THE SELECTION OF LOUDSPEAKERS FOR BBC RADIO & MUSIC

R. Walker, BBC Research and Development Department.

1 INTRODUCTION.

The monitoring loudspeaker is unlike any other part of an audio production chain. It is the means by which the audio signal is judged. Other parts of the production chain certainly have effects, but they can be judged, altered or adjusted based on the final sound as reproduced by the loudspeakers. Historically, the BBC satisfied its requirements for studio monitoring loudspeakers by designing and developing its own range. The requirements could not generally be met at that time by contemporary commercial products, though some were used in special cases.

Apart from the obvious need for high quality reproduction, the large number of studios and control rooms (more than 600 in the period around 1980) meant that the quality and character of the audio reproduction had to be maintained over a large number of units, over a range of loudspeaker sizes and for periods of many years. One of the main reasons for requiring consistency was (and still is) production practices that require operational staff to move between rooms for different jobs and for programme audio to be processed consecutively in different facilities (especially in television). Both of those factors require the sound monitoring in different rooms to be as similar as possible, otherwise much time and effort can be spent in doing and re-doing changes to the recorded material. That is unlike a commercial recording studio or small radio/TV station, where the technical staff work almost all the time in the same few rooms and can become familiar with their eccentricities.

Those considerations led, over the years, to a succession of BBC-designed loudspeakers that also achieved substantial national and international recognition. The BBC-designed loudspeakers introduced several new developments, such as the use of polyethylene/ polypropylene (PE/PP) co-polymers for the diaphragm and Kapton® voice coil formers. The unit costs of BBC-designed loudspeakers were also significantly lower than equivalent commercial products, at that time.

The last successful BBC-designed loudspeakers were the small LS3/5A (*c.* 1975), the large LS5/8 (*c.* 1982) and the medium LS5/9 (*c.* 1984). Those loudspeakers were the main ones in use from the dates of their original designs to the present. Very many are still in use at the present time (2004). In design, they may be well past their time and are, in any case, now irreplaceable. This author suggested in about 1992 that work should be started on their successors. At that time, the loudspeaker designs were between 8 and 17 years old. They are now between 20 and 29 years old.

Three main factors led to the demise of the loudspeaker development work. With the coming of a more commercial imperative to the BBC as a whole, it was no longer considered worthwhile to develop loudspeakers internally. At that time (around 1994), the individual Directorates were made autonomous and responsible for managing themselves. That led to a lack of support for the sort of centralised activities needed to support loudspeaker development. Also, each business unit was isolated and indeed had been deliberately set up to be in internal competition with other units. The choice of loudspeakers was seen as an individual selection, based on each unit's own activities or cost considerations. That policy has led, over a period of about the last 10 years, to significant fragmentation, with many different types of loudspeakers being used in different areas. That has in

turn led to the recognition (at least in some places) of the value of that earlier requirement for consistency.

A second factor has been the availability of improved commercial loudspeakers. Over the period of about the last 40 years there is no doubt that commercial loudspeakers have improved dramatically. Even over the last 20 years since the design of the LS5/9, developments in commercial loudspeakers have led to improved properties, perhaps most of all in consistency.

The third factor is the cost and complexity of modern loudspeakers. Years ago, a loudspeaker consisted of a few relatively simple components. Of course, the sound quality, reliability and other parameters depended on the choice of materials and their physical forms, but the principles were straightforward enough for a relatively small team to succeed in the design. Now, the additional refinements, the improvement in achievable quality and the complex design techniques make it less practicable for a small team to embark upon the development with any realistic hope of success. Designing high quality loudspeakers is an activity that can only be contemplated by a well-resourced team [1]. The addition of modern complex electronics, for room equalisation, for allowance for component variations or for the addition of digital control or input interfaces, further adds to the potential scale of the task.

BBC Radio and Music (BBC R&M) considered that the scale of the project made it necessary to carry out a study to select a range of loudspeakers with the objective of setting up a Framework Agreement for the procurement of 'standard' loudspeakers. They invited News and World Service to take part, so that they could take advantage of the results if they wished. The size of the requirement made it necessary to go through a formal tendering process. It also exceeded the EU threshold, which meant that the EU Procedure (Restricted) process had to be applied.

In order to include the quality of the loudspeaker as a selection parameter it was necessary to carry out formal listening tests. Even without that potential legal requirement, it was thought desirable to have a uniform loudspeaker installations throughout the new development. That required all potential user groups to support the selection process and to accept the outcome. It was also hoped that other Directorates might accept the idea of a new set of 'standard' BBC loudspeakers for future developments. Of course, under the existing separations between Directorates, there is no guarantee that will be so, but it would clearly be beneficial to the Corporation as a whole if it helped to keep down costs and contribute to uniformity of output.

2 THE TENDERING PROCESS.

In outline, the 'Restricted' EU tendering process consists of two stages. The first is an advertisement for potentially interested parties. The responses to the advertisement were studied and used to create a short list for the subsequent 'Invitations to Tender'.

The Invitation called for proposals for three different sizes of loudspeaker to replace the LS5/8, LS5/9 and LS3/5A. It was made clear that the audio quality selection would be on the basis of subjective assessment alone[2]. However, a brief technical specification was included in the Invitation as guide to the general performance requirements. The subjective assessment was to be made on the basis first of absolute audio quality for each loudspeaker size and then on the 'family' resemblances between the three different sizes[3]. The set of three loudspeakers was considered as

[1] It is true that there are still some individual designers working in the field. Generally, they produce loudspeakers with individualistic characteristics, which may well be excellent in their field of application, but which do not satisfy the BBC's requirements for a widely used, standard loudspeaker.

[2] Other non-audio factors, such as handling, maintainability, interchangeability, etc. were also important, as were business considerations not specifically related to loudspeakers, but they were not part of the audio quality selection process.

