31 judgments per participant per program item. All judgments for a single program item were made in one test session, requiring a total of seven sessions per participant. The program items were presented in a different random order for each participant. Within each session, the individual comparisons were presented on separate user interface pages and randomly ordered. The preference ratings—including further experimental details, a depiction of the user interface, and analysis of listener reliability—are further discussed by Francombe et al. [23].

The user interface was created in Max/MSP, with one paired comparison of stimuli on each test page. Participants were asked to make their preference ratings on a continuous horizontal slider and required to type their reasons for giving a particular preference rating on separate lines into a text box on the screen. The wording of the instructions pertaining to the free elicitation response was as follows.

"In the box provided, please type the factors that led to your preference choice. Please type each factor on a separate line. You are not asked to list all of the differences between the two stimuli; please list only those factors that contributed to your preference decision. However, please try to be as specific as possible. For example, it is not sufficient to simply say that you felt A was "better" or "worse" than B; please say which aspects of the stimuli led to you feeling this way. You may use positive or negative terms. It is possible that for a given stimulus pair you might have a large preference for some aspects of A and a large preference for other aspects of B; this might lead to an overall preference that is not especially large one way or the other, or no preference. If you preferred some aspects of one stimulus and some aspects of the other, please type all of these aspects."

3.1 Free Elicitation Results

Prior to the group discussion stage, simple analysis was performed on the text data to begin to understand the responses.

A total of 6806 responses were collected from the 15 listeners (39107 words). The mean phrase length was 5.75 words. Of the 6806 responses, 4220 were unique phrases. The experienced listeners produced a total of 3773 responses—of which 1967 were unique (52.13%)—with a mean of 4.72 words per phrase. The inexperienced listeners produced a total of 2740 responses—of which 2270 were unique (82.85%)—with a mean of 7.12 words per phrase. These results suggest that, on the whole, the inexperienced listeners were more verbose and their responses were more varied, which could be attributed to their lack of prior knowledge of suitable descriptive terminology. However, the number of responses given by each individual listener (Fig. 1) indicates that there are substantial individual differences even within the two listener groups, in terms

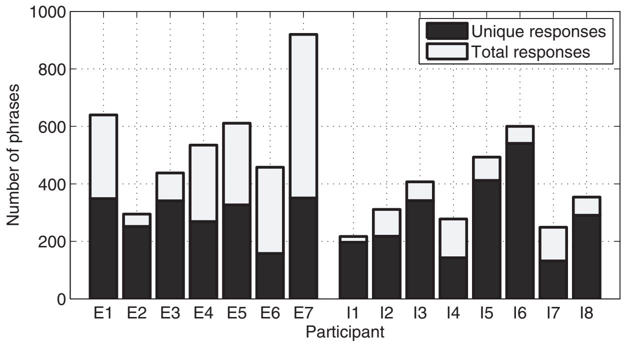


Fig. 1. Number of responses (total and unique) in the free elicitation task for each listener. The prefix "E" indicates an experienced listener; "I" indicates an inexperienced listener.

of the total number of responses and the proportion that are unique.

A quick, informal method for investigation of a large text data set is to create "word clouds," which visualize the data by representing the frequency of word use (after the removal of stop words) by the size of a word. Such word clouds were generated using *Wordle*[1] [24] for the experienced and inexperienced listener responses (including all elicited text, i.e., without filtering duplicated responses). The experienced listener cloud is dominated by the term *enveloping*, suggesting that this may be an important factor. Other technical terms (*mono*, *distorted*, *hf*, *image*, *wider*, and so on) also stand out, as well as some descriptive language (e.g., *sounds*, *better*, *much*). The inexperienced listener cloud highlights a lot of less specific language, more often focussing on describing aspects of the experimental methodology (words such as *sound(ed)*, *headphones*, *prefer*, *like*, *feels*, and *stimulus*); however, some descriptive terminology is also present (*unidirectional*, *surround*, *clearer*, and so on).

3.2 Free Elicitation Conclusions

The free elicitation was designed to create a pool of textual responses that described the listeners' reasons for giving particular preference judgments and could be used as the input to a group discussion stage in order to produce attribute sets. By asking listeners to write down only the factors that significantly contributed to their preference judgments, it was hoped that the responses would be succinct and avoid many terms that had only a small influence. However, the results from the first stage suggest that this may not have been the case, with 1967 and 2270 unique phrases written by the experienced and inexperienced listeners respectively. It was consequently considered necessary to perform a redundancy reduction stage prior to the group discussion (see Sec. 4). In the experiment upon which this methodology was based [17], the group discussion stage took approximately 5 hours to perform for each group with 263 and 317 unique phrases for the experienced and inexperienced listeners respectively. Extrapolating to the number of elicited phrases for this experiment indicated that the group discussion stage would be unfeasibly long and, therefore, likely to result in participants becoming bored, fatigued, and disillusioned with the process.

