A subjective comparison of discrete surround sound and soundbar technology by using mixed methods

T. Walton, M. Evans, F. Melchior & D. Kirk

BRITISH BROADCASTING CORPORATION
Abstract

In recent years, soundbars have seen a rise in interest from consumers of home audio. Such technology offers an alternative means to experience surround sound content compared to conventional discrete multichannel systems. This paper presents a subjective comparison between two soundbars, a discrete 5 channel surround system and a discrete stereo system for a range of content and listener experience in order to evaluate how soundbar technology compares to conventional discrete systems. A mixed methods approach, Open Profiling of Quality, was used in order to deeper understand preference ratings for the various reproduction systems. The results show that the discrete surround system was significantly preferred to the soundbars for all content due to a combination of timbral and spatial factors.

This document was originally published in the proceedings of the Audio Engineering Society 140th Convention, Paris, 2016.

http://www.aes.org/e-lib/browse.cfm?elib=18290

Additional key words: audio, attribute, elicitation, listening, home, qualitative, immersive
A Subjective Comparison of Discrete Surround Sound and Soundbar Technology by Using Mixed Methods

Tim Walton\textsuperscript{1,2}, Michael Evans\textsuperscript{2}, David Kirk\textsuperscript{1}, and Frank Melchior\textsuperscript{2}

\textsuperscript{1}Open Lab, Newcastle University, Newcastle Upon Tyne, United Kingdom
\textsuperscript{2}BBC Research and Development, Salford, United Kingdom

Correspondence should be addressed to Tim Walton (tim.walton@bbc.co.uk)

\textbf{ABSTRACT}

In recent years, soundbars have seen a rise in interest from consumers of home audio. Such technology offers an alternative means to experience surround sound content compared to conventional discrete multichannel systems. This paper presents a subjective comparison between two soundbars, a discrete 5 channel surround system and a discrete stereo system for a range of content and listener experience in order to evaluate how soundbar technology compares to conventional discrete systems. A mixed methods approach, Open Profiling of Quality, was used in order to deeper understand preference ratings for the various reproduction systems. The results show that the discrete surround system was significantly preferred to the soundbars for all content due to a combination of timbral and spatial factors.

\section{Introduction}

Surround sound content has been available to users in home environments since the early 1990s. The technology was initially domesticated with the concept of home cinema in mind \cite{1} - users had become familiar with surround sound in the cinema and welcomed the option of improving the home cinema experience with the added immersion of surround sound. Whereas surround content is still primarily related to film, over the years the range of content has increased so that today various television programmes and music recordings are also produced in the surround sound format.

Traditionally the default setup for surround sound reproduction is 5.1 (also called 3-2 format). In addition to a stereo pair at ±30°, this setup comprises a centre channel at 0°, two surround channels generally at ±110° and an optional low-frequency effects (LFE) channel (the .1 in 5.1) \cite{2}. However, this speaker layout is often inconvenient for the user resulting in non-optimal speaker positions and different types of speakers being used for the various channels \cite{3}.

A possible alternative approach to achieve surround sound is by means of soundbar technology. Such technology is generally advertised as being able to deliver a good spatial impression from a single enclosure of speakers. In recent years, soundbars have seen a large rise in interest \cite{4} as such devices are possibly viewed as being simpler to setup, more convenient and more aesthetically pleasing than traditional discrete surround setups. The definition of a soundbar in terms of products on the market is somewhat vague. The majority of low to mid-range soundbars are only capable of 2 channel stereo reproduction. At the higher end of the market, however, soundbars are advertised as being capable of delivering surround content \cite{5} and are thus a possible alternative to traditional discrete surround setups. This surround effect can be achieved in several ways. One method is to use beamforming with an array of speakers \cite{6} - the aim being to reflect sound around the room so that it appears as if there are speakers positioned behind the listener. Another method is to take advantage of psychoacoustic phenomena to create a virtual surround effect.

