Eliciting the most prominent perceived differences between microphones

Andy Pearce, a) Tim Brookes, Martin Dewhirst, and Russell Mason
Institute of Sound Recording, University of Surrey, Guildford, Surrey, GU2 7XH, United Kingdom

(Received 23 October 2015; revised 30 March 2016; accepted 30 April 2016; published online 25 May 2016)

The attributes contributing to the differences perceived between microphones (when auditioning recordings made with those microphones) are not clear from previous research. Consideration of technical specifications and expert opinions indicated that recording five programme items with eight studio and two microelectromechanical system microphones could allow determination of the attributes related to the most prominent inter-microphone differences. Pairwise listening comparisons between the resulting 50 recordings, followed by multi-dimensional scaling analysis, revealed up to 5 salient dimensions per programme item; 17 corresponding pairs of recordings were selected exemplifying the differences across those dimensions. Direct elicitation and panel discussions on the 17 pairs identified a hierarchy of 40 perceptual attributes. An attribute contribution experiment on the 31 lowest-level attributes in the hierarchy allowed them to be ordered by degree of contribution and showed brightness, harshness, and clarity to always contribute highly to perceived inter-microphone differences. This work enables the future development of objective models to predict these important attributes.

© 2016 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

http://dx.doi.org/10.1121/1.4950820

I. INTRODUCTION

To describe the sonic characteristics of a microphone, manufacturers can supply several standardised measurements that describe its objective performance (BS EN 60268-4, 2010). It has been noted by Olive and Toole (1989), Hebrock et al. (1996), and others that these measurements do not always correlate well with perceived characteristics, and they have suggested other objective measures (Green and Statham, 1998; Hebrock et al., 1996, 1997; Olive and Toole, 1989). However, even these suggested measures do not directly correlate with specific subjective attributes. Identifying the perceptual attributes that vary between microphones, and the extent to which these attributes vary, would be a step toward more perceptually meaningful microphone comparisons.

Relevant perceptual descriptors and attributes can be identified by using elicitation experiments, and these have been widely conducted for loudspeakers, musical acoustics, concert hall acoustics, and multi-channel audio systems, e.g., Disley et al. (2006); Francombe et al. (2014); Gabrielson (1979); Gabrielson and Sjögren (1979); Koivuniemi and Zacharov (2001); Lavandier et al. (2008); Lokki et al. (2011, 2012).

Gabrielson and Sjögren (1979) elicited 55 descriptors of the differences between loudspeakers, and analysed these descriptors to find the 8 most prominent corresponding attributes: clearness/distinctness, sharpness/hard-softness, brightness/darkness, fullness-thinness, feeling of space, nearness; disturbing sounds, and loudness. More recent work by Lavandier et al. (2008) found two salient dimensions relating to the perceived differences between loudspeakers from a multi-dimensional scaling (MDS) analysis. These were labeled: bass/treble balance and medium emergence. This research was followed by a study by Michaud et al. (2015) on a larger set of loudspeakers, finding three salient dimensions: bass/treble balance, medium emergence, and feeling of space. In the work of Koivuniemi and Zacharov (2001), which was focused on spatial sound systems, four timbral attributes were found: richness, hardness, emphasis, and tone colour.

In the research of musical acoustics, Disley et al. (2006) performed experiments into musical timbre using 15 attributes: bright, clear, warm, thin, harsh, dull, nasal, metallic, wooden, rich, gentle, ringing, pure, percussive, and evolving. Pratt and Doak (1976) identified three scales to describe the timbre of musical instruments: dull/brilliant, cold/warm, and pure/rich.

Although there may be some overlap between the dimensions, descriptors, scale labels, and attributes identified in these previous studies and those that would be pertinent to the assessment of microphones, it cannot be concluded that they will all be relevant, nor that they will be sufficient. There has, however, been a relatively small amount of research specifically into the perceived differences between microphones.

Research by De Man and Reiss (2013) was focused on the methods used to obtain subjective data, rather than on identifying specific perceptual attributes, and found both multi-stimulus and pairwise-comparison approaches to be
suitable. Hebrock et al. (1996) asked listeners to describe the characteristics of three dynamic microphones and reduced the responses to the descriptors harsh, edgy, warm, (not enough, or no) low-end and (not enough, or extended) high-end. As part of a subsequent study into the perceptual effects of ringing in microphones, Hebrock et al. (1997) found that, of 13 suggested attributes, the 9 most frequently used by listeners were detailed, dull, muffled, open, thin, warm, harsh, nasal, and smooth. In the first study, only three microphones were considered, and all had similar specifications. In the second study, listeners characterised stimuli only in terms of attributes specified by the experimenters. It is therefore possible that other attributes might have been found to be important had a wider range of microphones been evaluated in the first study and had listeners not been limited in their responses in the second.

Research by McKinnie (2006) aimed to identify some of the perceptual attributes that differ between perceptually very similar microphones, but the results of the listening experiments indicated the responses were no better than chance. This may have been due to the magnitude of the differences between the stimuli being too small.

It is apparent, therefore, that although multiple attributes have been found that contribute to, or that might contribute to, the perceived differences between microphones, the relative degrees of their contributions are untested and there might be other attributes that contribute equally or even more. Hence, this study aims to determine the full set of attributes in terms of which microphones differ, to label these attributes appropriately, and to find the relative contributions of these attributes to perceived inter-microphone differences. The challenge of modeling these attributes in terms of objective physical metrics can then be addressed in future studies. In order that the study should not be limited to experimenter-proposed attributes, the approach taken employs elicitation experiments whereby listeners report freely on perceived differences between recordings made using alternative microphones. To avoid the problems encountered by McKinnie (2006), a wide range of microphones is selected, programme items are chosen that exhibit a wide range of differences, and a hybrid elicitation method is employed.