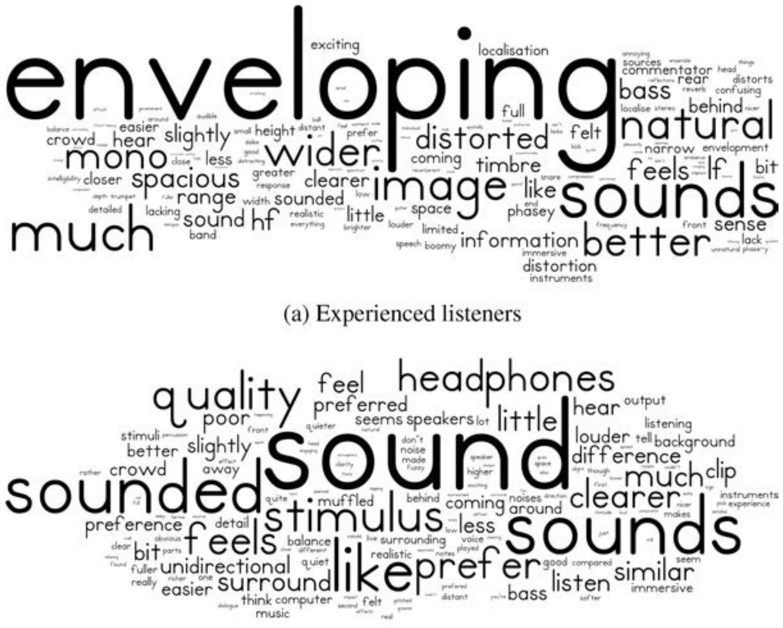
4 STAGE TWO: AUTOMATIC CLUSTERING

Using automatic reduction methods to reduce the number of responses from the free elicitation stage could significantly reduce the length of the discussion process and thus lead to more considered judgments and a higher quality outcome. However, using automatic reduction has the disadvantage of potentially obfuscating subtle nuances used in the language, especially when the participants that contributed the words are present in the group discussion and able to explain their meanings. Consequently, it is necessary to achieve a compromise in reduction that facilitates the group discussion while not obscuring unique attributes from the human participants.

Reduction methods have been used in previous studies: Zacharov and Koivuniemi [7] truncated words from a free elicitation to their first five letters and rejected words with similar roots in order to reduce the terms for a group discussion; and Guastavino and Katz [8] reduced elicited terms to their root forms ("lemmata"), grouped synonyms, and used a thesaurus to group terms by "semantic themes." While both of these methods can successfully reduce terms, they suffer from some disadvantages. Simply truncating words and completely rejecting other variants means that some terms are not carried forward to the group discussion; this reduces the benefit gained by holding a group discussion with participants who have listened to the stimuli and produced the descriptive terms. However, using humans to perform a preliminary reduction (possibly with the aid of a thesaurus) could potentially introduce a source of bias and is not objective or repeatable. It was therefore considered necessary to develop a repeatable algorithmic method of grouping the elicited responses, while still being able to present all of the data to the listeners so that the automatic grouping could be overridden if desired.

The process of analyzing large sets of textual data to search for relationships is known as "text mining"; it varies greatly in scope from the simple stemming described above to complex algorithms that aim to extract information from large data sets [25, pp. 1–13]. In this case, a clustering algorithm based on the similarity of word use between each response was implemented in MATLAB. The resulting clusters—containing elicited phrases that the algorithm considered to have similar content—were then presented to the participants. This had the effect of greatly reducing redundancy in the dataset while allowing participants to potentially identify terms that had been misidentified. The clustering algorithm is described in more detail in the following section.

[1] It should be noted that *Wordle* uses a set of stop words that may differ from those used in the clustering procedure described in this study.



(a) Experienced listeners

(b) Inexperienced listeners

Fig. 2. Word clouds showing all responses from both groups of listeners. The word size is proportional to frequency of use.

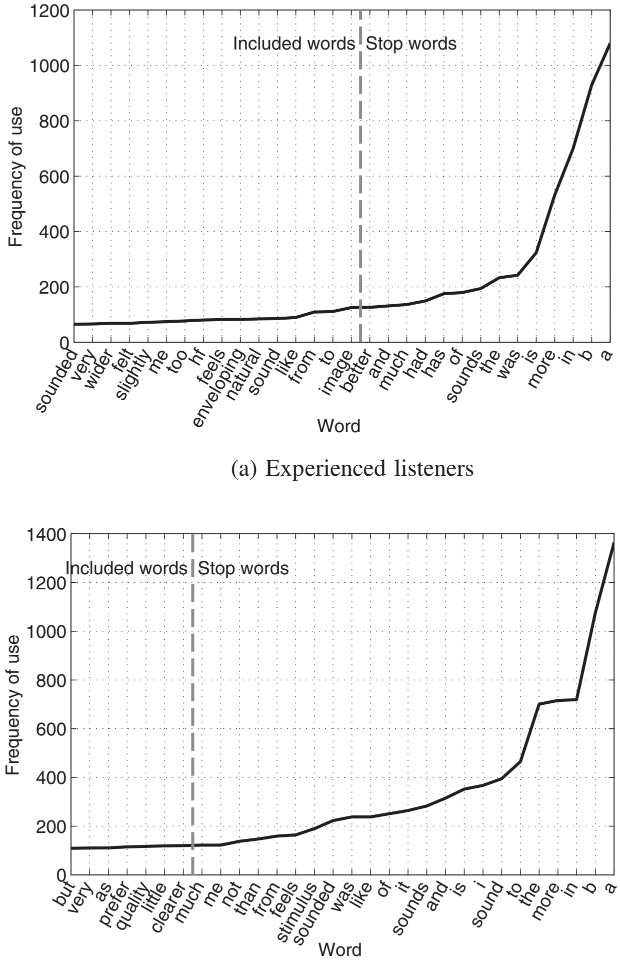
4.1 Clustering Algorithm

The following procedure was followed (separately for the results from each listener group) based on techniques described by Weiss et al. [25, pp. 114–116]. Prior to the clustering, the data were preprocessed to remove the exact duplicate responses described in Sec. 3.1 (the automatic clustering should combine any duplicates, so removing them at this stage aided in clarifying the output that was presented to listeners) and empty lines, and all responses were converted to lower case.

