Audio Engineering Society
Convention Paper 5998
Presented at the 116th Convention
2004 May 8–11 Berlin, Germany

Unidimensional Simulation of the Spatial Attribute 'Ensemble Depth' for Training Purposes – Part 2: Creation and Validation of Reference Stimuli

Tobias NEHER, Tim BROOKES and Francis RUMSEY

Institute of Sound Recording, University of Surrey, Guildford, Surrey, GU2 7XH, UK
{t.neher, t.brookes, f.rumsey}@surrey.ac.uk

ABSTRACT

In the context of devising a spatial ear-training system, a study into the perceptual construct 'ensemble depth' was executed. Based on the findings of a pilot study into the auditory effects of early reflection (ER) pattern characteristics, exemplary stimuli were created. Changes were highly controlled to allow unidimensional variation of the intended quality. To measure the psychological structure of the stimuli and hence to evaluate the success of the simulation, Multidimensional Scaling (MDS) techniques were employed. Supplementary qualitative data were collected to assist with the analyses of the perceptual (MDS) spaces. Results show (1) that syllabicity[1] of source material (rather than ER design) is crucial to depth hearing and (2) that unidimensionality was achieved, thus suggesting the stimuli to be suitable for training purposes.

1. INTRODUCTION

Over the last decade or so, the audio industry has been witnessing a constant growth in the use of multichannel audio systems. Consequently, there has arisen the need to assess the performance of such systems in terms of their spatial quality. In this respect, it might be ideal to be able to use objective measures that correlate well with subjective phenomena of spatial sound display. These could yield accurately repeatable results and make the evaluation process time- and cost-effective. However, due to the fact that such measures are not available yet [1], experimenters have to resort to subjective testing methods.

It is widely acknowledged that treating humans as measuring instruments has a number of drawbacks. Humans are known to be highly variable in their judgements, which causes subjective evaluations to be inefficient and prone to unreliability. Hence, to be able to conduct valid and reliable sensory tests, it may be essential to minimise the variability in order to obtain meaningful data on which well-founded decisions can be made. That is why subjects need to be put in a frame of mind to understand the characteristics they are asked

[1] The term 'syllabicity' is used to denote a discontinuous or erratic amplitude envelope characteristic.

to measure, which can be achieved by controlled practice and training [2]. This is especially important if stimuli are to be evaluated in terms of several specific qualitative attributes, as the risk of confusion or different understandings of semantic meanings on behalf of the subjects is even higher in that case.

Training is commonly applied in a wide range of disciplines. By simulating the perceptual phenomena of interest, subjects can be exposed to and hence familiarised with the characteristics that they are required to assess at a later stage. However, it is self-evident that for an optimum training effect to occur such simulations need to be able to provide clear and unmistakable demonstrations of these subjective effects.

At the *Institute of Sound Recording* (IoSR) work is under way on the development of a spatial ear-training toolkit to be used for instructing naïve listeners in the assessment of multichannel audio systems. In two previous papers [3, 4] the spatial attributes of 'source distance', 'source width' and 'ensemble width' were investigated and algorithms for their unidimensional synthesis proposed and validated. In a third paper [5] the simulation of 'ensemble depth' (ED) – another component taken from a scene-based paradigm devised by one of the authors for the evaluation of spatial sound reproduction [6] – was begun. In particular, several characteristics of early reflection (ER) patterns were scrutinised so as to determine their perceptual relevance. The work reported herein concludes this study by addressing the creation and verification of exemplary ED sound excerpts. In addition, the previously adopted methodology for validating reference stimuli (or, in fact, any other supposedly 'artefact-free' algorithm) is refined.

2. THE MULTIDIMENSIONAL NATURE OF SPATIAL QUALITY

In this section a brief outline of the structure of spatial quality will be given to acquaint the unfamiliar reader with some elementary ideas and the terminology to be used in subsequent sections.