The experiment methods are further explained in Sec. II, and broken down into five distinct phases. The specific procedures and results for the five phases are presented in Secs. III–VII, and the results are discussed in Secs. VIII and IX.

II. EXPERIMENT METHODS

There are two approaches to elicitation experiments: direct and indirect (Bech and Zacharov, 2006). Direct elicitation methods involve asking participants to verbally describe the perceptual sensations evoked by stimuli, whereas indirect elicitation experiments require subjects to rate these sensations without explicit description.

A common indirect elicitation method is MDS in which participants rate the similarity of every pairwise combination of a set of stimuli, and an analysis is then conducted which attempts to position each stimulus in a multidimensional group space so that the pairwise distances match the pairwise similarity ratings. This has been used in several studies on audio codecs, tools, and products to find the number of salient dimensions across which a stimulus set varies (Gabrielsson, 1979; Hall, 2001; Neher et al., 2006). However, MDS analysis is only a data reduction tool, reducing the potentially large number of perceptual differences between stimuli into a smaller number of orthogonal dimensions (Hair et al., 2010). Each dimension does not necessarily correlate with a single perceptual attribute and dimensions are not identified as relating to particular attributes but are, instead, simply numbered.

One of the most common direct elicitation methods is free choice profiling (FCP), in which each participant develops his/her own set of words or phrases to describe the attributes that they can perceive to differ between stimuli. This is often followed by a panel discussion stage where common and similar terms are grouped, across all participants, to arrive at a single list of agreed descriptors that might each correspond to a particular attribute (Francombe et al., 2014; Zacharov and Koivuniemi, 2001). In the FCP stage, to ensure that all differences within a stimulus set are elicited, each stimulus must be compared directly with every other stimulus. This can be a difficult and time-consuming task for subjects and can lead to listener fatigue, potentially resulting in noisy data and/or missed attributes.

The hybrid method used in this study combined both approaches to make the elicitation task simpler and thereby increase the likely quality of the results. First, a similarity rating experiment and MDS analysis was conducted to identify stimulus pairs exhibiting large differences. This analysis was conducted similarly to the work of Neher et al. (2006), Hall (2001), and Williams (2010). These stimulus pairs (rather than the full stimulus set) were then used in an FCP experiment, similar to that conducted by Francombe et al. (2014). This was followed by a panel discussion to group the elicited terms and to agree on an attribute label to represent the terms in each group.

Previous studies have used the frequency with which a term has been elicited as an indicator of the importance of the attributes that it describes (Francombe et al., 2014). However, this approach has the potential to underestimate the importance of attributes which listeners hear clearly but find difficult to describe. In the current study, therefore, once attribute labels had been agreed, a novel attribute contribution experiment was conducted, asking subjects to explicitly rate the degree to which each of the attributes contribute to the overall difference between each stimulus pair.

The full study was conducted in five phases.

- Phase 1: Determined suitable microphones and programme items based on objective factors that are known to differ between microphones, in order to make recordings likely to be able to reveal the attributes comprising the most prominent inter-microphone differences. This is described in Sec. III.
- Phase 2: Employed pairwise similarity ratings and MDS to reveal the number of salient dimensions and to identify exemplary stimulus pairs. This is described in Sec. IV.
• Phase 3: Used an FCP approach to elicit terms from listeners that describe the differences between the exemplary stimuli. This is described in Sec. V.
• Phase 4: Used panel discussions to group the elicited terms to reduce redundancy and to identify and label the underlying perceptual attributes. This is described in Sec. VI.
• Phase 5: Determined the degree to which each agreed attribute contributes to perceived inter-microphone differences, by way of an attribute contribution experiment. This is described in Sec. VII.

III. PHASE 1: MICROPHONES AND SOURCES

There are over 1500 studio microphones in existence, with varying degrees of similarity and difference (Microphone Database, 2013). For this study, microphones were selected that were expected to exemplify the full range of potential inter-microphone differences.

A. Microphone selection

The selection of microphones was conducted in two parts. First, in order to select microphones based on objective parameters, examples were listed that exemplified the main differences in microphone design. Second, since some perceptual differences might not correlate with standard objective measures, recording engineers were asked to suggest additional microphones that they felt sounded significantly different from those listed.

1. Objective selection

BS EN 60268-4 (2010) describes manufacturer guidelines for the measurement and documentation of the transduction type, sensitivity, frequency response, directivity pattern, and self-noise of a microphone. In addition to these five factors, other research has suggested that the diaphragm size and transient response of a microphone may affect perceived sound quality (Ballou, 2009; Bartlett, 1987; Hebrock et al., 1997). It has also been suggested that the head-basket and, for a condenser microphone, the capsule termination type may be relevant (Combs, 2006; Joly, 2015).

From a list of commonly used studio microphones, microphones were selected that represented the extremes for each of these objective parameters. Thus, for each continuously-variable parameter, one microphone in the selected set was chosen due to it having a particularly high value of that parameter; one was chosen due to it having a particularly low value; and the other microphones will have intermediate values of that parameter (but each will represent an extreme for another parameter). For categorical parameters (e.g., transduction type) at least one microphone was chosen in each category.