To the author’s knowledge, to date there has only been limited published work on the subjective comparison of discrete surround systems and soundbars. Moulin et al. \cite{7} compared a discrete surround system, headphones and a soundbar in the context of an audiovisual scenario. The soundbar was rated as having a lesser degree of sound spatialisation than the two other systems although the test took place in an acoustically treated room - possibly influencing the performance of the soundbar, as discussed further on in this paper. Additionally, the term ‘degree of sound spatialisation’ could be seen to encompass various aspects of the sound, limiting the scope of the results.

The aim of the study presented here is to assess the effectiveness of soundbar technology by subjectively comparing a discrete surround system, a discrete stereo
An advantage of such methods is that, compared to various other elicited construct methods, participants may be able to better relate to the attributes being used [12]. Within elicited construct methods, a distinction can be made between consensus vocabulary (CV) techniques and individual vocabulary (IV) techniques. CV techniques aim to define a consensually agreed attribute list for the entire panel to use. With IV techniques each participant develops and employs their own attribute list and the data is subsequently analysed with multivariate statistical techniques. The differences between CV and IV techniques are discussed in greater detail in [13].

Lorho [14] used an individual vocabulary profiling technique inspired by Free-Choice Profiling (FCP) and the Flash Profile technique to study spatial enhancement systems for stereo headphone reproduction. The method used was similar to OPQ however did not involve a preference rating section. The Repertory Grid Technique (RGT) [15] [16] [17] is another elicited construct method whereby participants elicit and rate their own individual set of attributes. RGT by itself, however, does not aim to correlate overall preference ratings to individual elicited attributes. Preference mapping has previously been used for spatial audio assessment, such as in the Audio Descriptive Analysis and Mapping (ADAM) technique [18]. Unlike the previous two methods mentioned, this method employed consensus vocabulary profiling.

In addition to evaluating the effectiveness of soundbar technology, this study aims to introduce OPQ to the field of spatial audio evaluation so that it may be used as an alternative to the methods mentioned above.

2 Experimental Design

2.1 Reproduction systems

A variety of technology exists when it comes to audio reproduction via soundbars; the exact details of which are often not publicised by manufacturers or discussed in academic literature. However, it is possible to make several distinctions between the various technologies available. The first is a distinction between stereo soundbars, which downmix additional channels to stereo, and soundbars which can process surround channels individually. It is the latter of these that is of interest in this study. Another distinction is made between the methods that soundbars use to reproduce the spatial characteristics of audio content. It appears that one group of soundbars focuses on beamforming technology, by which sound is reflected off the walls of the room in order to replicate discrete speaker locations, as discussed in [6]. Another group focuses on psychoacoustic filtering to create sound zones around the listener. It is with this in mind that two soundbars were chosen for this study. Soundbar 1, a Focal Dimension, predominantly uses filtering to achieve a surround sound experience [5]. Soundbar 2 on the other hand, a Yamaha YSP-4300 ‘digital sound projector’, predominantly uses beamforming [19]. To assess the effectiveness of soundbar technology, these two soundbars were compared to a discrete surround system and also a discrete stereo system.

2.2 Stimuli

Six audio excerpts in the 5.0 channel format were used for the main sections of this experiment with an additional two being used for a familiarisation stage. These...
spanned a range of genres (ambient sound, pop music, classical music, radio drama, documentary and film) in order to cover the range of material likely to be experienced by users of the systems under study. The excerpts chosen contained both foreground-foreground (F-F) characteristics and foreground-background (F-B) characteristics [20] so that the excerpts had varying degrees of spatial information. A basic measure of how much surround information each stimuli contained was quantified with two ratios: a root mean square (RMS) frontal-surround ratio and a peak frontal-surround ratio. The RMS frontal-surround ratio was calculated by dividing the sum of RMS values for the L, R and C channels with the sum of RMS values for the LS and RS channels. The peak frontal-surround ratio was calculated in the same way but using peak values. The higher the ratio, the less surround information a sample has. Details of the stimuli are presented in Table 1.