2.1. What constitutes spatial quality?

Sound quality has been assumed to be a multidimensional phenomenon for a long time. In the field of concert hall acoustics, researchers like Barron, Beranek and Schroeder identified and studied fundamental components such as timbre, loudness and spatial impression (e.g. [7, 8, 9]). Eventually, their findings were also adopted and scrutinised in the context of reproduced sound (e.g. [10]). Due to the increased interest in multichannel audio in the recent past, research has also been carried out in this area, proving that *spatial quality* itself has a complex perceptual structure, too. Two independently conducted studies [11, 12] seem to confirm the existence of several spatial characteristics, which are mostly descriptive (rather than attitudinal or emotive [13]) in nature. These describe discrete sound scene components including the distance, depth and width of single or groups of sources as well as spatial features of the reproduced environment. A detailed discussion of associated findings can be found in [6]. In Figure 2.1 some of these characteristics are depicted graphically.

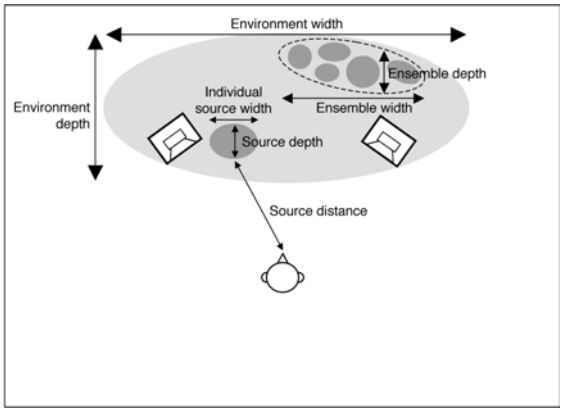


Figure 2.1: Graphical illustration of various spatial attributes as proposed in [6].

Note that the arrows indicating the various spatial components have different sizes to demonstrate the transition from the micro-level (i.e. individual sources, small arrows) to the macro-level (i.e. the entire environment, large arrows) with the ensemble-level (i.e. groups of sources, medium-sized arrows) in-between.

3. SPATIAL ATTRIBUTE SYNTHESISER

The ultimate objective of this project is to produce a tool that can be used to train listeners in the assessment of spatial sound displays. A stand-alone ear-training program, offering independent real-time control over each of the chosen spatial characteristics, is envisaged. As a first step towards this goal, the framework of a processing platform has been implemented with the help of the DSP-development software *MSP* [14]. The software runs on an *Apple Macintosh G4* computer and, in its current configuration, allows for the simultaneous processing of up to seven monophonic input signals. For each of these 12 early reflections can be rendered and individually adjusted in terms of level, delay and angle of incidence. This corresponds to all 1st and 2nd order specular reflections in the horizontal plane. Since the height dimension is not taken into account by most reproduction systems and the relevance of floor and ceiling reflections to perceived sound quality has not been investigated in detail yet, it was decided to ignore them because of the need for CPU 'housekeeping'. As regards the generation of specular reflections only, Martin emphasised the perceptual benefits of including a diffusion control in such a simulation [15]. Yet, he also acknowledged a resultant steep rise in computational cost, which is why this issue has been neglected so far. For each reflection order a biquad filter can be inserted into the signal path to imitate the effects of wall absorption. A 4-channel decorrelated reverberation stream is also computed, the level and decay time of which can be independently controlled in three separate, adjustable frequency bands. The reverb processor is a slightly modified version of an algorithm developed by Jot [16]. Different techniques can be employed for directionally encoding the direct sound and/or the early reflections. Currently, 5-channel pairwise constant power panning and a 4th order ambisonic panner [4, 17] are supported. At the reproduction end an ITU set-up [18] is used. A block diagram of the processing platform is shown in Appendix A. Due to its modular structure modifications and extensions can be accomplished fairly easily, as different requirements arise and more processing power becomes available.