For the parameters diaphragm size and frequency response, categories were selected. For diaphragm size, microphones were categorised as either large diaphragm (diameter ≥16 mm) or small diaphragm (diameter <16 mm). For frequency response, microphones were categorised as either flat or tailored, where tailored refers to any microphone whose on-axis frequency response includes a region exhibiting gain more than 3 dB greater than that at 1 kHz (Microphone Data, 2015). Since measurements of transient response are not included in BS EN 60268-4 (2010), transient response is estimated from transduction type and diaphragm size: a small-diaphragm condenser microphone is likely to have a fast transient response; a large-diaphragm dynamic microphone is likely to have a slow transient response (Ballou, 2009).

Application of these criteria resulted in the selection of the eight studio microphones shown in Table I.

2. Expert suggestions

A list of the eight selected microphones was presented to five experienced audio engineers. The engineers were asked if they felt that any perceptual characteristics that differ between microphones were not accounted for by the microphones on the list and, if they did, to suggest additional microphones to illustrate these characteristics. All five engineers responded that the list exemplified all relevant perceptual differences. To avoid any bias arising from knowledge of the microphones chosen, none of these engineers took part in subsequent phases of the study.

3. Additional non-studio microphones

It is unlikely that the audio engineers would have considered non-studio microphones. MEMS (microelectromechanical systems) microphones are used in a wide range of low-power devices, such as tablets and mobile telephones. They can be designed to have similar frequency response and self-noise to that of a typical studio microphone (Kardous and Shaw, 2014; Sessler, 1991), but the perceived quality of the recorded signal can be very different.

A pilot experiment was conducted with 12 commercially available MEMS microphones from 4 manufacturers. Ten subjects were asked to rate the basic audio quality, as defined in ITU-R BS.1116-1 (1994) and ITU-R BS.1534-1 (2003), of recordings made with the 12 microphones using a multiple stimulus comparison test interface. Recordings were made of pop music, classical music, and speech. The two MEMS microphones reported as having the highest and lowest quality as rated by all three programme items were added to the list for the current study:

• Wolfson WM7131 (Edinburgh, Scotland): MEMS microphone reported as having the highest quality in a pilot study.
• Knowles SPU0410HR5H (Itasca, IL): MEMS microphone reported as having the lowest quality in a pilot study.

B. Programme item selection

In previous work, it was found that musical sources were good at revealing the perceptual differences between microphones (better than vocal sources) (Pearce et al., 2015). Therefore only musical sources were considered for this study.

The sources were selected to have characteristics likely to reveal the objective inter-microphone differences listed in Sec. III A 1: double bass, drums, acoustic guitar, string
quartet, and trumpet, each played unaccompanied. The double bass was plucked, playing a jazz turn-around (duration 8 s) resulting in a stimulus with little high-frequency content and low sound pressure level (SPL). Drums consisted of a snare, hi-hats, and cymbals playing a simple rhythm (duration 7 s) across all pieces of the kit; this produced a non-harmonic frequency spectrum and high level of high-frequency energy. The acoustic guitar played continuous six-string strummed chords (duration 9 s) with a pick; this produced fast transients, a high level of high-frequency content, and a harmonic frequency spectrum. The string quartet played the first four bars (duration 9 s) of Vivaldi’s Summer (mvt. 1), producing a large dynamic range, broadband harmonic frequency spectrum, and slow transients. Finally the trumpet played a loud fanfare (duration 12 s), generating a high SPL at the microphone, and was intended to excite as many distortions in the microphones as possible.

C. Recording of the stimuli

Previous experiments (Pearce et al., 2015) have identified a suitable method for recording stimuli for intermicrophone perceptual comparisons: a multi-microphone array with a maximum inter-microphone spacing of no more than 150 mm, in an ITU-R BS 1116 compliant listening room (ITU-R BS.1116-1, 1994).

The five selected programme items were therefore recorded in this way, using the microphone arrangement shown in Fig. 1, to provide 50 stimuli for Phase 2. All microphones were recorded with a Presonus Digimax FS (Baton Rouge, LA) microphone preamplifier feeding an RME Fireface 800 (Haimhausen, Germany) audio interface. The MEMS microphones were supplied with 2.7 V power and recorded through the instrument inputs of the preamplifier due to their high output impedances. The input gain on the preamplifier was adjusted for each microphone to produce the same digital input level when excited with pink noise replayed through a Genelec 1032 (Iisalmi, Finland).

![FIG. 1. Layout of the recorded microphones. (1) AKG C12, (2) AKG C414 B-XLS, (3) AKG C451, (4) Coles 4038, (5) DPA 4006-TL, (6) Electrovoice RE20, (7) sE 2200a, (8) Shure SM58, (9) Wolfson WM7131, and (10) Knowles SPU0410HR5H.]