[3] Throughout this selection process, the terms 'large', 'medium' and 'small' were used to identify the three broad categories of loudspeaker. In the design, it is inevitable that more compromises have to be made to obtain good performance from a smaller loudspeaker. The use of loudspeakers is frequently constrained by the space available and

a family of products needed to fulfil the BBC requirements. They did not necessarily all have to be provided by a single manufacturer.

The first level of assessment was the response to the initial advertisement. The accompanying questionnaire covered aspects of the supply of loudspeakers not directly related to their acoustic performance, such as company size and financial controls, staff health and welfare and company history. That was intended to eliminate commercially inappropriate proposals. At that stage, there was no exclusion of any potential supplier who might otherwise provide suitable loudspeakers. All respondents who met the commercial requirements were included, even though that resulted in more loudspeakers to be included in the assessments than had been thought practicable at the outset. Some of the proposals included the same loudspeakers in different combinations. It was clear that the final total of 19 different proposals was too many to assess. By removing the repetitions and the impractically expensive loudspeakers, the total number of proposals was reduced to 10. Each of the proposals included submissions for all three of the target sizes, so the total number of loudspeaker pairs actually to be assessed subjectively was nine large, ten medium and nine small (28 in total).

3 TEST METHODOLOGY.

3.1 The Parameters of an Ideal Test.

A rigorous subjective test should consist of double blind, individual assessments using a suitable number of descriptive axes by a suitable number of test subjects. Probably five to ten subjective axes would be used. That is likely to be higher than the true number of different quality attributes, but would allow some scope for subsequent statistical factor reduction. Other tests doing similar things have commonly used around 20 descriptors. If the set of descriptors is too limited then test subjects feel constrained whilst having too many leads to confusion and overload of the test subjects.

It was thought that not less than 10 test subjects would be needed, if they were highly skilled and could demonstrate a reasonable degree of consistency. For less able subjects, at least two or three times that number would have been necessary. The subjects were also being asked to assess many different aspects of loudspeaker quality at the same time. Less expert subjects would have found that difficult. Because of the background to the tests, it was thought that the test subjects should also be, in some sense, 'chosen' by their colleagues, so that the tests would be seen as reasonable and the final results more readily accepted by the majority of production staff who had not taken part directly.

The total number of loudspeakers was also an issue. At the outset, it had been anticipated that as many as eight suppliers might provide samples of each of the three sizes, resulting in 24 loudspeakers (stereo pairs) to be assessed. In the event, the final number was 28 pairs. Some dummy or repeat assessments also had to be included. In particular, the first assessment each day had to be rejected because there was no way of indicating to the test subjects the likely range of quality. Subjective tests, at least absolute assessments, usually include 'anchor points' to guide the subjects in the range of responses to be expected. That was not feasible in this case because the loudspeakers had to be moved between presentations. It had also been decided at the outset that the existing range of BBC loudspeakers would not be included as references because that would unfairly bias the results towards the existing and familiar products. It was felt that the proposals should be assessed entirely on their own merits, not in comparison with existing products. With one dummy test added to each set, the final total number of quality assessments was 31.

---

the accompanying reductions in performance have to be accepted. For cost-sensitive applications, smaller loudspeakers also usually cost less.

The time allowed for each person's assessment also has to be reasonable. It was thought unreasonable to ask subjects to begin immediately with completely unknown material so some time was allowed for familiarisation using the subject's own material. It was estimated that the familiarisation and assessment could not be done in less than 30 minutes.

In loudspeaker listening, the acoustics of the room have a significant influence on the perceived sound. That, at least, would be the same for all loudspeakers. However, more significantly, the position of the loudspeaker in the room also has a profound effect. That meant the assessments had to be done with all of the loudspeakers in the same positions in the room and that each test had to include changing the physical positions of the loudspeakers. In addition, some people think that having another loudspeaker present anywhere in the room affects the perceived sound[1]. That meant not only did the loudspeakers have to be changed in position but actually removed from the room when not being assessed. All of that had to take place without the subject knowing what was happening. In practice, the test subjects would have to be moved to a separate waiting area while the loudspeakers were changed over and the new ones connected, tested and their levels set. All of that was thought likely to add at least 30 minutes to each assessment.

So, it was thought that a full set of tests would consist of about 240 assessments. With 60 minutes each that would total about 240 hours of testing, or about 30 working days.

There were two possibilities :-

1) One test could be carried out per day with one loudspeaker and all ten subjects. With the final number of loudspeakers, that would have taken a total of 28 days to test all of the loudspeakers. It would have meant all ten subjects attending the test venue on 28 separate occasions for one hour on each occasion. Taking into account the significant travelling time, even from central London never mind from the Regions, that would have been logistically impossible.
2) Alternatively, each of the test subjects could be present at the test venue for an extended period and the loudspeakers changed over. About eight tests per day might have been achieved. That would have made a total of about four days. The total time for the testing would therefore have been about 40 working days. However, it would have been utterly exhausting for the project team, who would have to arrange the loudspeakers in the room a total of 310 times and probably too exhausting for the subjects (who would have to concentrate very hard for most of the four days, admittedly with reasonable breaks every 30 minutes or so).

Both of these ideal options were considered to be unworkable. Option '1' had the serious problem that no dummy or 'ranging' tests could be sensibly included. Each test subject would listen to only one loudspeaker each day and would have to carry mental recollections of the quality over a period of more than a month. That was clearly not sensible. Option '2' had the advantage of minimising the commitments for the test subjects, but involved an unacceptable amount of work for the project team. Both of the options would have taken too much time overall. It was clear that the procedure was going to have to be compromised in some way in order to make the tests practicable.

3.2 Possible Simplifications and Their Defects.

A number of considerations applied to the simplification of the test procedure:-

1. Whatever else happened, there had be some randomisation of the order of presentation to reduce the effects of order as a factor in the results. If that were not done, the results would depend very heavily on the presentation sequence. Even expert listeners would find it difficult to avoid their assessment being relative to the previous samples rather than independent of them.