4.1.1 Stop Word Determination

Stop words are words that have little to no predictive value in the dataset. This can occur for a number of reasons: words may be very common in the language, therefore likely contributing more to the syntax of the response than to the semantic content; words may occur very frequently in the dataset (a word that appears in every response will have no power to group similar responses); or words may occur very rarely in the dataset (a word that only appears once in the dataset also has no power to group similar responses) [25, pp. 26–28]. Removing stop words facilitates clustering by the essence of each response rather than by superfluous words that might be syntactically necessary but are not semantically important. The following criteria were used to determine stop words.

- A list of all individual words used and their frequencies was generated (this is known as a "dictionary").
- The most frequently used terms were plotted in order of frequency of use (Fig. 3), and the most-used word that could potentially have predictive power (determined by the first author given knowledge of the stimuli under test) was identified. Words more common than this were



(a) Experienced listeners

(b) Inexperienced listeners

Fig. 3. Most frequently used words in the free elicitation stage. The vertical line indicates the cutoff point for stop words.

considered to be stop words: 14 and 23 words were identified from the experienced and inexperienced datasets respectively.
- The words only used once (383 by the experienced listeners and 552 by the inexperienced listeners) were considered to be stop words.
- Additionally, a standard set of 571 stop words (from the SMART information retrieval project [26]) was used. However, this list was modified to remove any terms that might in fact be relevant (as determined by the first author) given the stimulus set. The following terms were removed from the stop word list: *above*, *around*, *behind*, *below*, *beside*, *besides*, *clearly*, *far*, *outside*, and *sub*. The stop word list used is included in the dataset that accompanies this paper.

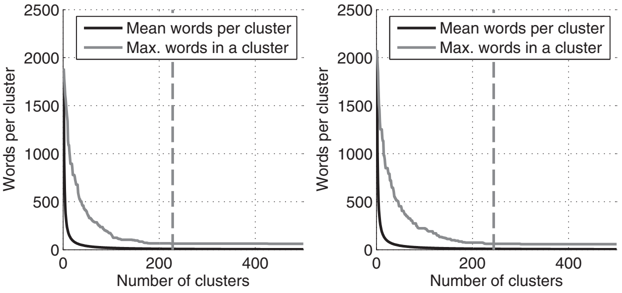
4.1.2 Dictionary Generation

A dictionary was generated after removal of the stop words. Any words felt to have limited predictive power, but that had not been identified as stop words, were manually added to the stop word list (52 and 61 words for the experienced and inexperienced listeners respectively). The words removed included intensifiers and instrument/sound names.

4.1.3 Clustering

The clustering was performed using the following algorithm.

1) Each word of the responses, the dictionary, and the stop word list was "stemmed"; that is, the words were truncated to the first N letters (where N was set at 5 based upon the value used by Zacharov and Koivuniemi [7] as well as experience with using the clustering algorithm). Stemming words ensures that similar terms are grouped together and also helps to mitigate the effect of spelling mistakes on the clustering (e.g., *enveloped*, *enveloping*, and *envelpoing* [sic] would all be coded as *envel*).
2) The dictionary and stop word list were regenerated to account for any new overlap caused by the stemming process.
3) An n-by-m matrix was generated, where n was the number of responses and m the number of words in the dictionary.
4) For each row—that is, each response—each cell of the matrix was populated with a one if the dictionary word in that column was contained in the phrase, and a zero if it was not.
5) Any all-zero rows (i.e., phrases where all of the words had been classified as stop words) were removed from the matrix (and therefore these phrases were unclustered).
6) Agglomerative hierarchical clustering was performed on the resulting matrix using the "Ward" linkage method. This technique initiates each row of the matrix as an independent cluster and merges the two clusters that result in the minimum increase in within-cluster variance at each step.



(a) Experienced (b) Inexperienced

Fig. 4. Mean number of words per cluster and maximum number of words in a cluster for 1 to 500 clusters. The dashed vertical lines show the final number of clusters used.

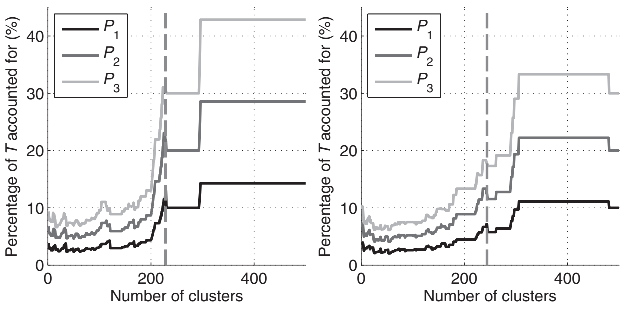
It should be noted that words that are designated as stop words are not used for determining the cluster into which a response should fall; however, they do remain present in the final clustered responses (i.e., the responses are unchanged by the procedure).