4. CREATION OF STIMULI I

4.1. Source material

As part of an earlier study dealing with a spatial attribute representative of the ensemble-level category (see Section 2.1) speech recordings had been chosen for the source material, mainly because of the human voice's criticalness as a test signal [4]. Since the recordings were contextually unrelated, an inevitable and undesirable side-effect of this approach was that the subjects did not conceive of the speakers as an 'ensemble'[2]. Despite the fact that unidimensionality could still be achieved, it would be beneficial to this work if subjects not only perceived intra-source characteristics correctly, but also the sources' 'roles' within their intended frame of reference. This should make the concepts and definitions of the associated attributes more intuitive, hence resulting in a more effective training procedure.

Undoubtedly, the term 'ensemble' bears strong musical connotations. Therefore, non-musical signals seem less suitable for conjuring up an impression of a group rather than one of several discrete sources. Thus, for this experiment it was decided to use instrumental recordings, hypothesising that this would lead to a more unitary result with regard to the sources' psychological relatedness.

To reduce stimulus complexity as well as processing cost, an ensemble comprising four instruments only was envisaged. Yet, this was believed sufficient to evoke the cognitive cues necessary for producing the desired sensation. The ideal candidate for this job appeared to be a string quartet, as it seemed to exhibit all the desirable properties for an ensemble-level attribute simulation. Hence, a 4-bar recording (~8s in length) was used for this study, featuring the standard instrumental line-up of violin 1, violin 2, viola and cello. The instruments had been recorded separately under acoustically 'dry' conditions, thereby lending themselves to the superposition of a synthetically created acoustic context. Figure 4.1 displays a screenshot of the string quartet recording.

[2] This was evident from their verbal responses, which had been collected as part of the associated validation experiment (see also Section 5.3), i.e. no descriptors related to the group as a whole were used by the listeners.

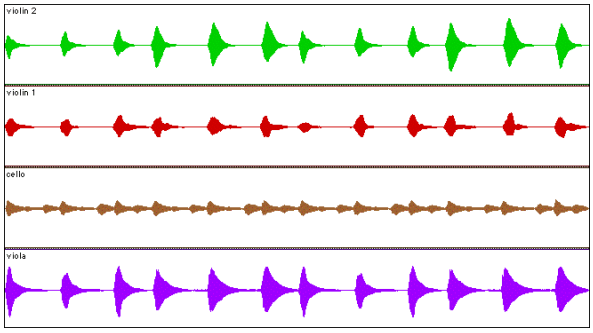


Figure 4.1: Screenshot of *string quartet* source material

4.2. Simulation strategy

In principle, altering the distance of one component source at a time is sufficient to deepen/flatten an ensemble. However, to increase the likelihood that listeners perceive depth as opposed to just relative source distance changes, it might be helpful to move all sources simultaneously. This assumption was based on the experience that when presented with complex stimuli, subjects tended to focus on whatever they perceived to change first rather than 'scanning' the whole sound scene for differences. Suffice it to say that this could lead to an oversimplified representation of the perceptual organisation of the stimuli to be generated and validated as part of this study. Therefore, to counter such auditory complacency the decision was made to have two pairs of sources being displaced from a straight-line (or 'flat') arrangement with one dyad (the outer sources) gradually getting closer relative to the listener and the other one (the inner sources) progressively getting more distant. Even though it was anticipated that this would increase the difficulty of the listeners' task, it was considered an appropriate measure to prevent the auditory system from undertaking a superficial analysis. At the same time, it was felt that the source grouping should enable listeners to detect the pattern more easily, which seemed less likely with all instruments moving independently. Figure 4.2 illustrates the applied notion graphically.