[1] The author does not subscribe to this view. Most loudspeakers are so inefficient electrically and are in any case normally driven from a virtually zero impedance source that the possibility of received sound being re-radiated after being converted to electrical signals and back again is entirely negligible. The box itself is an object in the room, but no more significant than any other piece of furniture. Some loudspeakers have open resonant ports for low frequency tuning, but the damping/reverberation time of those resonators is entirely swamped by the room itself (in any reasonable room and with any reasonable loudspeakers).

That would be impossible with scheme '1' but possible to do with a test programme based on scheme '2'.

2. More than one loudspeaker pair could be put in the room at the same time in as near to the optimum positions as possible (restricted significantly in some cases because of the size of some of the loudspeakers). That would simplify the changeover. However, that would unfairly influence the results because of room position effects, apart from any questions over interactions between the loudspeakers.

3. The obvious simplification was to allow several subjects to assess the loudspeakers at the same time. If that were done without allowing them to change seating positions then the results would again be heavily biased by the test subject's seat position. If they were allowed to move around during the session then there would inevitably be some risk of collusion. Collusion leads to effectively fewer test subjects, especially if there are one or two dominant personalities in the room. However, the test subjects were all experienced operators, with independent opinions. It was thought likely that they could form their own views, independent of their colleagues. (In the event, the results showed that most test subjects did form reasonably individual opinions.) That simplification would not have been acceptable with inexperienced test subjects.

4. Given the time-scale and scope of the testing, the only practicable test procedure would be to have all the sample loudspeakers on site and to invite all of the test subjects to be at the test venue for an intensive period of three or four days.

5. It was not possible to carry out the tests entirely 'double-blind'. Inevitably, the project team would know which loudspeakers were being assessed. However, only one member of the project team remained in the room to supervise the tests – to ensure that the test subjects did not confer or try to look behind the screen to identify the loudspeakers. That person was often actually unaware of the loudspeaker manufacturer because code letters were used to identify the loudspeakers. In any case, under the pressure of physically moving the loudspeakers and setting the levels, the members of the project team had little time (or inclination) to take note of the identity. The supervisor was also under strict instruction not to exhibit any reactions.

3.3 The Final Test Procedure – Absolute Quality.

The final procedure for the quality tests consisted of one day's testing for each of the three sizes of loudspeaker. Each day consisted of 10 or 11 assessments (depending on the number of loudspeakers in the group). On occasions, errors in set up procedure or defects in equipment caused some assessments to be re-started. When the fault was dearly noticeable to the subjects and made it unrealistic to continue, for example one channel not working or a 'ringing' loudspeaker stand [1], the test was re-started immediately. When the fault was not directly evident, for example if a mistake had been made in the overall level setting, the whole test was repeated without the test subjects being aware and the first results discarded. The final programme allowed 45 minutes for each assessment, 30 minutes for the actual assessment and 15 minutes for the technical changeover. Longer breaks were included for refreshments and meals. Overall, the test programme was very demanding of both the project team and the subjects, with additional breaks having to be added sometimes.

The test panel was nominally 12 subjects, but not all were present all of the time. Eleven test subjects attended all of the tests. Because it was thought that 12 at once were too many for the room, the test panel was split into two groups. The whole test procedure for quality assessment was therefore carried out twice, once for each group. The two test sessions were held in mid-December 2003 and early January 2004. The orders of presentation were different for the two groups.

[1] The suppliers had been asked to provide stands if they thought it necessary to do so. Otherwise, ordinary metal frame stands were used. In more than one case, the manufacturer's own stand was found to be unacceptable.

The reproduction level was controlled to within ?0.5 dB using a pre-recorded pink noise track on the test disk and a Bruel and Kjær 2231 Precision Sound Level meter set to 'flat', located at the central front seat position. The pink noise recording was band-limited to 200 Hz – 8 kHz, removing the lowest frequencies to provide fair comparison signals for the smaller loudspeakers. The individual tracks of the test disk had been set subjectively to appropriate levels for the type of material by members of the project team. During the tests it was not necessary to adjust the levels. The test disk was simply allowed to play from beginning to end, with 13 tracks of about 1 minute each. A period of about 10 minutes was allowed before the formal test for the test subjects to use their own material for familiarisation.

The quality assessments were carried out using two-channel stereophonic material, though some tracks were presented in left or right channels only.

3.4 The Test Material.

The test material was selected to cover a range of genres. It included male and female speech, drama and several musical items. Not all of the material was the highest quality. Monitoring loudspeakers have to be able to present defective material accurately in order for the defects to be identified. Each extract lasted for about 1 minute. Thirteen tracks were finally selected for presentation. The recording medium was audio CD. Replay was by Studer A730 CD player (digital output) with a Prism 'Dream DA–1' digital-to analogue decoder. Most of the loudspeakers included built-in amplifiers. For the remainder, analogue power amplifiers were supplied by the system proposers. All of the systems tested used analogue audio inputs to the loudspeakers/amplifiers.

3.5 Absolute Quality Assement Parameters.

The absolute quality was assessed on a total of 19 individual parameters divided into five main attributes and an overall score. The parameters were :-

SPECTRAL UNIFORMITY – uncoloured tonal reproduction of the audio material, evenness of frequency response, granularity of frequency response, adequate, but non-emphasised low frequency response, adequate, but non-emphasised high frequency response.

SOUND-STAGE IMAGING – Correctly located sound stage, continuous, even and wide left-right sound stage, localisation and separation of individual sound images, stable and open sound stage.

AMBIENCE REPRODUCTION – spacious and diffuse reverberant sound, rendition of direct-to-reverberant sound ratio, rendition of recording sound space, uncoloured reproduction of reverberation.

DYNAMICS AND DISTORTION – consistency of performance at reduced level,

adequate rendition of material recorded at low level, level, purity and impact of loud sounds, lack of range compression and frequency intermodulation.

OFF-AXIS PERFORMANCE – performance off the centre line, loudspeaker plane or with head movements.

SUMMARY – Overall Impression.