The excerpts ranged in duration from 10 to 15 seconds and included an added 1.5 seconds of silence at the beginning of each clip. This silence ensured that the reproduction systems were not identifiable by their fade-in characteristics once they started to receive a signal. For each of the stimuli, stereo downmixes were created in accordance with ITU-R BS.775 [2] for playback over the discrete stereo system. All stimuli were initially aligned to -23 LUFS [21] although were altered in the calibration procedure described in section 2.5.3. In order for the 5.0 material to be used with the two soundbars and to be transmitted via digital optical cables, it was necessary to encode the stimuli. The stimuli were encoded to DTS digital surround with a SurCode DTS software encoder. The encoded 5.0 DTS WAV files had a sample rate of 44.1 kHz and a data rate of 1.234 Mb/s. All 4 reproduction systems received encoded files that were encoded with the same settings.

2.3 Participants

A total of 18 participants (age range: 21–42, mean 28, gender: 13 male, 5 female) participated in this study. All participants were fluent in English and self-reported normal hearing. A distinction between experienced and naïve assessors was made. Experienced assessors were professionals or academics in the fields of audio, acoustics or music and had previously participated in at least one other critical listening test. Naïve assessors were those who did not meet this requirement.

2.4 Experimental method

The experimental method employed in this study was a variation of OPQ, as presented by Strohmeier [9]. This is a mixed methods approach whereby both quantitative and qualitative data are collected. An overview of the steps involved in this method and how they were arranged in this particular study is presented in Figure 1, with each of the stages being subsequently described. The procedure was divided into two sessions for each participant. The two sessions were completed on separate days and within 5 days.

2.4.1 Attribute elicitation introduction

After reading an information sheet and filling in a short demographic survey, the first task presented to the participants was an introductory verbal exercise on attribute elicitation. The purpose of this task was to familiarise participants with the attribute elicitation process and the kind of words that may be used to describe sensory differences and similarities between two objects. The question posed to the participants was “Imagine a basket full of apples. What kind of attributes, properties or factors can you use to describe similarities and differences of two randomly picked?” The researcher could help the participants find attributes but never suggested specific examples. It was decided not to use an introduction exercise related to audio so as not to bias the future auditory elicitation task.

2.4.2 Familiarisation

A familiarisation page was then presented to the participants. On this page were 2 rows of 4 excerpts - A, B, C and D (corresponding to the 4 reproduction methods) for two different familiarisation stimuli. The participants were instructed to listen through the 2 series of 4 excerpts thinking about the differences between them, both in terms of timbral and spatial aspects. For naïve listeners the distinction between timbral and spatial aspects was explained and discussed. Participants were also encouraged to think about which clip they preferred and why. It was stated to the participants that the degree of difference between the audio clips presented here was representative of the rest of the experiment.

2.4.3 Preference rating & attribute elicitation

The preference rating and attribute elicitation stages were undertaken simultaneously. This differs from traditional OPQ where these are normally completed in
Table 1: Overview of content items used including spatial categorisation and frontal-surround ratios. Starred items were used in the familiarisation stage only.

<table>
<thead>
<tr>
<th>Item Genre</th>
<th>Title</th>
<th>Categorisation</th>
<th>Frontal-surround Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documentary</td>
<td>Africa</td>
<td>F-B</td>
<td>6.6</td>
</tr>
<tr>
<td>Ambient sound</td>
<td>Applause from Last Night of the Proms</td>
<td>F-F</td>
<td>1.0</td>
</tr>
<tr>
<td>Pop music</td>
<td>Lady Gaga - Pokerface</td>
<td>F-F</td>
<td>3.1</td>
</tr>
<tr>
<td>Radio drama</td>
<td>The Hitchhikers Guide to the Galaxy</td>
<td>F-F</td>
<td>1.3</td>
</tr>
<tr>
<td>Classical music</td>
<td>Last Night of the Proms</td>
<td>F-B</td>
<td>2.1</td>
</tr>
<tr>
<td>Film</td>
<td>Tropic Thunder</td>
<td>F-F</td>
<td>7.1</td>
</tr>
<tr>
<td>Ambient sound*</td>
<td>Thunderstorm</td>
<td>F-F</td>
<td>1.1</td>
</tr>
<tr>
<td>Pop music*</td>
<td>The Doobie Brothers - Long Train Runnin'</td>
<td>F-F</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Fig. 1: Overview of the method employed for this study. Session 1 had a total duration of 90-120 minutes and session 2 had a total duration of 60-90 minutes.
each attribute meant.