<table>
<thead>
<tr>
<th>Microphone</th>
<th>Directivity</th>
<th>Transducer</th>
<th>Diaphragm size</th>
<th>Frequency response</th>
<th>Self-noise (dBA)</th>
<th>Sensitivity (mV/Pa)</th>
<th>Diaphragm termination</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKG C12 (Vienna, Austria)</td>
<td>C</td>
<td>C</td>
<td>L</td>
<td>F</td>
<td>10</td>
<td>25</td>
<td>Edge-Terminated</td>
</tr>
<tr>
<td>AKG C414 B-XLS (Vienna, Austria)</td>
<td>C</td>
<td>C</td>
<td>L</td>
<td>F</td>
<td>10</td>
<td>25</td>
<td>Cylindrical</td>
</tr>
<tr>
<td>AKG C451 (Vienna, Austria)</td>
<td>B</td>
<td>D</td>
<td>S</td>
<td>T</td>
<td>10</td>
<td>25</td>
<td>Edge-Terminated</td>
</tr>
<tr>
<td>Coles 4038 (Hoddesdon, England)</td>
<td>B</td>
<td>R</td>
<td>L</td>
<td>F</td>
<td>10</td>
<td>25</td>
<td>Cylindrical</td>
</tr>
<tr>
<td>DPA 4006-TL (Alleroed, Denmark)</td>
<td>C</td>
<td>C</td>
<td>S</td>
<td>T</td>
<td>10</td>
<td>25</td>
<td>Centre-Terminated</td>
</tr>
<tr>
<td>Electrovoice RE20 (Bachman, MD)</td>
<td>C</td>
<td>C</td>
<td>L</td>
<td>F</td>
<td>10</td>
<td>25</td>
<td>Edge-Terminated</td>
</tr>
<tr>
<td>sE 2200a (Hitchin, England)</td>
<td>C</td>
<td>C</td>
<td>S</td>
<td>F</td>
<td>10</td>
<td>25</td>
<td>Centre-Terminated</td>
</tr>
<tr>
<td>Shure SM58 (Niles, IL)</td>
<td>C</td>
<td>D</td>
<td>L</td>
<td>F</td>
<td>10</td>
<td>25</td>
<td>Edge-Terminated</td>
</tr>
</tbody>
</table>

Small-diaphragm condenser microphones likely to have a fast transient response. Large-diaphragm dynamic microphones likely to have a slow transient response.
loudspeaker 1.5 m from the array, at a measured level of 74 dB\(_{\text{SPL}}\) at the array. Each source was positioned with its acoustic centre 1.5–2 m directly in front of the array.

It is acknowledged that placing a microphone in an array in close proximity to other microphones may alter its off-axis response; however, this study does not seek to determine the off-axis characteristics of a particular microphone but, rather, to compare on-axis characteristics across microphones.

**IV. PHASE 2: SIMILARITY RATINGS**

Using the recordings made in Phase 1, pairwise-comparison tests were conducted in order to find programme items and microphone pairs that exhibited differences across each salient dimension, for use in an FCP experiment.

Stimuli were presented diotically over a pair of Sennheiser HD650 headphones with a Focusrite VRM Box (High Wycombe, England) interface, with the VRM feature disabled. All stimuli were loudness-matched by a panel of five listeners, using a method-of-adjustment test, to a listening level judged to be comfortable.

**A. Similarity ratings**

To reduce the potential for listener fatigue, each of the five programme items was presented in a separate listening test. Prior to each test, subjects were presented with all ten stimuli, which could be auditioned at will, for the programme item under assessment in order to allow them to familiarise themselves with these stimuli. They were then presented with a smaller version of the test interface in order to allow them to familiarise themselves with the rating task. Nine listeners were asked to rate the similarity of each pair of stimuli on a 100 point scale with endpoints labeled as “most similar” and “least similar” taking into consideration only the range of similarities within the ten-stimulus set. All listeners were undergraduate students on the Music and Sound Recording course at the University of Surrey; all had participated in multiple listening tests previously, and all had passed a taught module in technical listening.

Each of the five tests used a graphical user interface which comprised one page of recordings of a single programme item. Each test contained all 45 pair-wise comparisons, and the listener moved a slider to indicate the perceived similarity between the two stimuli in each pair. To ensure that each pair was considered and rated, each slider had to be moved from its original position, least similar, before the test software would show the test as completed. If a listener wished to rate a pair as least similar then they were required to move the slider away from this point and back again. No restrictions were placed on the order in which listeners rated stimulus pairs nor on the number of times each pair could be auditioned. Ordering of the tests, stimulus pair ordering within each page, and stimulus ordering within each pair, were all randomised for each listener.

**B. Multi-dimensional scaling analysis**

MDS analysis of the similarity ratings was conducted for each programme item independently to find: (i) for each programme item, the number of salient dimensions across which the ten recordings differed; and (ii) for each dimension within each programme item, a pair of stimuli exhibiting a large difference, for use in Phase 3. MDS analysis was conducted in SPSS version 21 using the PROXSCAL algorithm.

1. Dimension analysis

After the work of Kruskal and Wish (1978), Martens and Zacharov (2000), Neher et al. (2006), and Brookes and Williams (2010), the number of salient dimensions in a dataset was deemed to be the dimensionality of the simplest MDS solution having a normalised raw stress of <0.1 where adding a further dimension would increase the squared correlation (\(r^2\)) by no more than 0.05.

Using these criteria, the number of salient dimensions used by the listeners to differentiate between the tested stimuli was found individually for the bass (three), drums (two), guitar (three), strings (four), and trumpet (five) programme items. Plots of the normalised raw stress for each programme item are included in the supplementary material.\(^1\) The maximum number of dimensions for an MDS solution for a set of \(N\) stimuli is \(N – 1\); however, as the number of dimensions increases above \((N – 1)/4\) the risk of a degenerate solution increases (Kruskal and Wish, 1978).

One indication of a degenerate solution is that the distances between the data are equal, or nearly equal, and thus the data lie on a circle (Takane, 2007). Visual inspection of each of the MDS solutions confirmed that this was not the case.