The 19 parameters were to be given a numerical score between 1 and 5, with '1' representing 'poor' and '5' representing 'good'. A space was left at the end for free-form comments.

3.6 Test Procedure – Family Resemblances.

The second stage of the tests, assessing family resemblances was thought to be technically less problematical. It was arranged as a follow-on test, after the loudspeakers had been put in order of preference by the quality assessments.

It was inescapable that all three of the loudspeakers being assessed for similarity should be in the room at the same time, to allow rapid switching between them. It was considered quite impossible for test subjects to assess or remember differences or similarities over longer periods of time. The absolute quality had already been decided, so that the effects of the room and the other loudspeakers on quality were less important. Because it was considered less critically dependent on the test conditions, all 11 of the test subjects were present at the same time. The three loudspeakers being compared were placed close together, so that room position effects would be minimised. The tests were carried out with monophonic signals, the source being the same test recordings with the two channels bridged together.

The scoring sheet for the assessment of family resemblance consisted simply of three boxes for the two-way comparisons between the large, medium and small loudspeakers. The points were awarded as +2 for "very good match", 1 for "fairly good match", -1 for "slightly different" and –2 for "very different". The instructions were to assess and score each of the three comparisons. There was no fixed time limit; the test subjects were allowed to continue until all were satisfied that they had produced a fair assessment.

The selection of groups for the family resemblance tests was guided by two main factors. Firstly, the best individual loudspeakers should be selected. Secondly, if the first consideration did not lead to an optimum selection for family resemblance then loudspeakers not scoring the highest in absolute quality might have to be included. A third and minor consideration, to be used to resolve otherwise closely-matched selections, was that preference should be given to loudspeakers from the same manufacturer.

The family matching was considered to be a high priority. It was not so high that the absolute quality should be seriously compromised but high enough to allow for the rejection of the best individual loudspeakers. This was expected to be a difficult compromise to achieve, and so it proved to be. The technique adopted was first to try a few obvious groupings and then discuss with the test subjects other possible groupings. The test subjects remained unaware of the identity of the loudspeakers throughout. All discussion was carried out using the coded identifications.

4 ROOM ACOUSTICS AND THE TEST ROOM.

The room to be used for the tests was the subject of much discussion during the preparations for the tests. In the past, there had been a reasonably clear view of what a typical BBC control room was like, acoustically. Usually, they were (and many still are) heavily treated to produce much shorter reverberation times than most audience environments and to control early reflections by absorption. Many people have felt for some time that there should be a change in that basic room design. Indeed this author, in around 1991, began a process for change in the development of room acoustics that allowed much more 'lively' rooms, whilst still maintaining control of early reflections[1,2,3,4,5]. The new studio developments, and in particular, the development for which this loudspeaker selection process was being carried out, were (allegedly) to include much less acoustic treatment. It was thought best if a room more like the proposed new rooms could be used for the loudspeaker selection.

In the event, that proved to be impossible. Firstly, it was difficult to find a room of any reasonable sort that was free for the necessary period. The use of reference listening rooms at BBC R&D had been ruled out on the grounds that they were too unrepresentative. Secondly, the messages

coming from the design teams for new buildings suggested that the higher quality rooms at least would not be quite as different to present rooms as at first thought. That was more or less inevitable. Critical assessment of sound quality cannot be carried out in rooms with large areas of specularly-reflecting surfaces and long and uneven reverberations times, however attractive they might be architecturally.

A room was eventually found that was available, centrally located and that had adjacent storage areas for the test loudspeakers[1]. The room had been a control cubicle for radio drama and so had the 'standard' BBC acoustic treatment already installed, though it had been disused for some time and was visually unattractive. To modify the room acoustically, it was made acoustically symmetrical by the addition of similar acoustic treatment over the existing observation window and some of the wall treatment was removed where it would not result in discrete early reflections in the listening area. Some patches of high frequency absorption were added to areas where equipment had been removed and the resulting bare patches would have caused reflections. To hide the wall surfaces, to make the room somewhat more attractive and thus not cause too much distraction for the test subjects, the entire inner wall surface was lined with lightweight fabric drapes. The final mid-band reverberation time was about 0.2 s.

To provide for blind testing, an 'acoustically transparent' screen was erected across one end of the room, as an enclosure for the loudspeakers and amplifiers under test. The screen was made of lightweight black net fabric and lit by downlights so that the test subject could not see what was behind. The acoustic transmission properties of several different samples of material were measured and the one showing the least acoustic effect was selected. It showed a uniform progressive fall in transmitted sound level with increasing frequency, amounting to about 1dB at 20 kHz. That would probably have been detectable in a direct comparison but it was the same for all of the loudspeakers tested and was considered to be acceptable.

5 RESULTS.

5.1 Absolute Quality.

The quality scores from the test subjects were averaged. That took into account the varying numbers of test subjects and their varying numbers of responses. It had not been insisted on that all subjects should score all attributes. Sometimes, subjects were unable to give numerical scores for all of the 19 attributes. Sometimes, they just gave an overall score for a group of attributes. In fact, there were only a very few instances of no score being given at all but there were perhaps 20% of cases where an aggregate score had been given for two or more attributes. The use of averages allowed the data processing to take account of those different numbers. The scores were averaged for the five quality attributes individually and overall. The overall scores were also calculated for the two groups of test subjects separately.

The results from the quality assessments are shown in Figs. 1 to 4. Fig. 1 shows the overall average results for each of the loudspeakers for all three sizes. Figs. 2, 3 and 4 show the average results by attribute for each of the three sizes of loudspeaker. The 10 proposals were given code numbers 1 – 10 and the three sizes were given code letters 'a' for large, 'b' for medium and 'c' for small. In those cases where the same loudspeaker had been included in more than one proposal, the data was duplicated (e.g. a4/a8 and c4/c8). The loudspeakers themselves had not been assessed twice (as discussed above).