4.1.4 Determination of the Number of Clusters

The agglomerative clustering algorithm produced the maximum number of clusters that it was allowed to in each case. Therefore, it was necessary to determine an appropriate number of clusters to allow. This number was a compromise between too few clusters (which increases the chance of obscuring unique attributes) and too many clusters (which would leave the subjects with many clusters to group). The clustering algorithm was run multiple times to generate from 1 to 500 clusters (selected as a suitable upper limit based on the time taken to perform group discussion in a similar experiment [17]), and various statistics that described the clusters were examined (as well as manual observation of the clusters to check their suitability).

To ensure that participants were able to easily classify the terms in each cluster, it was desirable that each cluster contained a reasonable number of responses and that no cluster contained a very large number of responses. If too few phrases were clustered, the participants would waste time grouping similar responses; if too many phrases were clustered, it would be time consuming and complicated for participants to divide clusters. It was also notable that when the clustering algorithm produced a single cluster with a large number of responses, the content was much less focused. Therefore, the mean and maximum cluster size was considered. Fig. 4 shows the mean and maximum cluster size for 1 to 500 clusters for both groups of listeners. As the number of clusters increases, the mean number of words per cluster drops off quickly; the maximum cluster size stabilizes at approximately 60 after around 200 clusters.

It may also be expected that the clustering is most effective (i.e., the created clusters contain separate attributes) when the number of dictionary words within each cluster is small. In order to evaluate this, the proportion of each dictionary word's contribution to each cluster was evaluated. The total number of dictionary words T in any particular



(a) Experienced (b) Inexperienced

Fig. 5. Minimum percentage of clustering terms accounted for by P_1, P_2, and P_3. The dashed vertical lines show the local maxima that were used to determine the number of clusters.

cluster is the total number of words in the cluster minus the number of stop words in the cluster. The proportion of T accounted for by the most frequently used dictionary word in a cluster is denoted P_1, the second most frequently used dictionary word by P_2, and the nth most frequently used word by P_n, so that $\sum_{n=1}^{N} P_n = T$, where N is the number of unique dictionary words.

Where a cluster contains a large number of dictionary words with an approximately equal proportion, the cluster does not clearly reflect a single perceptual attribute. Conversely, where a cluster contains a small set of dictionary words that account for the majority of the words in that cluster, the cluster tends to more clearly reflect a single perceptual attribute. Therefore, the proportion of each cluster accounted for by the three most common dictionary words was evaluated. The minimum percentage of clustering words across all clusters accounted for by P_1, P_2, and P_3 was plotted (Fig. 5), again for 1 to 500 clusters. The figures show a distinctive pattern of a steady increase in percentage followed by two plateaus. For both listener groups, the most pronounced local maximum fell above 200 clusters, which had been identified above as the point required for a reasonable maximum cluster size. Therefore, this point was used to specify the number of clusters that were generated (228 clusters for the experienced listeners and 244 for the inexperienced listeners).

The statistical analysis detailed above coupled with observation of the produced clusters was intended to ensure a balance between producing useful clusters that would enable the participants to perform the group discussion while providing enough redundancy reduction to make the discussion procedure feasible.

4.1.5 Final Adjustments

After deciding on the number of clusters, some small adjustments were made before rerunning the algorithm. The unclustered items were manually reviewed in order that any obvious spelling mistakes could be corrected. Following these corrections, the number of words only used once was recalculated. The final clusters were generated with these minor adjustments.

The clusters were labelled with up to three of the most commonly used clustering word stems for each cluster. The output that was presented to the participants also included a bar chart showing all clustering words and the percentage of T that each word accounted for. Clustering words were shown in bold text. For example, the experienced listener cluster labelled "**locat**, **preci**" included five responses: "easier to **precisely locate** things in b," "can **locate** more **precisely** in b," "easier to **locate** more **precisely** in a," "easier to **precisely locate** in a," and "easier to **locate** more **precisely** in b." Similarly, the inexperienced cluster labelled "**source**, **uncer**, **varie**" included five responses: "definite **separated sources** in b," "mono **source** in a less preferable than b," "varying **location** of the **source** of b **creates** more **uncertainty**," "b feels like smaler [sic] **source** of the **noise**," and "b **varies** more in where the **source** is, **uncertainty**." This is an example of a cluster in which there may be more than one concept expressed, highlighting the benefit of using a clustering and presentation method in which participants can see the individual sentences in the cluster and split them into different groups if desired.

4.2 Clustering Conclusions

An automatic text clustering algorithm was used to reduce redundancy in the data produced in the free elicitation. The method focused on determining relevant words from the content and then grouping responses that contained similar words. The outcome was a reduction of 1967 unique responses to 228 clusters (88.5% reduction) for the experienced listeners, and 2270 unique responses to 244 clusters (89.3% reduction) for the inexperienced listeners. It is difficult to assess the success of the method with no ground truth data available; however, one advantage of clustering rather than merging responses is that all of the original responses were presented to the participants, who were able to make any corrections where there were obvious mistakes. Therefore, no responses were lost; participants could view all responses and split or merge clusters as they found appropriate.

The experienced listeners' clusters ranged from 1 to 68 responses, with a mean of 8.4 responses and standard deviation of 8.9 responses per cluster. The inexperienced listeners' clusters ranged from 1 to 61 responses, with a mean of 8.5 responses and a standard deviation of 7.7 responses per cluster.