2.4.5 Attribute rating

As this stage took place on a different day, the familiarisation stage was repeated prior to the attribute rating. Participants were handed a list of their developed attributes and were asked to think about the differences between the familiarisation clips with respect to the listed attributes.

The aim of the attribute rating stage is to quantify the strength of the developed attributes for each stimuli. As with the preference ratings, a full paired comparison method was used to achieve this. Each participant used an individualised interface with rating scales corresponding to the developed attributes from the previous stage. They were instructed to rate which clip, A or B, had more of the listed attributes on a rating scale ranging from ‘A has much more’ to ‘B has much more’. Again, 36 comparisons were made with a 10-15 minute break half way through.

It should be noted that the attribute elicitation and rating techniques employed in this study differ from conventional OPQ. In [9], a single stimulus FCP method was used to elicit and rate attributes. The paired method employed in this study, however, is more akin to pairwise RGT, as used in [22]. It was decided to use a paired approach as various studies have suggested that structured methods are easier for inexperienced assessors, e.g. [23].

2.5 Experimental setup

2.5.1 Room

Due to the nature of the technology used by the soundbars, an ITU-R BS.1116 [24] standardised listening room was deemed unsuitable. Such a room has minimal side wall reflections so could possibly prevent lateral reflections produced by the soundbars. Instead, a user testing lab was used which was initially designed to represent a typical living space. It’s dimensions on the horizontal plane can be seen in Figure 2, with a height of 2.9 m. The floor of the room was carpeted, the front and rear walls (from the perspective of the participant) were made of glass, and the side walls were typical plaster walls, one of which had a large mirror. A heavy curtain was drawn across the front wall. Acoustic measurements of the room were carried out although full results of these are not presented in this paper. The reverberation time ($T_{30}$) measured in octave bands was below 0.5 s for all frequencies, and between 0.2 and 0.3 s for frequencies 500 Hz to 8 kHz.

2.5.2 Setup

The laptop used to administer the listening test software was connected to a Roland UA-25EX interface running at a sample rate of 44.1 kHz and a bit depth of 16-bit. The digital audio signal from the interface was routed to a Friend-chip DMX-12 optical matrix via an optical Toslink cable. This optical matrix enabled switching between the various reproduction systems and was controlled via a MIDI signal output from the listening test software and a second UA-25EX interface. The optical matrix was connected to the various reproduction systems via optical Toslink cables. For the discrete surround and stereo systems, a Denon DN-A7100 AV receiver was used to decode the audio files and route the channels to the corresponding speakers.

A schematic of the setup can be seen in Figure 2. The discrete surround system was composed of 5 PMC DB1S-A speakers and a PMC TLE1 subwoofer. It was setup with a radius of 2.5 m and with angles specified in ITU-R BS.775 [2] (i.e. $0^\circ$, $\pm 30^\circ$, $\pm 110^\circ$). To minimise elevation differences between the various systems, the speakers were placed upside down with the tweeters below the mid-range drivers. The tweeter height was 1.07 m which corresponded to ear height. The 2 soundbars, the centre speaker for the discrete system and the sub for soundbar 1 were all placed on a desk with dimensions 1.5 x 0.6 x 0.85 m (w x d x h). The soundbars and discrete centre speaker were stacked with soundbar 1 at 0.92 m from the floor, soundbar 2 at 1.01 m from the floor and the tweeter from the discrete centre channel at 1.07 m from the floor. There was therefore an elevation difference between the systems of 15 cm ($\sim 3^\circ$).
relative to the listener). All systems were 2.5 m from the listening position and an acoustically transparent curtain was used to prevent visual bias.