Because the scale for scoring quality was from 1 to 5, an 'average' loudspeaker would score 3. There was a strong tendency for the test panel to avoid scoring either very high or very low. Individual scores of 1 were not given often, most test subjects probably thinking that they would reserve '1' for really poor loudspeakers, which did not in fact appear very often. At the other end of

[1] Maida Vale, Studio 7 control cubicle.

the range, scores of '5' were also comparatively rare, individuals probably reserving that for better loudspeakers than actual appeared. Certainly, there were many instances of loudspeakers being given scores of '1' (or even 0 in extreme cases) or '5' by individual test subjects who particularly disliked or liked a specific loudspeaker. However, those very polarised results tended to be swamped in the averaging, so that the overall results tended to be compressed into the middle range. An average of '2' would have actually indicated a poor quality loudspeaker and '4' an exceptionally good one [1]. Fig. 1 shows that, in fact, the entire range of average results for all three sizes was from 2.42 to 3.72. For the individual attributes, the total range of values was 2.08 to 4.0. That illustrates the compression of the results into a reduced range.

In Fig. 1, some of the proposals scored either all poor (e.g. No. 7) or all good (e.g. Nos. 8, 9) over the range of sizes but in the majority of cases, the three sizes scored quite differently. That suggested that it might be difficult to select the best loudspeakers to make up family groups.

In Figs. 2 to 4, the individual attributes were more consistent. Generally, those loudspeakers with poor overall results also scored poorly in all attributes, and vice versa. That was encouraging, as it showed loudspeakers could not be selected for particular attributes. Thereafter, only the overall scores were taken into account. The converse of that would have been very difficult to manage. If it really had been the case that some loudspeakers were especially suitable for some applications then selection of a 'standard' set of loudspeakers might have been impossible.

5.2 Family Resemblance.

The Table 1 shows the results for the family resemblance tests, ordered by the overall family resemblance score. The total possible range for the score was –2 to +2, showing once again a compression of the actual results into a restricted range.

		Loudspeakers in test . . .			**Loudspeaker quality scores . . .**			
Test #	Overall average score	Large	Medium	Small	Large	Medium	Small	Total-9
6	0.93	a5	b4	c5	3.55	3.21	3.61	1.37
9	0.67	a5	b5	c5	3.55	2.95	3.61	1.11
8	0.44	a2	b2	c5	3.49	3.35	3.61	1.45
3	0.23	a2	b2	c2	3.49	3.35	3.09	0.93
5	0.17	a9	b9	c9	3.28	3.34	3.42	1.04
7	0.13	a6	b4	c4	3.72	3.21	3.21	1.14
2	-0.30	a4	b4	c4	3.15	3.21	3.21	0.57
4	-0.87	a6	b3	c3	3.72	3.56	3.48	1.76
1	-1.13	a5	b3	c3	3.55	3.56	3.48	1.59

Table 1. Family resemblance results sorted by overall score.

The columns are arranged as 'Number of the test' followed by the 'Score', followed by three columns listing the actual loudspeakers included in that test. Columns 6 to 8 give the overall average scores for those loudspeakers from the quality tests. The final column gives the total average quality score. In the final column, the differences have been emphasised by subtracting the average quality score for three loudspeakers of 9.

The order of preference for family resemblance depended on which of the comparisons – large/medium, medium/small, large/small or overall – was used to order the list. The table shows the list ordered by the overall result. Other tables were generated (not shown) with ordering by the three individual two-way comparisons. The top four family test results for each of the four orders

[1] All of the loudspeakers assessed in this trial were from the upper end of high quality professional monitoring loudspeakers and that a 'poor' score might still represent a loudspeaker that was good or excellent in terms of absolute performance for other applications.

gave 6,9,3,8 / 6,5,9,8 / 8,6,9,5 and 6,9,8,3 as the preferences, for respectively the large/medium, medium/small, large/small comparisons. The shaded section in Table 1 indicates those best five candidates. The others scored too low on either the absolute quality or the family resemblance tests to be considered as contenders.

Test 6 was obviously a good combination. It came top (three times) or second (once) in the family resemblance test results and had a reasonably high quality score, coming fourth overall. However, it did include a mix of manufacturers (the middle size was different). Test 9 produced lower results from the quality tests. It was thought to be too far down the quality preference order to be a first choice, even though the family resemblance positions were 2nd, 3rd, 3rd, and 2nd. Test 8 had very good results from the quality tests (3rd place overall and top out of the five best results for family resemblance). However, it produced lower results from the family resemblance tests, coming 4th, 4th, 1st and 3rd respectively. Test 3 only featured in two of the top four family resemblance results (sixth and seventh in the other two). It also had low results for quality. Test 5 produced poor result from quality tests, although all sizes got about the same score. It also featured in only two of the top four family resemblance test orders, coming 5th and 8th in the others.

Of these results, Test 8 appeared to be a reasonable first choice. It achieved 3rd place overall in the quality tests and 3rd in the overall family resemblance results, but was only 4th in the important Large-Medium ordered list.

Tests 6 and 9 were both though to be possibilities for a second choice. They differed only in the selection for the medium sized loudspeaker. They both came high in the results for family resemblance. Test 9 did have lower quality results but had the advantage of being a single manufacturer.

6 FINAL SELECTION.

The loudspeakers in Test #8 of the family resemblance tests were a2, b2 and c5, i.e. a mixed manufacturer group. The small loudspeaker from the same manufacturer as the two larger sizes, c2, only obtained an average score in the quality assessments and the family resemblance score was not high (Test #3). However, a mixed set was not considered unacceptable. The large and medium sizes, a2 and b2, were certainly very well liked. Taken together, they scored higher than any other pair of large/medium loudspeakers (Fig. 1). The small loudspeaker, c5, was the highest scoring one in its size category. Overall, that should have been a fairly clear winning combination, with perhaps a bias towards absolute quality.