Although the shapes of the group spaces appear sensible, it is still possible that a solution might be degenerate. If this is the case then the range of microphones taken forward to subsequent phases might be sub-optimal, increasing the statistical noise in their results. If results of subsequent phases appear overly noisy, or if listeners report significant difficulty with their tasks, then the question of degeneracy will be revisited.

2. Selection of stimulus pairs

The MDS solution with the identified dimensionality for each programme item was plotted as a set of two-dimensional group space projections, and these plots are included in the supplementary material.\(^1\) For each programme item, a pair of microphones spaced most widely across each orthogonal dimension was identified. Across all 5 programme items, a total of 17 microphone pairs were selected. These are shown in Table II along with the mean dissimilarity scores which will be used in Section VII.

V. PHASE 3: ELICITATION OF DESCRIPTORS

With stimulus pairs selected that exhibit differences across each of the main perceptual dimensions of inter-microphone difference, a direct elicitation experiment was conducted to find terms describing these differences. The FCP method was used, where the response format is not limited and subjects are
free to use as many terms as required to describe the differences between stimuli (Bech and Zacharov, 2006).

Fifteen final-year undergraduate students on the Music and Sound recording course at the University of Surrey participated in this experiment. All had participated in multiple listening tests previously, and all had passed a taught module in technical listening, but none had previously participated in this study. For the avoidance of potential bias, subjects were given no information about the nature of the stimuli.

The test interface comprised 17 pages; 1 stimulus pair per page. On each page, subjects were asked to type in a text box as many terms or short phrases as required to describe the differences between the stimuli. Presentation order of the pages was randomised between subjects.

Listeners had commented previously that in performing some comparisons they focused primarily on particular portions of the stimuli. To facilitate this practice, ease the task, and potentially reduce statistical noise in the results, subjects were provided with the facility to define and loop sections of the audio. Stimuli were reproduced with the same setup as Phase 2.

A total of 768 descriptive terms were elicited.

VI. PHASE 4: DESCRIPTOR GROUPING & ATTRIBUTE LABELLING

To remove redundancy from the elicited terms panel discussions were held, similar to those in the work of Zacharov and Koivuniemi (2001) and Francombe et al. (2014). All 15 subjects who participated in the FCP experiment participated in these panel discussions.

Each of the 768 elicited terms was printed onto an individual card, with the associated stimulus pair on the rear of the card. The cards were then presented to the subjects one at a time, asking the subjects to group together any elicited terms referring to the same perceptual attribute. During the discussions, two sets of the original reproduction setup were available for auditioning each stimulus pair upon request.

TABLE II. Mean dissimilarity scores from Phase 2 experiment for the 17 selected stimulus pairs.

<table>
<thead>
<tr>
<th>Programme item</th>
<th>Microphone pair</th>
<th>Mean dissimilarity score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bass</td>
<td>SM58 and Knowles</td>
<td>70.44</td>
</tr>
<tr>
<td>Bass</td>
<td>4006 and WM7131</td>
<td>68.56</td>
</tr>
<tr>
<td>Bass</td>
<td>C12 and 4038</td>
<td>52.22</td>
</tr>
<tr>
<td>Drums</td>
<td>4006 and RE20</td>
<td>69.78</td>
</tr>
<tr>
<td>Drums</td>
<td>SM58 and Knowles</td>
<td>82.89</td>
</tr>
<tr>
<td>Guitar</td>
<td>C12 and Knowles</td>
<td>79.00</td>
</tr>
<tr>
<td>Guitar</td>
<td>SM58 and WM7131</td>
<td>62.44</td>
</tr>
<tr>
<td>Guitar</td>
<td>4038 and 4e2200</td>
<td>57.78</td>
</tr>
<tr>
<td>Strings</td>
<td>4e2200 and Knowles</td>
<td>63.89</td>
</tr>
<tr>
<td>Strings</td>
<td>4006 and RE20</td>
<td>49.56</td>
</tr>
<tr>
<td>Strings</td>
<td>414 and WM7131</td>
<td>63.56</td>
</tr>
<tr>
<td>Strings</td>
<td>451 and 4038</td>
<td>49.89</td>
</tr>
<tr>
<td>Trumpet</td>
<td>451 and Knowles</td>
<td>66.67</td>
</tr>
<tr>
<td>Trumpet</td>
<td>C12 and RE20</td>
<td>59.44</td>
</tr>
<tr>
<td>Trumpet</td>
<td>4e2200 and 4038</td>
<td>65.00</td>
</tr>
<tr>
<td>Trumpet</td>
<td>4006 and SM58</td>
<td>57.33</td>
</tr>
<tr>
<td>Trumpet</td>
<td>C12 and WM7131</td>
<td>37.56</td>
</tr>
</tbody>
</table>

The discussions reduced the 768 elicited terms into 38 groups. Subjects were then asked, for each group, to produce a label and a description for the corresponding perceptual attribute. Whilst conducting this stage of the discussion, subjects decided to arrange the groups into a hierarchy, and two additional, mid-level, empty groups were added to help structure the hierarchy. This resulted in a total of 40 groups and associated attributes, shown in Fig. 2. The attributes corresponding to the two additional groups are denoted with an asterisk.

From Fig. 2, it can be seen that each higher-level attribute, such as spectral content, can be considered as a combination of the lower level attributes: low-frequency content (LF content), mid-frequency content (MF content), and brightness in this case.