2.5.3 Calibration

The speakers in the discrete system were level aligned relative to each other (to within 0.5 dB) using pink noise measured at the listening position. The subwoofer for the discrete system had a cutoff frequency of 85 Hz and was level aligned using pink noise and a 1/3-octave real time analyser to achieve a flat frequency response. The bass management was done via the AV receiver. A comfortable listening level was used for the absolute level of the discrete system (and therefore the other systems), as subjectively decided by the researcher and verified by participants in a pilot study. This corresponded to a mean stimuli level of 64.8 dBA. Soundbar 1 included several manual options to calibrate the soundbar for the room of use. These were set as - distance: 2 (medium), position: 6 (free standing), room: 8 (medium), subwoofer: 11 (dimension). The level of the subwoofer was adjusted in the same manner as with the discrete system. Soundbar 2 was automatically calibrated using the inbuilt calibration function. Inter-system level alignment was carried out subjectively by the researcher. A subjective approach was chosen so as to take into consideration any psychoacoustic filtering employed by the soundbars. This was checked for every stimulus and gain was applied to the pre-encoded source files accordingly.

2.5.4 Test administration

The test was administered via a laptop running a custom designed listening test patch on Max MSP software. For both rating sessions each page contained two excerpts: ‘A’ and ‘B’. The participants could play these as many times as they liked. It was possible to switch between the excerpts, although when a new excerpt was selected, it would always play from the beginning. The order of the comparisons was randomised, as was the assignment of stimuli to either button ‘A’ or ‘B’. The rating scale used for both sessions was based on a Comparison Category Rating scale [25] and had a range of values from -35 to +35 in interval increments. There were labelled anchors at -30 (‘Much prefer A’), 0 (‘No preference’) and +30 (‘Much prefer B’) and unlabelled anchors at ±10 and ±20. The labelled anchors were offset from the end so as to reduce end-of-scale effects [26]. The scale had to be clicked before the participant could move on to the next page. Below the rating scale was a text box in which the participants could type any differences between the stimuli they heard. They were encouraged to list words or short phrases, not full sentences in order to reduce the test duration. The attribute rating interface had scales corresponding to the developed attributes of each participant. They were asked to ‘compare clips ‘A’ and ‘B’ in terms of which has more of the following attributes’. The labelled anchors were ‘A has much more’ and ‘B has more’. One rating scale had to be clicked before moving on to the next page.

3 Results

3.1 Participant reliability

In paired comparison tests, it is possible to assess intra-participant reliability by using circular error rates [27]. A circular error occurs when a participant makes an inconsistent judgment on a triad of stimuli. For example, a circular error would occur if a participant preferred stimulus A to B, preferred stimulus B to C, but preferred stimulus C to A. Such errors indicate that either the participant was not paying attention, that they altered their assessment criteria as the test progressed, or that they found the test challenging resulting in inconsistent judgments. By comparing the number of circular errors associated with each participant and the maximum possible number of circular errors, a circular error rate in percent can be calculated. Figure 3 shows the circular error rates from the preference rating session for all participants. An error tolerance threshold of $A = 0.06$ was used for the calculations. If $A = 0$, no inverted preference can be accepted, whereas if $A > 0$, only preference inversions greater than the value of A are counted as circular errors. As a continuous slider
was used in this study, a value of $A = 0.06$ was chosen (for scores scaled to ±1) so as to discount slider inaccuracies of participants who intended to show no preference. It is seen that the majority of participants could make reliable judgments with 8 participants making no errors and 14 participants having an error rate of less than 10%. Participants 7, 10, 11 and 12 (all naïve) show an error rate above 15%, which is higher than the average of 6.5%. To ensure reliability of results, it was decided to exclude the results from these participants for the subsequent analysis. An error limit of 10% was also used in [28] to ensure consistent results from paired comparison ratings.