However, there were some serious reservations about the maximum sound level capabilities of the large loudspeaker. It was physically much smaller than others in the 'large' category. The original guide specification had indicated a maximum output level of 108 dB spl, which it was easily capable of at mid and high frequencies. However, even the manufacturer's own data showed very limited low frequency output – with measured indications of stress and distortion by 90/92 dB spl at 50/63 Hz. Listening tests confirmed that the limited low frequency output would be an operational restriction in many areas, though adequate for some others.

One option would have been to select that loudspeaker as the standard for the large size and to specify something else in areas that required higher output. However, it was decided that would lead to too many special cases, maybe even the majority of 'large' loudspeaker installations, and would partially defeat the objective of selecting a standard set of loudspeakers. Therefore the selection was rejected.

The next reasonable option was either Tests #6 or Test #9. They were the same except for the choice of the middle-sized loudspeaker. They both came high on the list of family resemblance results, but their results for absolute quality were substantially different. That led to an investigation of some very different results obtained for the 'b5' loudspeaker from the two groups of test subjects.

The first group had given the loudspeaker an overall score of 3.84 and the second group 1.91. The first group thought the budspeaker better than any other by a large margin. The second highest-placed medium loudspeaker by that group scored 3.50. The second group thought the loudspeaker worse than any other, though by a smaller margin. The next lowest placed medium loudspeaker by that group scored 2.17 [1]. That seemed to be a perverse anomaly that required further investigation.

The discrepancy needed to be resolved, though the issue actually only involved the selection of the medium-sized loudspeaker, whether it was to be 'b4' or 'b5'. The two samples of loudspeaker 'b5' were listened to extensively by the project team in the test environment. It became clear that the two samples supplied were significantly different in sound quality. When placed together, the difference was striking. Of course, there was no indication of which one was more 'correct'. It was decided that further samples of the type should be obtained and assessed. An additional pair was acquired and assessed by the project team. That pair was more consistent and also matched one of the original pair fairly well. It was concluded that one of the original pair had been somewhat different. It was still not clear whether that represented normal manufacturing tolerances or a defective sample. The effect of the difference in the responses of the original pair might well have been to give a 'spread' in the stereo image quality, corresponding to a "pulling apart" or "stretching out" of the stereophonic images reported by some members of the project team. Such an effect might have been picked up by one group of test subjects but not by the other, especially if their attention was also being taken up by a long list of other factors. Further assessments, and some measurements, were carried out at BBC R&D Department, using another listening room and some other test subjects, comparing all four of the samples.

The conclusion at the end of all of the supplementary tests was that the anomalous score given by the second group of test subjects was probably correct but that it had been based on a defective or atypical loudspeaker sample.

The family of loudspeakers represented by 'a5', 'b5' and 'c5' had very closely matched manufacturing processes. Indeed, the project team were quite convinced that some of the actual drivers were the same in some cases, though the manufacturers did state that they were not precisely the same. In that context, it seemed unlikely that the loudspeakers should score significantly differently in the quality tests. In the end, it was recognised and accepted that the final choice for the medium-sized loudspeaker was based on somewhat uncertain evidence. The balance of probability, taking all the factors into account, was felt to be in favour of loudspeaker 'b5' as the preferred choice over 'b4'.

It was also the case that loudspeaker 'b4' was already in use in significant numbers in some parts of Radio. Therefore, the existing loudspeakers could be re-deployed without conflicting with the requirement for family resemblance as it provided a good match and the existing stock would remain of some value. Thus, the final preferred selection for new purchases was 'a5' for the large sized loudspeaker, 'b5' for the medium sized loudspeaker and 'c5' for the small sized loudspeaker.

7 SUMMARY AND CONCLUSIONS.

This report has presented the background to the requirements for a range of new, standard loudspeakers for BBC Radio & Music. That required a set of three sizes, to replace the BBC designed LS5/8, LS5/9 and LS3/5A still in use in many areas. Because of the size of the potential order, there was in any case a requirement to pursue a formal tendering and selection process for the supply of the loudspeakers.

[1] It was also true that the second group gave lower scores overall. The overall average scores from the quality tests for the two groups were 3.34 and 2.97 respectively. Given the compression of scores into a relatively narrow range, that represented quite a marked difference.

A project was set up to carry out the tendering process and to organise formal subjective tests on potential candidates as part of the selection process. The selection was to be made not only on the absolute quality of the loudspeakers but also on their family resemblance so that a more uniform sound quality could be achieved in a range of applications. The tests involved operational studio managers from many areas of radio as test subjects, proposed by their colleagues as being expert and credible judges of sound quality. The intention was to try to ensure that all staff potentially affected would have contributed in some way to the process and would, as a result, more readily accept the outcome. Loudspeakers are a sensitive topic on which many people have very different views and it was anticipated that reservations could be expressed by some. By trying to involve as many people as possible, even only in a peripheral sense, it was hoped that the transition to new loudspeakers could be made as smooth and trouble-free as possible.

Inevitably, the final outcome was not entirely clear-cut. It would have been simple if a single set of loudspeakers from one manufacturer had stood out from the rest in quality and family matching. In the event, the best loudspeakers in each size category were not selected because of the family matching requirements and other technical performance issues, but the final compromise was still close to the best. The final choice consisted of three different sizes of loudspeaker from the same manufacturer. That should have benefits for consistency in the future, especially if that manufacturer were to change the production method or introduce new models. It should be noted that many other factors were also included in the selection process. They included weight, size, maintainability and many others. The acoustic differences were taken into consideration amongst those other factors in the overall selection.

8 ACKNOWLEDGEMENTS.

This work was managed and led by Stuart Collins of Radio Resources. Many other people from different departments contributed to the selection process. The author acknowledges the substantial contributions of all of those people, whose work this paper mainly represents. The author gratefully acknowledges the permission granted by the BBC for publication of this paper.