5 STAGE THREE: GROUP DISCUSSIONS

The aim of the group discussion stage was to reduce the responses from the free elicitation experiment—via the clustering procedure—into attributes that describe the important perceptual differences between spatial audio reproduction methods. The group discussion procedure was based on the methods employed by Francombe et al. [17] and Zacharov and Koivuniemi [7]. The discussions were performed separately by the two listener groups. The following tasks were required of the participants: grouping, attribute labelling, attribute definition, and end-point labelling.

In the grouping stage, the clusters were presented sequentially to the group, who were asked to put into sets the clusters that contained terms relating to the same perceptual experience. The following wording was used to introduce the task.

"The aim of these sessions is to determine the characteristics (or 'attributes') of audio replay that contribute to listener preference. In the first stage, you each made preference ratings and gave reasons for your judgments. There were a large number of reasons collected, and it's likely that some actually refer to the same perceptual experience (even if the word used was not necessarily the same). In these group discussion sessions, the reasons you gave will be presented back to you, and your first task is to group them into sets of terms that describe essentially the same experience or reason for preferring different types of audio replay. Because there was such a large number of terms elicited, they have been pre-clustered into sets using an automatic process designed to attempt to predict similar groups. Because the process is automated, it may not always get it right, so it might be necessary for you to divide or join clusters. The clusters have been labeled with up to three of the most commonly used word stems in that cluster, and there is also a bar chart showing the frequency of word use within the cluster. Words that were used to create the cluster are shown in bold font."

During the grouping, participants were allowed to discard clusters if they felt that none of the responses were relevant or if they felt that it was difficult to determine exactly which group a cluster fell in yet they were confident that the percepts referred to had all been covered in other groups. This helped to streamline the process by preventing long, complicated discussions that would not have uncovered any novel attributes.

Following the grouping task, participants were asked to suggest an attribute label, attribute definition, and scale end-point labels for each group. The following wording was used to introduce the task.

"Once all of the terms (or clusters of terms) have been put into groups, the next task is to produce: one word or short phrase to label that group; a brief description that you could use to explain to someone that hadn't taken part in the original experiment what is meant by that group; and a set of scale end-points if that attribute was to be rated on a scale."

5.1 Results

In the following subsections data from the experienced and inexperienced listeners are analyzed in turn.

5.1.1 Experienced Listeners

The experienced listeners completed the grouping task in two sessions (1 hr 25 mins and 1 hr 43 mins) and the attribute development tasks in two further sessions (1 hr 27 mins and 1 hr 35 mins). The four sessions took a total of 6 hrs 10 mins. The participants converted the 228 clusters into 27 attributes, which are listed alongside their descriptions and scale end-points in Table 1.

5.1.2 Inexperienced Listeners

The inexperienced listeners spent two sessions performing the grouping task (1 hr 31 mins and 1 hr 26 mins), before finishing the grouping and performing the attribute development in one further extended session (2 hrs 59 mins plus breaks). The three sessions took a total of 5 hrs 56 mins. The participants converted 244 clusters into 24 attributes, which are listed alongside their descriptions and scale end-points in Table 2.

5.2 Attribute Set Comparison

The two attribute sets were produced by listeners with different backgrounds and experience, and it is therefore possible that the two types of listeners find different aspects of the experience to be important when comparing spatial audio reproduction methods. However, it is also likely that there is at least some overlap between the two attribute sets. To investigate this, participants were asked to compare the two attribute sets. The participants were each sent a spreadsheet with the attribute labels and descriptions from the set that their group elicited in rows and the attribute labels and descriptions that the other group elicited in columns (both lists were randomly ordered, with different orders for each participant). They were asked individually to complete the spreadsheet by entering a number from 0–2 into each cell, where 0 indicated that there was no overlap between the two attributes; 1 indicated that there was some overlap between the attributes but that they were not identical; and 2 indicated that the attributes were identical (they described the same aspect of the listening experience, even if they used different words). Participants were asked to make judgments based primarily on the attribute descriptions, using the labels for guidance if necessary. It should be noted that the inexperienced listeners did not receive any training on technical vocabulary before filling out the spreadsheet; they had the attribute descriptions to refer to and were instructed to contact the experimenter if they had any questions (no such questions were asked).

From the 15 participants, 10 spreadsheets were returned (6 by experienced listeners, and 4 by inexperienced listeners). Table 3 contains the attributes that were felt to be identical (i.e., given a score of 2) by 50% or more of the participants that returned the spreadsheets. The results are unsurprising given the similarities of the definitions. Looking at the attributes that were felt to be similar but not identical presents a much more complex picture; only 146 of the 648 (22.5%) attribute combinations were never considered to have some degree of overlap. This points to the difficulty of using language to describe the characteristics of spatial sound and possibly goes some way towards explaining the difficulty of determining a small set of useful attributes (as discussed in Sec. 1).