### 3.2 Preference ratings

The preference ratings for each paired comparison were first scaled to lie in the range of ±1, where -1 corresponds to full preference for stimulus ‘A’ and +1 corresponds to full preference for stimulus ‘B’. If $P_{ij}$ is the preference probability of stimulus $i$ versus stimulus $j$, it is assumed that

$$P_{ij} = -P_{ji}. \quad (1)$$

That is, a negative probability of preference $P_{ij}$ means stimulus $j$ is preferred to stimulus $i$. From these preference probabilities, preference scores can be calculated with

$$S_i = \sum_{j \neq i} P_{ij}, \quad (2)$$

where $S_i$ is the preference score for stimulus $i$. As 4 reproduction systems were used, possible values of $S_i$ lie in the range of ±3. These preference scores were scaled to lie in the range of ±1 so that +1 corresponds to full preference towards a reproduction system. In total, 24 preference scores were calculated (4 systems by 6 content items).

After the exclusion of 4 participants, data from 14 participants remained (7 experienced, 7 naïve) who rated 4 reproduction systems for 6 content items. To investigate the effect and interaction of the independent variables, a three-way mixed ANOVA was carried out on this data. The between-subject factor was participant experience (2 levels) and the within-subject factors were system (4 levels) and content (6 levels). Prior to this, the main assumptions underlying the ANOVA were checked for each group of participants. Normality was checked using a Shapiro-Wilk test for the 24 combined within-subject factors for both groups. The data was not significantly different from normal ($p > 0.5$) for 22 out of the 24 factors for both the experienced and naïve groups. The data is therefore said to have a predominantly normal distribution. Homogeneity of variance was checked via Levene’s test and the data was found to have equal variance ($p > 0.5$) for both within-subject and between-subject factors. Finally, Mauchly’s test of sphericity was conducted which gave non-significant ($p > 0.5$) results.

The between-subject factor of participant experience was found to have a non-significant influence on the preference scores [$F(1, 12) = 2.68, p = .127]$. However, a partial eta-squared value of $\eta_p^2 = .183$ indicates that participant experience does have an effect on the preference scores. This suggests that if a larger study was conducted with more participants per group, the participant experience could be a significant factor.

An overview of the within-subject factor results from the mixed ANOVA model in order of decreasing significance.

<table>
<thead>
<tr>
<th>Source</th>
<th>$df$</th>
<th>$F$</th>
<th>$\eta_p^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
<td>Hyp</td>
<td>3</td>
<td>84.323</td>
<td>.875</td>
</tr>
<tr>
<td></td>
<td>Err</td>
<td>36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content*System</td>
<td>Hyp</td>
<td>15</td>
<td>5.557</td>
<td>.317</td>
</tr>
<tr>
<td></td>
<td>Err</td>
<td>180</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content<em>System</em>Experience</td>
<td>Hyp</td>
<td>15</td>
<td>1.681</td>
<td>.123</td>
</tr>
<tr>
<td></td>
<td>Err</td>
<td>180</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content</td>
<td>Hyp</td>
<td>5</td>
<td>2.101</td>
<td>.149</td>
</tr>
<tr>
<td></td>
<td>Err</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>System*Experience</td>
<td>Hyp</td>
<td>3</td>
<td>2.403</td>
<td>.167</td>
</tr>
<tr>
<td></td>
<td>Err</td>
<td>36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content*Experience</td>
<td>Hyp</td>
<td>5</td>
<td>0.303</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td>Err</td>
<td>60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Within-subject factor results from the mixed ANOVA model in order of decreasing significance.
were calculated. Figure 4 shows the marginal means for both the systems averaged over content and for each content individually. These preference scores, which can have values in the range of ±1, are averaged over listener. Additionally, matrices that show which pairwise comparisons are statistically significant are presented. For the content average plot, all systems have corresponding marginal means significantly different from each other; the discrete surround system is the most preferred followed by the stereo downmix. As an interaction was found between the reproduction systems and content, it is necessary to examine the marginal means for each system with respect to content. It is seen that the discrete surround system is significantly preferred over the stereo downmix for the items Ambience, Pop Music, Film and Radio Drama. These items are all F-F items whereas the items which do not show a significant difference are F-B items. The discrete surround system is significantly preferred to both soundbars for all content. Likewise, the stereo downmix is significantly preferred to soundbar 1 for all content. For comparisons between the stereo downmix and soundbar 2, only items Pop Music and Radio Drama show significant differences - the stereo downmix being preferred in both cases. The only content which displays a significant difference between the two soundbars is the item Ambience for which soundbar 2 is preferred over soundbar 1. It should be noted that the greatest difference in preference scores is seen for the content item with the lowest frontal-surround RMS ratio (content Ambience), i.e. the content item with the most surround information.