9 REFERENCES.

1. Walker, R. A new approach to the design of control room acoustics for stereophony. AES Preprint No. 3543 (G1-1), 94th AES Convention, Berlin, March 16-19, 1993.
2. Walker, R. Early reflections in studio control rooms: the results from the first Controlled Image design installations. AES Preprint No. 3853 (P12-6), 96th AES Convention, Amsterdam, February 26-March 01, 1994.
3. Walker, R. Controlled Image Designs: the management of stereophonic image quality. BBC R&D Report No. 1995/4.
4. Walker, R. Controlled Image Designs: the measurement of time-frequency responses. BBC R&D Report No. 1995/3.
5. Walker, R. Controlled Image Designs: results from the first installations. BBC R&D Report No. 1995/5.

RW
19/08/04

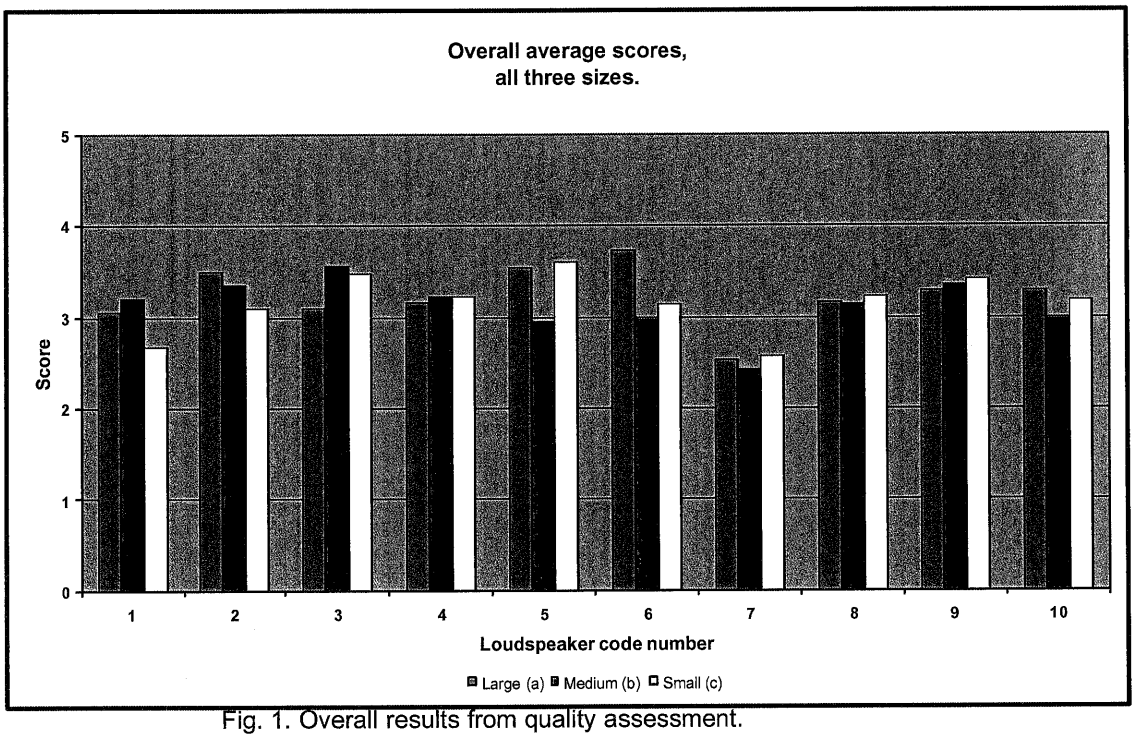

Fig. 1. Overall results from quality assessment.

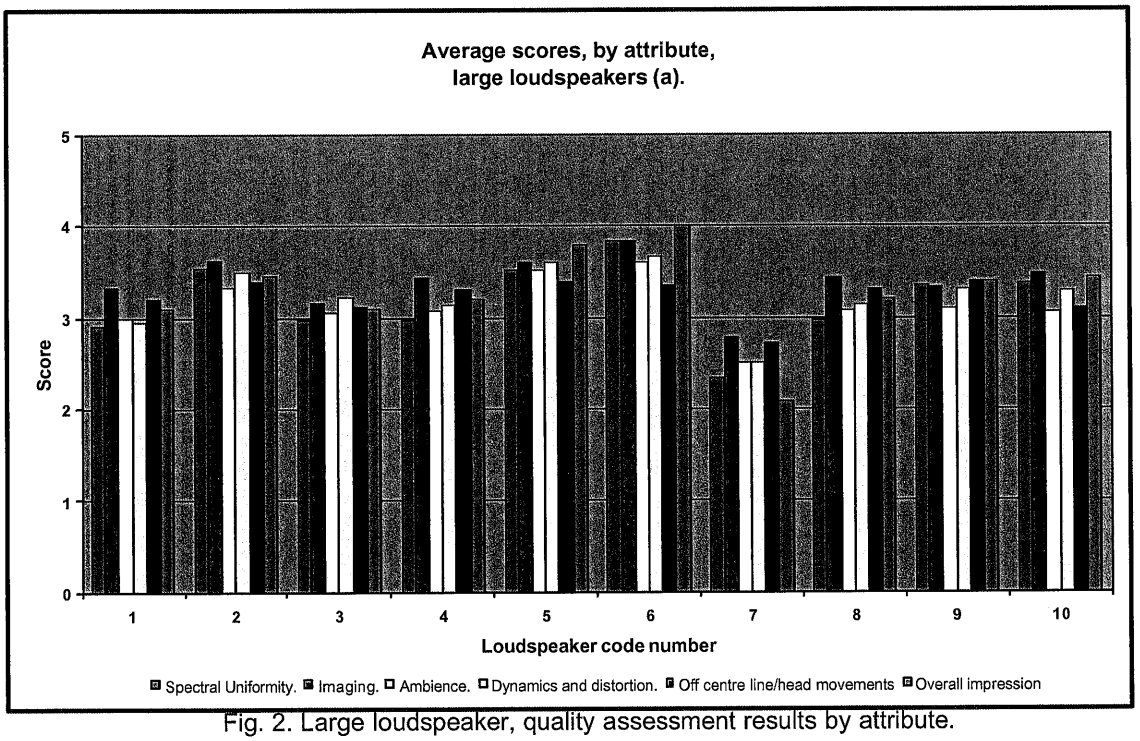

Fig. 2. Large loudspeaker, quality assessment results by attribute.