It is interesting to note that the number of attributes identified is greater than the total number of dimensions revealed by the MDS analysis in Phase 2. Although this could suggest that there is a degree of remaining redundancy, it is likely to indicate that at least some of those dimensions correspond to multiple attributes varying in parallel within the tested stimulus set. It is also interesting that several of the agreed perceptual attributes might more commonly be considered to be acoustic parameters (e.g., noise level, dynamic range, LF
content). However, the panel felt that these acoustic parameters were directly perceptible and identifiable and therefore could also be considered to be perceptual attributes.

VII. PHASE 5: ATTRIBUTE CONTRIBUTION

Phase 4 generated a list of 40 attributes in a hierarchy (with 31 attributes at the lowest level) that exemplify the differences between the recorded stimuli. However, it is not clear the extent to which each of these contributes to the overall difference between stimuli.

In order to determine which attributes contribute the most, an attribute contribution experiment was conducted. In this experiment, subjects were presented with each of the 17 selected pairs of stimuli in turn and asked to rate the contribution of each attribute to the overall perceived difference between the 2 stimuli in the pair. The test interface comprised a separate page for each stimulus pair (Fig. 3). Thirty-one sliders were presented to the subject, each pertaining to one of the lowest-level attributes and assigned a unique colour. The order of these was randomised for each subject, but maintained across different stimulus pairs. When hovering the mouse cursor over a slider, the definition of the attribute, agreed upon in Phase 4, would appear.

The attribute contribution chart, shown in Fig. 3, updated in real time to display a pie chart that represents the contribution of each of the sliders to the overall differences. Subjects were asked to make this pie chart representative of the overall difference. As in Phase 3, listeners were able to define and loop corresponding sections of the two stimuli to facilitate their preferred mode of listening.

The tests were split over three individual sessions containing six, six, and five pages, respectively. The pages were randomly ordered for each subject, and each subject rated each stimulus pair only once. Stimuli were again reproduced as in Phases 2 and 3.

The attribute contribution test was completed by the same 15 subjects as in Phases 3 and 4. Four additional listeners, with the same level of listening experience and training, completed the experiment in order to check for bias in the original subjects due to their involvement with the panel discussions.

A. Attribute contribution results

Since the subjects were asked to make the attribute contribution pie chart representative of the overall difference between the stimuli, the absolute position of each slider has no meaning (e.g., all sliders set to 20 will produce the same pie chart as all sliders set to 100). For the analysis, it was the percentage contribution of each attribute to the overall difference that was analysed.

1. Subject comparison

A univariate analysis of variance (ANOVA) was performed with the percentage contribution as the dependent variable, and the independent variables of stimulus pair, attribute, and the subject group (original 15 subjects or the additional 4 subjects). The subject group did not have a statistically significant effect ($p = 1.000$): subject group did not affect the overall result.

However, the interaction between the attribute and the subject group had a statistically significant effect ($p = 0.001$). Although this implied that the two groups of subjects were responding differently for some attributes, the $F$ statistic was low ($F = 2.010$) compared to that for other

![FIG. 3. (Color online) Listening test interface for the Phase 5 attribute contribution experiment.](image-url)
significant variables, such as attribute ($p < 0.001$, $F = 36.138$). Partial eta squared for this interaction was also low ($\eta^2_p = 0.07$). It was therefore concluded that this interaction effect was very small compared to other factors. The full ANOVA table is shown in Table III.

From this analysis, it was concluded that the original subjects were largely unbiased in their judgements. Therefore, all 19 subjects were considered as a single group in all subsequent analysis.

### 2. Overall attribute contributions

The relative contribution of each attribute to the perceived differences between the tested microphones was sought. This contribution can be determined from the product of two factors for each tested stimulus pair: (i) the attribute’s contribution to the overall difference between the two stimuli (i.e., the percentage contribution in the Phase 5 results), and (ii) the relative magnitude of the overall difference between those two stimuli (i.e., the mean dissimilarity score shown in Table II, divided by the mean of all mean dissimilarity scores).

The mean contributions (across stimulus pairs) and 95% confidence intervals for each attribute are shown in Fig. 4. From this it can be seen that brightness contributed the most overall to the differences between the microphones. The second-highest-contributing factor was noise level. However, it can be seen from Fig. 4 that the 95% confidence intervals are larger for noise level than for any other attribute. This suggests that the ratings of noise level contribution were not consistent across programme items. The full rank ordering of attributes, and mean contributions, are shown in Table IV.

### 3. Attribute contributions by microphone type

There was a concern that brightness was contributing highly to the overall difference due primarily to the MEMS microphones having a high-frequency resonance. Additionally, the large 95% confidence intervals for noise level implied that the ratings for this attribute differed greatly across stimulus pairs and it was felt that the noise performances of the MEMS microphones might have been largely responsible for this. To investigate further, analyses were conducted by microphone type, considering separately: (i) studio-vs-studio microphone pairs and (ii) MEMS-vs-studio microphone pairs.

Figure 5 shows the mean contribution of each attribute for the studio-studio and MEMS-studio comparisons separately.