3.3 Sensory profiling

In the attribute elicitation and refinement stage the 14 participants whose preference ratings were analysed in the previous section produced a total of 102 refined attributes (max 8, min 5). The raw paired attribute ratings were converted to attribute scores using the same method as with the preference scores. For each participant this resulted in an $M \times N$ matrix (or configuration) of attribute ratings, where $M$ is the number of test items (24 in this case) and $N$ is the number of individual attributes. The individual participant matrices were concatenated to form a complete attribute matrix of 24 items x 102 attributes. The following analysis was run on this dataset and implemented in XLSTAT.

The first stage of the analysis was to run Generalised Procrustes Analysis (GPA) on the dataset of attribute ratings. The aim of GPA is to reduce scale effects and to obtain a consensus configuration. This is achieved by rotating and transforming the configurations by minimising the residual distance between the configurations and their consensus. The second part of the analysis was to run Principal Component Analysis (PCA) on the dataset from the GPA procedure.

A total of 8 components were needed to explain 100% variance in the GPA model. The cumulative variance of these 8 components are (in %) 75.7, 88.0, 91.7, 94.4, 96.5, 98.3, 99.3 and 100 respectively. The first two components describe the majority of the variance, 88%, so were used to form the perceptual space.

Figure 5 shows all the rated attributes in the perceptual space of PC1 and PC2. The further the attributes are from the centre, the greater their associated explained variance. The inner and outer circles represent 50% and 100% explained variance respectively. Several clusters of attributes can be identified and are labelled with relevant descriptions. It is seen that PC1 (75.71% explained variance) is positively loaded with attributes related to width (‘width’, ‘wide’, ‘width of frontal image’, ‘width of image’) and negatively loaded with the attribute ‘focused’. A cluster related to envelopment is found in very close proximity to the width cluster (‘envelopment’, ‘enveloping’) and a cluster related to immersion is also found close by (‘immersion’, ‘immersive’, ‘localisation of audio elements’, ‘changing distance and movement’, ‘movement’). The grouping of these attributes suggests that envelopment terms were used in a similar way as width terms, however, immersion terms were used slightly differently with attributes related to movement and localisation of audio elements playing a role. This distinction between immersion and envelopment was also commented on by several of the participants during the attribute elicitation stage. Clusters related to timbral aspects are found on the diagonals of the perceptual space. Negative timbral aspects (‘bandlimited’, ‘tinny’, ‘colouration’, ‘distortion’) are found at negative PC1 and PC2 values whereas positive timbral factors (‘rich’, ‘clarity’, ‘depth’, ‘balanced’) are found at positive PC1 and PC2 values. The fact that spatial and timbral aspects of the same polarity are located on the same sides of the perceptual space suggests that stimuli related to positive spatial attributes are also related to positive timbral attributes and vice versa. When comparing the attributes generated by experienced listeners (55) to those generated by naïve listeners (47), it is seen that similar terms are used, although the attributes from experienced listeners generally describe more variance in the data. This suggests...
that experienced listeners produce stronger attributes and are more confident in their use.