The attribute comparison data can also be used to suggest attributes that are unique to each set. No single attribute from the experienced or inexperienced listener set scored

Table 1. Attributes produced by the experienced listeners

Label	Description	Scale end-points
Amount of distortion	Overall level of distortion	Distorted → clean
Audibility of compression	How audible any artificial compression is	No audible compression → very highly compressed
Bandwidth	The difference between the highest and lowest frequency	Band limited → full spectrum
Depth of field	Perceived proximity of sources	Close → distant
Dynamic range	The difference in level between the loudest and quietest point	Small range → large range
Ensemble balance	Relative levels of the different sources	Balanced → unbalanced
Enveloping	How immersed/enveloped you feel in the sound field	Fully enveloping → not at all enveloping
Horizontal width	360 degree horizontal width	Point source → surrounding
Level of reverb	Amount of reverb	Wet → dry
Overall level	The overall loudness of the reproduction	Quiet → loud
Overall spectral balance	The magnitude of broad cuts and boosts in the spectrum	None (perceived as flat) → significant
Overall subjective preference	Aspects such as excitement, engagement, impressiveness, and impact	Dislike → like
Perceived transducer quality	The perceived quality of the speaker over which it was reproduced	Good → bad, low quality → high quality
Perceptibility of noise	How perceptible the noise is	Perceptible → imperceptible
Phasiness	Level of phasiness and corresponding discomfort	Not phasey → uncomfortably phasey
Physical sensation	Physical sensations created by the reproduction system	None → many
Realism	Overall, how realistic it sounds	Unnatural → natural, unreal → real
Realism of reverb	How realistic the reverb sounds	Natural → unnatural
Sense of space	The extent to which you feel you are in the same space in which the music/event was performed	Not at all → very much
Spatial balance	How biased the reproduction is to particular area(s) of the sound field (including hole in the middle)	Even → biased
Spatial clarity	Ease of localization of individual sources	Distinct → indistinct
Spatial movement	Degree of movement of sound sources	Static → dynamic
Spatial naturalness	How natural the source position is within the 3D image	Natural → unnatural
Spatial openness	How claustrophobic the sound feels. The proximity of the 3D sound field. A sense of air/openness	Suffocating → open
Spectral clarity	The ability to distinguish different sources based on their spectral content (timbre)	Indistinct → precise
Spectral resonances	Presence of unpleasant resonances/presence of sharp peaks in the spectrum	Resonant → not resonant
Subjective quality of reverb	How pleasant the reverb sounds	Pleasant → unpleasant

zero overlap with every other attribute; however, the sum (over participants and attributes) of similarity values for each attribute was calculated and the attributes put into rank order. The attributes that scored lower than the lowest scoring attribute from Table 3 (*overall level* and *echo* for the experienced and inexperienced listeners respectively) were *audibility of compression*, *dynamic range*, *spatial movement*, *bandwidth*, *overall spectral balance*, *phasiness*, *subjective quality of reverb*, *horizontal width*, *spatial naturalness*, and *amount of distortion* for the experienced listeners, and *can't hear difference*, *treble*, *emotional reaction*, *bass*, and *headphones* for the inexperienced listeners. The unique experienced listener attributes are mainly technical terms, while the unique inexperienced listener attributes are either broad descriptive categories or, potentially, sub-attributes of others in the experienced listener set.

With the available data, it is not possible to say definitively whether the differences in listener groups arose due to differences in perception, preference, or vocabulary. However, the fact that some attributes appeared to have little overlap even with the simple descriptions suggests that different attributes are important to the two listener groups. This is supported by similar findings in literature; for example, Rumsey et al. [6] found that frontal spatial fidelity was more important to experienced listeners than inexperienced listeners, and that surround spatial fidelity was more important to inexperienced listeners than experienced listeners.

5.3 Group Discussion Conclusions

In the group discussions, a total of 51 attributes were developed: 27 for the experienced listeners and 24 for the inexperienced listeners. There is undoubtedly some overlap between the two attribute sets, as highlighted by the results of the comparison analysis. The attributes produced are similar to attributes seen in previous studies, with some exceptions that are likely due to the more extended range of reproduction methods and program material items included as well as the use of experienced and inexperienced

Table 2. Attributes produced by the inexperienced listeners

Label	Description	Scale end-points
Bass	The level of bass in the sound	Low → high
Can't hear difference	The two things sound the same	
Clarity	How clear a sound is to the listener	Muffled → clear
Detail	The amount of details within the individual sounds that you can discern	Little detail → lots of detail
Discernibility	Whether you can pick out individual sounds	I can pick out the individual sounds → I can't pick out the individual sounds
Distance	Perceived distance of sound from the listener	Close → far
Ease of listening	The effort required to listen to the sound	Easy → hard
Echo	How echoey the sound is	No echo → very echoey
Emotional reaction	The emotional response that the track elicits, e.g., threatening or calming	Very positive → no effect → very negative
Harshness	Is the sound soft, gentle, mellow, or is it harsh, piercing, painful?	Soft → harsh
Headphones	Does wearing headphones enhance the experience?	Not at all → a lot
Immersion	How immersed/involved you feel in the sound	Uninvolved → involved
Odd sounds	The presence of odd, unusual, or unnatural sounds	Nothing odd → very odd
Output quality	The speaker/recording quality. Is it crisp, or is it tinny or fuzzy?	Poor → good
Physical reaction	A reaction to the physical components of the music, i.e., vibrations	No reaction → strong reaction
Position of sound	Your preference depending on the position of the sound(s)	Like → dislike
Prominence	The prominence of a sound relative to other sounds in the track	No prominent sound → a prominent sound
Realism	Closer to the original sound/the real experience	Artificial → real
Richness of sound	How much body the sound has. Is it full, rich, deep?	Flat → full
Spatial balance	Left/right, front/back, up/down distribution of sound	Balanced → unbalanced
Surrounding	The spread of sound around the space	Unidirectional → surrounding
Targeting	The degree to which the sound is targeted towards the listener. How personal it is to the listener	Untargeted → targeted
Treble	The level of treble in the sound	Low → high
Volume	The volume of the track	Quiet → loud

Table 3. Attributes that were felt to be identical by 50% of participants or more.