---

**TABLE III.** Full factorial ANOVA table for the attribute contribution experiment.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
<th>p</th>
<th>$\eta^2_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stimulus</td>
<td>16</td>
<td>5.341 $\times 10^{-5}$</td>
<td>3.338 $\times 10^{-6}$</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Attribute</td>
<td>30</td>
<td>36625.444</td>
<td>1220.848</td>
<td>36.138</td>
<td>0.000</td>
</tr>
<tr>
<td>Subject Group</td>
<td>1</td>
<td>4.051 $\times 10^{-6}$</td>
<td>4.051 $\times 10^{-6}$</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Stimulus * Attribute</td>
<td>480</td>
<td>72688.289</td>
<td>151.434</td>
<td>4.483</td>
<td>0.000</td>
</tr>
<tr>
<td>Stimulus * Subject Group</td>
<td>16</td>
<td>3.463 $\times 10^{-5}$</td>
<td>2.164 $\times 10^{-6}$</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Attribute * Subject Group</td>
<td>30</td>
<td>2037.314</td>
<td>67.910</td>
<td>2.010</td>
<td>0.001</td>
</tr>
<tr>
<td>Stimulus * Attribute * Subject Group</td>
<td>480</td>
<td>16192.163</td>
<td>33.734</td>
<td>0.999</td>
<td>0.501</td>
</tr>
<tr>
<td>Error</td>
<td>8959</td>
<td>302658.076</td>
<td>33.783</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**FIG. 4.** Results of the Phase 5 attribute contribution experiment averaged over all stimulus pairs, arranged by rank order of the mean percentage contributions.
TABLE IV. Attributes ordered by overall contribution to inter-microphone differences.

<table>
<thead>
<tr>
<th>Rank order</th>
<th>Attribute</th>
<th>Mean contribution (%)</th>
<th>Rank order</th>
<th>Attribute</th>
<th>Mean contribution (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Brightness</td>
<td>12.849</td>
<td>17</td>
<td>Boominess</td>
<td>2.670</td>
</tr>
<tr>
<td>2</td>
<td>Noise level</td>
<td>6.654</td>
<td>18</td>
<td>Honky</td>
<td>2.519</td>
</tr>
<tr>
<td>3</td>
<td>Harshness</td>
<td>6.382</td>
<td>19</td>
<td>Ensemble balance</td>
<td>2.471</td>
</tr>
<tr>
<td>4</td>
<td>Clarity</td>
<td>6.161</td>
<td>20</td>
<td>Source distance</td>
<td>2.417</td>
</tr>
<tr>
<td>5</td>
<td>Piercing</td>
<td>5.378</td>
<td>21</td>
<td>Room timbre</td>
<td>1.716</td>
</tr>
<tr>
<td>6</td>
<td>Body</td>
<td>5.334</td>
<td>22</td>
<td>Punchiness</td>
<td>1.407</td>
</tr>
<tr>
<td>7</td>
<td>Tinny-ness</td>
<td>4.886</td>
<td>23</td>
<td>Dynamic range</td>
<td>1.225</td>
</tr>
<tr>
<td>8</td>
<td>MF content</td>
<td>4.259</td>
<td>24</td>
<td>Recording noise</td>
<td>1.025</td>
</tr>
<tr>
<td>9</td>
<td>Roundness</td>
<td>3.788</td>
<td>25</td>
<td>Image width</td>
<td>0.981</td>
</tr>
<tr>
<td>10</td>
<td>Instrument noise</td>
<td>3.708</td>
<td>26</td>
<td>Noise spectrum</td>
<td>0.909</td>
</tr>
<tr>
<td>11</td>
<td>Raspy</td>
<td>3.699</td>
<td>27</td>
<td>Reverberation</td>
<td>0.895</td>
</tr>
<tr>
<td>12</td>
<td>Realism</td>
<td>3.533</td>
<td>28</td>
<td>Perceived pitch</td>
<td>0.659</td>
</tr>
<tr>
<td>13</td>
<td>Lower mids</td>
<td>3.484</td>
<td>29</td>
<td>Reediness</td>
<td>0.643</td>
</tr>
<tr>
<td>14</td>
<td>Harmonics</td>
<td>3.331</td>
<td>30</td>
<td>Distortion</td>
<td>0.637</td>
</tr>
<tr>
<td>15</td>
<td>LF content</td>
<td>3.258</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Nasal</td>
<td>2.748</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

An ANOVA indicated that the effect of comparison type (studio- or MEMS-studio) was statistically significant \((p = 0.001, F = 10.93)\).

A one-way ANOVA, performed for each attribute individually, with comparison type as the factor, showed that the contributions of brightness, honky, nasal, tinny-ness, harshness, noise level, noise spectrum, recording noise, and instrument noise differed significantly according to comparison type. These attributes are shaded grey in Fig. 5.

Even though brightness was rated differently in the studio- and MEMS-studio comparisons, this factor is rated the highest in both comparison groups. However, the second most prominent attribute overall, the noise level, contributes very little to the difference in the studio-studio comparisons.

4. Attribute contribution by programme item

To analyse the effect of programme item on the contribution of each attribute, the results broken down by programme item are shown in Fig. 6. The range covered by the y axis on Fig. 6 is much larger than that on Figs. 4 and 5. This is because noise level (the second-highest contributor overall) contributes a large percentage to the overall difference for the bass programme item, but contributes very little for the other programme items. This might be due to the low SPL produced by the bass and/or to the absence of high-frequency programme content to mask the microphone’s self-noise.

This explains the large confidence intervals for noise level in the overall analysis, Fig. 4. Brightness, harshness, and clarity (the highest, third-highest, and fourth-highest overall contributors, respectively) contribute relatively large percentages to the inter-microphone differences for the majority of the programme items.