Figure 6 shows all the stimuli in the perceptual space of PC1 and PC2 with participants’ preferences mapped through external preference mapping. The attribute cluster labels identified from Figure 5 are also shown. It is seen that items are generally clustered by the reproduction system. The discrete surround system stimuli are located in the area of the perceptual space related to width, envelopment and immersion. Stimuli such as ‘Classical_ds’ and ‘Ambient_ds’ are perceived as wide and enveloping whereas ‘Radio_ds’ is the most immersive stimulus. At negative PC1 values soundbar 1 stimuli are found. These stimuli are also the most related to negative timbral factors. This suggests that soundbar 1 was not perceived as immersive or enveloping and had negative timbral characteristics. In particular, ‘Ambient_s1’ is seen to be the most focussed and worst rated in terms of timbral aspects. This agrees with discussions from several participants. Soundbar 2 stimuli are also found at negative PC1 values suggesting that they are also not as wide, enveloping or immersive as the discrete surround system stimuli. ‘Ambient_s2’, however, is found at a similar PC1 value to ‘Ambient_dm’ which suggests that they are perceived as having similar widths. The fact that soundbar 2 stimuli have higher PC2 values suggests that they are less related to negative timbral factors as soundbar 1 stimuli. In terms of spatial factors, soundbar 2 is seen to be perceived as slightly wider than soundbar 1. Finally, the stereo downmix stimuli are perceived as less wide, enveloping and immersive than the discrete surround stimuli.

The participants’ preferences mapped onto Figure 6 show that the majority prefer wide, enveloping and immersive stimuli (i.e. the discrete surround system stimuli). Participant 15, however, appears to have based their preference judgments more on timbral aspects than spatial aspects. This therefore means that their preference towards the discrete surround stimuli are less pronounced than with the other participants.

As the spatial and timbral factors shown in the perceptual space are not orthogonal, it is hard to separate their influence on the preference ratings of the various reproduction systems. It appears that the discrete surround system is preferred over the stereo down mix due to spatial aspects, and this is expected as both systems use the same speakers. The two soundbars are rated negatively compared to the other two systems due to a combination of both timbral and spatial factors. In terms of spatial factors, soundbar 2 is perceived as slightly wider than soundbar 1, although the only significantly different preference rating between these two systems is for the content ‘Ambience’.

4 Conclusions

A subjective comparison of a discrete 5 channel surround system, a discrete stereo system and two soundbars was made by means of a mixed methods approach.
Walton et al.  
Comparison of Discrete Surround Sound and Soundbar Technology

**Fig. 5:** PCA correlation loadings with attributes in the space of PC1 and PC2. The inner and outer circles represent 50% and 100% explained variance respectively.

**Fig. 6:** Objects and participants’ preferences in the preference map of PC1 and PC2. ‘ds’ = discrete surround, ‘dm’ = stereo downmix, ‘s1/2’ = soundbar 1/2.
When averaged over content, preference ratings for the two soundbars were significantly lower than the discrete surround system and the discrete stereo system. Additionally, a significant difference in preference ratings between the two soundbars was observed; the soundbar which employs beamforming to achieve a surround effect was significantly preferred to the soundbar which predominantly uses filtering. With respect to content, the greatest difference between the systems was observed for the content item with the lowest frontal-surround ratio, i.e. the content item with the most surround information. This, along with the qualitative analysis, suggests that the two soundbars under study did not effectively replicate discrete surround channels.

Qualitative analysis showed that the given preference ratings were due to a combination of both timbral and spatial factors. Participants preferences were mapped to wide, enveloping and immersive items which correlated to the discrete surround system. It was not possible to fully separate the influence of spatial and timbral factors on the preference ratings as these clusters were not orthogonal on the perceptual space. This suggests that the systems with positive spatial attributes also had positive timbral attributes and vice versa.

The modified OPQ method employed for this study was shown to be an effective method for investigating subjective differences between reproduction systems and could therefore be used for further studies on spatial audio perception. Participants with a range of listening experience were able to give preference ratings, develop and refine individual attributes and rate stimuli on these attributes over two sessions. The qualitative data collected allowed an in-depth analysis of the reasoning behind the preference ratings.

Care must be taken when generalising the above results to the two types of technology as a whole. For a more comprehensive comparison a greater number of discrete systems and soundbars should be investigated to see if results are consistent with those found in this study. Additionally, it is thought that the room of use could have a significant influence on the effectiveness of soundbar technology. This should also be investigated in further studies.

Acknowledgements

The first author would like to thank Thomas Nixon for the useful discussions on the setup and for his assistance with the room acoustic measurements. This work was supported by an ESPRC industrial CASE studentship in partnership with the BBC.

References

[12] Berg, J., “How do we determine the attribute scales and questions that we should ask of sub-