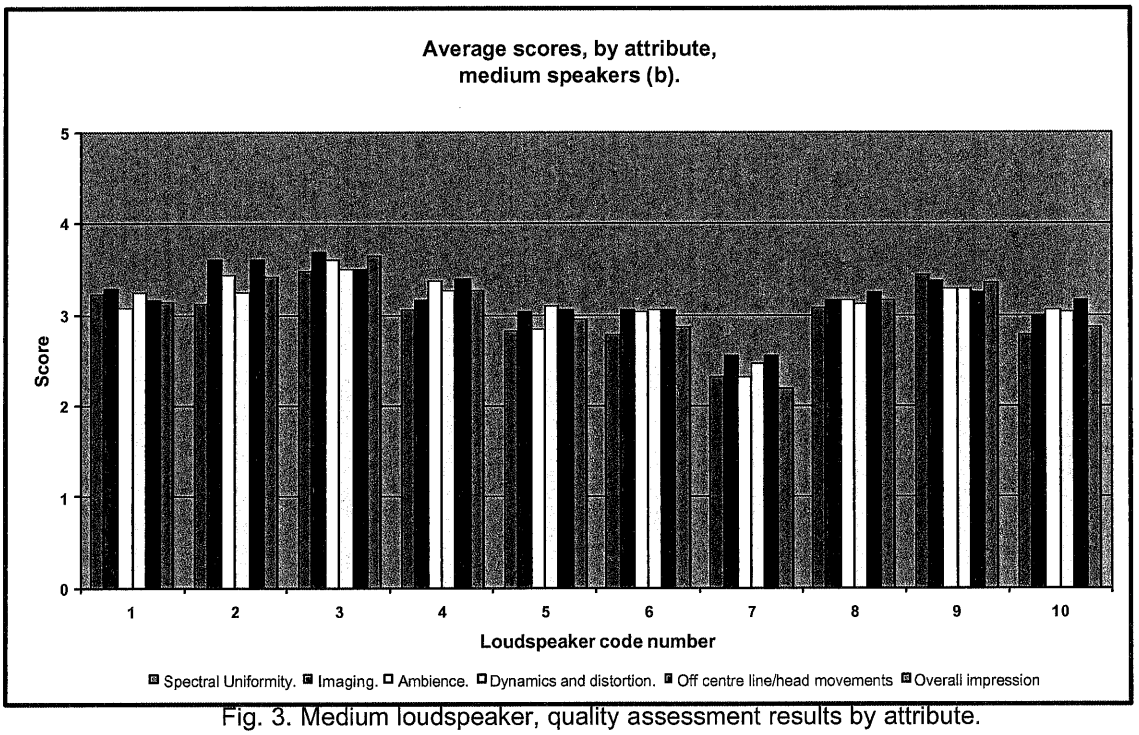

Fig. 3. Medium loudspeaker, quality assessment results by attribute.

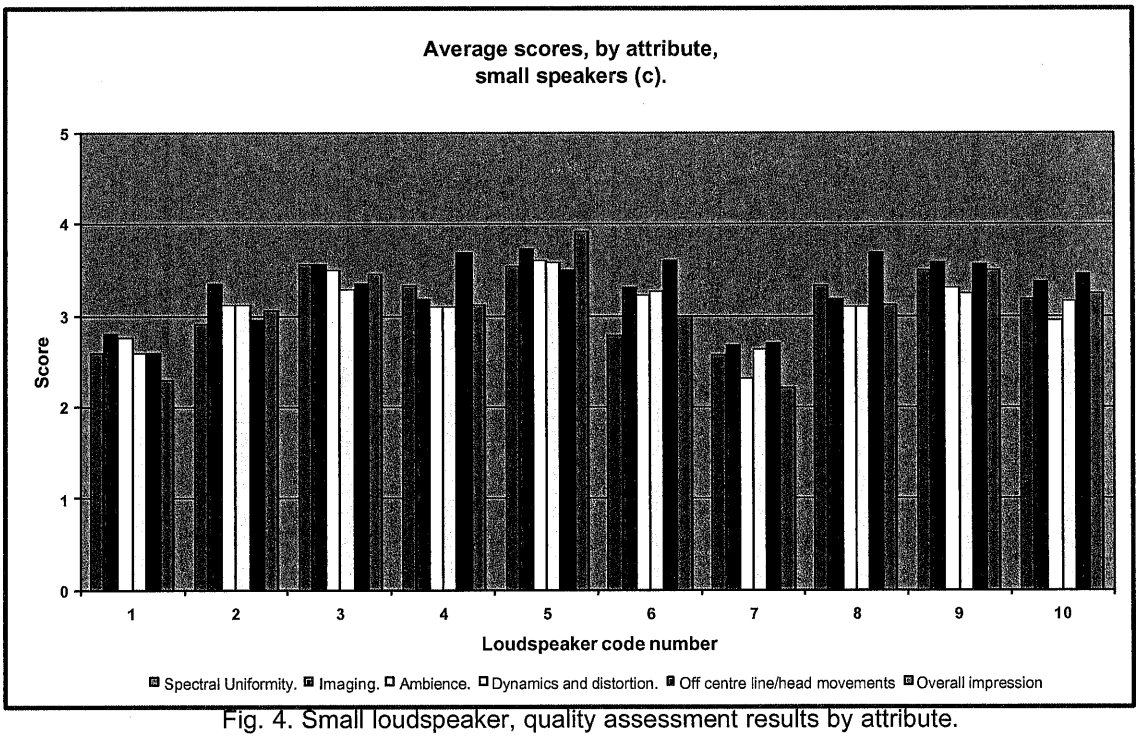

Fig. 4. Small loudspeaker, quality assessment results by attribute.