VIII. DISCUSSION

Comparing the microphone descriptors and attributes highlighted by Hebrock et al. (1996, 1997) against the attributes identified in the current study, it can be seen that there is little commonality; harshness, warmth, and LF content being the only shared attributes. It is notable, however, that this list includes the third most highly contributing attribute identified in the current study (harshness), as well as the higher-level attribute warmth, which is split into sub-attributes in the current study.

It is also worth noting that some other attributes of Hebrock et al. might in fact be similar (or even equivalent) to some in the current study. A further panel discussion such as that employed in Phase 4 might lead to, for example, extended high-end being grouped with brightness, or detailed being grouped with clarity. It might even group dull and muffled (Hebrock et al.) and identify them each as being equivalent to a lack of brightness. These possibilities underline the importance of Phase 4 to the current study.

Even allowing for these potential similarities or equivalences, however, the current study still identifies several additional attributes as being important. This confirms the value of the use of a wide range of microphones and programme items and of the adopted free elicitation approach.

The situation is similar when considering the findings of the loudspeaker-based studies reviewed in Sec. I. Clarity and brightness are the only attributes shared with the lowest level of the current study’s hierarchy, but both are in the top four according to degree of contribution. The higher-level attribute tone is also shared but the current study splits this into six component attributes. Again, there is also a possibility of similarity or equivalence between seemingly different attributes; for example, bass/treble balance, could be considered as a combination of LF content and brightness.

When considering the musical acoustics based studies reviewed in Sec. I, the attributes brightness, clarity, harshness, and nasal are shared with the lowest level of the current study’s hierarchy, and the first three of these are in the top four highest contributing (top three if noise level is discounted). The higher-level attributes warmth and ringing are also shared but are split into multiple sub-attributes in the current study. Finally, again, there is a possibility of similarity/ equivalence; for example, the dull/brilliant scale may refer to the same perceptual attribute as brightness from this study.
It is interesting to note that the attribute *noise level*, found to be the second largest contributing factor in this study, was not revealed in any of the previous studies. It was only revealed in the current study due to the deliberate selection of a very wide range of microphones (including MEMS microphones) and of programme items with a very wide range of characteristics (including the bass, which produced very little high frequency sound).

Thus, it seems that: (i) three of the four attributes found by the current study to contribute the most to perceived inter-microphone differences were also identified in previous studies; (ii) additional attributes were revealed by the current study, as a result of it focusing specifically on microphones, evaluating a wide range of microphones and programme items and employing free elicitation rather than allowing listeners to choose only from a limited prescribed attribute list; (iii) higher level attributes identified in previous studies can be broken down into multiple sub-attributes, each making a specific contribution; and (iv) panel discussions have the potential to identify equivalences between elicited descriptors and thereby reduce redundancy in attribute sets.

![FIG. 5. Results of the Phase 5 attribute contribution experiment broken down by microphone comparison type. Highlighted columns show significant differences between the studio-studio and MEMS-studio comparisons.](image)

![FIG. 6. Results of the Phase 5 attribute contribution experiment broken down by programme item.](image)
IX. CONCLUSIONS

The attributes contributing to the differences perceived between microphones (when auditioning recordings made with those microphones) are not clear from previous research, and perceived microphone characteristics do not always correlate well with manufacturers’ standard measurements. As a step toward developing a perceptually relevant set of measures of microphone quality, a five-phase study was conducted to determine the perceptual dimensions across which microphones differ and to find the relative contributions of the corresponding attributes to perceived inter-microphone differences.

In Phase 1, consideration of microphone technical specifications and expert opinions from audio engineers indicated that recording five programme items (double bass, drums, acoustic guitar, string quartet, and trumpet) with eight studio and two MEMS microphones (listed in Sec. III.A) would provide suitable stimuli to reveal the attributes comprising the most prominent inter-microphone differences. Such recordings were therefore made for use as stimuli in a listening test.

In Phase 2, pairwise listening comparisons between the resulting 50 stimuli, followed by multi-dimensional scaling analysis, revealed 17 salient dimensions and 17 corresponding pairs of stimuli exemplifying the differences across those dimensions.

In the FCP elicitation, in Phase 3, a total of 768 terms described the differences that listeners heard between the stimuli in each exemplary pair. Phase 4 then employed panel discussions to group the elicited terms and reduce redundancy, and identified a hierarchy of 40 perceptual attributes (Fig. 2).

Finally, in Phase 5, an attribute contribution experiment determined, for the 31 descriptors at the lowest level of the hierarchy, the degree to which each of them contributed to perceived inter-microphone differences. The results of this experiment allowed the attributes to be ordered by degree of contribution, and this ordering is shown in Table IV.

Further analysis revealed that, overall, brightness is the attribute contributing the most to inter-microphone differences (this was the case for all programme items and for the majority of microphone pairs). Noise level, although ranked second overall, only contributes highly when microphones differing greatly in self-noise are used to record a source that lacks high frequency content. Brightness, harshness, and clarity were shown to contribute highly for all programme items and for all microphone pairs.

Future work will develop models of the attributes contributing the most to perceived inter-microphone differences, in terms of objectively-measurable parameters. Such models could facilitate microphone development and testing by manufacturers, and microphone selection by users.

ACKNOWLEDGMENTS

The work reported in this paper was supported by the UK’s Engineering and Physical Sciences Research Council (EPSRC) and Cirrus Logic. Special thanks to Tsjerk Hoekstra and Martin Ross at Cirrus Logic for supplying and mounting the MEMS microphones in this project. The data underlying the findings presented in this paper are available from DOI 10.5281/zenodo.30621. Further project information is available at http://itosr.uk/MicQuality.