Attribute		
Experienced	Inexperienced	Pct. (%)
Physical sensation	Physical reaction	100
Overall level	Volume	100
Realism	Realism	100
Perceived transducer quality	Output quality	100
Enveloping	Immersion	90
Depth of field	Distance	80
Spatial balance	Spatial balance	70
Ensemble balance	Prominence	60
Spatial clarity	Discernibility	50
Level of reverb	Echo	50

listeners; for example, *spatial movement*, *spatial balance*, and *depth of field* from the experienced attributes, and *ease of listening*, *emotional reaction*, *headphones*, and *targeting* from the inexperienced listeners.

With such large attribute sets, it is difficult to suggest which of the attributes are the most important ones; this is the subject of further investigation [23].

The definitions suggest that there are multiple types of attribute. Some are very general and could be considered "umbrella categories" under which other more specific attributes are placed; for example, *realism of reverb* could fall under the category *realism*. Others relate to aspects of the stimulus such as the position of sounds or the presence of particular technical elements. Finally, some attributes relate to aspects of the overall listening experience; for example, *ease of listening* from the inexperienced listener attributes.

The participants reported that the task was understandable and achievable although difficult to complete. Reasons for this included the large number of terms (even after clustering) and the high cognitive load involved, resulting in participants becoming fatigued. Consequently, the process was time-consuming (approximately six hours for each group).

6 DISCUSSION AND CONCLUSIONS

In the following sections the experimental work presented above is summarized and discussed, and then the paper's overall conclusions are drawn.

6.1 Stage One: Free Elicitation

The free elicitation was performed alongside a preference rating task in an attempt to elicit only those attributes that contributed to listener preference. However, a very large dataset was collected, and it is likely that some of the elicited responses are more important than others. The large set of terms may be due to the fact that the elicitation was long and listeners felt the need to write different terms rather than repeating themselves, regardless of the instructions given. Using a multiple stimulus presentation when conducting a free elicitation (as in Francombe et al. [17]) may help to mitigate this. However, the methodology was successful at producing text data that could feed into a group discussion and also facilitated detailed analysis of the written responses alongside the preference data. Word clouds of the free elicitation responses suggested differences between the types of language used by experienced and inexperienced listeners; the experienced listeners were more technical while the inexperienced listeners were more descriptive. The term “enveloping” stood out as being very frequently used by experienced listeners.

6.2 Stage Two: Automatic Clustering

The automatic clustering stage was required to reduce the vast text dataset collected in the first stage in order that group discussions could reasonably be performed. Simple reduction strategies have been used in the past; however, the authors are not aware of such a clustering method being employed in an audio attribute elicitation experiment and feel that the method possesses significant advantages. The method gave a large reduction of redundancy in the data (approximately 90%), providing time savings that allowed the group discussions to be manageable. Clustering redundant terms rather than merging or removing them also enabled participants to access all of the data and make their own decisions as to any responses that belonged in different groups. Using such an algorithm also produces results that are objective and repeatable. Any use of an automatic clustering method can result in the loss of some nuance in the text data, but this was a necessary trade-off in order to allow the participants in the group discussion to perform the task without becoming fatigued or bored. The method appeared to work well; the group discussions were performed in a comparable time to those reported by Francombe et al. [17] even given the higher number of responses (before clustering).

6.3 Stage Three: Group Discussion

In the group discussion, a total of 51 attributes were developed. It is likely that participants can differentiate between stimuli on all of the attributes that were produced, but it is still of interest to determine those that contribute most to listener preference. The attributes that were produced in this study are in many cases similar to those produced in previous elicitation experiments, but some stand out as being notably different; for example, *spatial movement*, *audibility of dynamic compression*, *physical sensation*, and *depth of field* from the experienced listeners’ attributes, and *targeting*, *distance*, *ease of listening*, *odd sounds*, *emotional response*, *headphones*, and *physical reaction* from the inexperienced listeners’ attributes.

6.4 Conclusions and Future Work

Many spatial audio reproduction systems are currently in use. Understanding why listeners prefer some systems to others could help to determine the best system for any particular application and to steer future system development towards a perceptually optimal listening experience. The aim of the work documented in this paper was to determine the attributes that differentiate spatial audio systems in terms of listener preference.

A literature review revealed a number of limitations in the existing work: (i) spatial and timbral attributes were not often considered together; (ii) most studies included a limited range of reproduction methods, often without emerging channel-based methods such as 9.1 and 22.2; (iii) preference was often evaluated by inexperienced listeners, while attributes were elicited by experienced listeners; and (iv) the resulting attribute lists were complex and not necessarily relevant to listener preference. Consequently, an experiment was designed in which trained and untrained listeners auditioned program material embodying a wide range of both timbral and spatial characteristics, reproduced over a range of current reproduction systems (from low-quality mono to surround sound systems with height loudspeakers, and including headphones), and gave the reasons for their preferences for one system or another. An automatic clustering procedure was followed by group discussions to generate clear, concise attribute labels, descriptions, and scale end-points.

The perceptual differences between reproduction methods are described by the two sets of attributes in Tables 1 and 2. There is a degree of agreement between experienced and inexperienced listeners (described in Sec. 5.2) but some attributes were unique to the experienced listeners (mainly technical terms) and others were unique to the inexperienced listeners (broad descriptive categories and possible sub-attributes of those in the experienced listener set).

Further work will provide a fuller understanding of how the elicited attributes correspond to the preference ratings that were collected simultaneously. Future work might also focus on determining correlations between physical properties of the reproduced sound and the perceptual attributes, enabling the creation of perceptual models. Such models are beneficial when testing new or existing systems in order to ensure that the systems are optimized to produce the best possible listening experience. Predictive models can also be used in adaptive systems so that parameters can be adjusted in real-time to ensure an optimal listening experience. Identifying the important perceptual characteristics of spatial audio reproduction is important as it enables development of new systems (or improvement to existing systems) to focus on improving the characteristics that matter most to listeners.

7 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.00808897.

8 REFERENCES

[1] ITU-R rec. BS.775-3, “Multichannel Stereophonic Sound System With and Without Accompanying Picture,” Tech. rep., ITU-R Broadcasing Service (Sound) Series (2012).

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

[3] J. Francombe, T. Brookes, and R. Mason, “Perceptual Evaluation of Spatial Audio: Where Next?” *Proceedings of the 22nd International Congress on Sound and Vibration, Florence, Italy, 12–16 July* (2015).

[4] A. Gabrielsson and H. Sjogren, “Perceived Sound Quality of Sound-Reproducing Systems,” *J. Acoust. Soc. Am.*, vol. 65, pp. 1019–1033 (1979), https://doi.org/10.1121/1.382579.

[5] T. Letowski, “Sound Quality Assessment: Concepts and Criteria,” presented at the *87th Convention of the Audio Engineering Society* (1989 Oct.), convention paper 2825.

[6] 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.

[7] 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).

[8] C. Guastavino and B. Katz, “Perceptual Evaluation of Multi-Dimensional Spatial Audio Reproduction,” *J. Acoust. Soc. Am.*, vol. 116, pp. 1105–1115 (2004), https://doi.org/10.1121/1.1763973.

[9] J. Berg and F. Rumsey, “Verification and Correlation of Attributes Used for Describing the Spatial Quality of Reproduced Sound,” presented at the *AES 19th International Conference: Surround Sound—Techniques, Technology, and Perception* (2001 Jun.), conference paper 1932.

[10] 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.).

[11] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, “A Spatial Audio Quality Inventory (SAQI),” *Acta Acustica united with Acustica*, vol. 100, pp. 984–994 (2014), https://doi.org/10.3813/AAA.918778.

[12] T. Pedersen and N. Zacharov, “The Development of a Sound Wheel for Reproduced Sound,” presented at the *138th Convention of the Audio Engineering Society* (2015 May), convention paper 9310.

[13] F. Rumsey, “Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm,” *J. Audio Eng. Soc.*, vol. 50, pp. 651–666 (2002 Sep.).

[14] J. Berg, “The Contrasting and Conflicting Definitions of Envelopment,” presented at the *126th Convention of the Audio Engineering Society* (2009 May), convention paper 7808.

[15] G. Lorho, “Perceptual Evaluation of Mobile Multimedia Loudspeakers,” presented at the *122nd Convention of the Audio Engineering Society* (2007 May), convention paper 7050.

[16] H. T. Lawless and H. Heymann, *Sensory Evaluation of Food: Principles and Practices* (Springer, New York, USA, 1999), p. 343.

[17] 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).

[18] J. Francombe, T. Brookes, and R. Mason, “Elicitation of the Differences between Real and Reproduced Audio,” presented at the *138th Convention of the Audio Engineering Society* (2015 May), convention paper 9307.

[19] K. Koivuniemi and N. Zacharov, “Unravelling the Perception of Spatial Sound Reproduction: Language Development, Verbal Protocol Analysis and Listener Training,” presented at the *111th Convention of the Audio Engineering Society* (2001 Nov.), convention paper 5424.

[20] Genelec, “8020A Data Sheet,” http://www.genelec.com/documents/datasheets/DS8020a.pdf, accessed 06/01/2015 (2015).

[21] AES, “Multichannel Surround Sound Systems and Operations,” Tech. Rep. AESTD1001.1.01-10, AES Technical Council (2001).

[22] 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.

[23] J. Francombe, T. Brookes, R. Mason, and J. Woodcock, “Evaluation of Spatial Audio Reproduction Methods (Part 2): Analysis of Listener Preference,” *J. Audio. Eng. Soc.* vol. 65, pp. 212–225 (2017 Mar.).

[24] Wordle, “Wordle – Beautiful Word Clouds,” http://www.wordle.net/, accessed 27/05/2015 (2015).

[25] S. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, *Text Mining: Predictive Methods for Analyzing Unstructured Information* (Springer, New York, USA, 2005), pp. 1–116.

[26] G. Salton, *The SMART Retrieval System: Experiments in Automatic Document Processing* (Prentice-Hall, New Jersey, USA, 1971).

THE AUTHORS

Jon Francombe

Tim Brookes

Russell Mason

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.