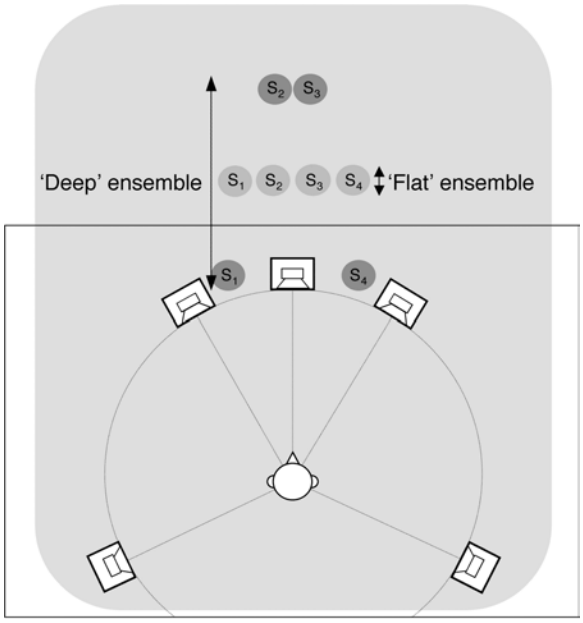


Figure 4.2: Graphical illustration of the concept of *ensemble depth* as simulated for this study

4.3. Applied processing

While source distance hearing has been the subject of considerable psychoacoustic research, little work has been done so far to specifically investigate depth perception. It is well known that the acoustic range stimulus depends on several independent physical parameters, including the level of the direct sound (DS), the direct-to-reverberant sound ratio (D/R) and spectral changes (e.g. see [19]). Evidently, to perceive ensemble depth the human hearing system has to deal with the acoustic information of at least two sources simultaneously. In terms of (additional) salient cues, some researchers have claimed that early lateral reflections are of paramount importance to depth hearing [20, 21]. Yet, as part of a pilot study designed by the authors to test this claim [5] it was found that the finer structure of ER patterns is not crucial to ED hearing. In any case, to allow for the perceptibility of depth in a reproduced sound scene, the audio engineer faces the challenge of having to harmonise numerous psychophysical factors. These, in turn, have to be brought into agreement with the respective cognitive processes (see Section 4.1) as well to be able to evoke the envisaged spatial impression. Thus, it is probably not too far-fetched to assume that the depth attribute

comes close to timbre in terms of physical and perceptual complexity (e.g. [22]).

The first step in the generation of the ED sounds involved determining the maximally possible 'dynamic range' of the DS levels. That is, it was ascertained how quiet the inner instruments could be made in relation to the outer ones whilst still being localisable. For this purpose, source material properties were exploited as much as possible, i.e. the first violin and the cello were placed in the centre of the ensemble, since they were melodically (violin 1) and rhythmically (cello) distinctive and thus less susceptible to being masked by the other two instruments. The resultant level difference (16dB and 19dB, respectively) was then subdivided into nine steps as required for the validation phase. To achieve subjectively equal step sizes it was found necessary to change the DS levels in a non-linear manner. More precisely, the levels of the inner instruments had to be reduced more as they receded into the distance. This substantiates the findings of other researchers, e.g. the 'distance compression' phenomenon described by von Békésy [23] or Begault's [19] finding that, in the absence of other cues, listeners prefer a 9dB over a 6dB difference for the sensation of half distance.

In addition, changes were made in terms of the lateral positioning of the instruments to incorporate a perceptual cue in the simulation commonly utilised in 2-dimensional imagery. Converging lines in a 2-D drawing convey parallel lines and hence depth in three dimensions. This *linear perspective* phenomenon forces the brain to automatically infer a 3-D context on the basis of such information being contained in the 2-D input of the retina [24]. Attempting to emulate this aspect aurally, a given DS was increasingly lateralised as the corresponding source appeared to get closer. To avoid any extra, unwanted qualitative changes, the ambisonic panner was used for the spatial encoding, because it had previously been found to be superior to pairwise constant power panning in terms of image width and timbre constancy across the frontal sector [4].

In order to provide the listeners with a supplementary distance/depth cue and hence to make the simulation more vivid, a first-order low-pass filter was inserted into the signal paths of the inner instruments. This had the effect of gradually rolling off the high-frequency (HF) content of the direct sounds, reaching a maximum cut-off frequency of 8kHz for the deepest stimulus. Thus, the applied processing was broadly in line with data published in [25] where the magnitude response of a typical air absorption filter was shown to be 3dB down at 10kHz for a source distance of 10m.

Regarding the design of the ER patterns, reflection levels were adjusted so as not to impair the previously established DS changes. In particular, it was found that for each instrument the ER levels had to follow the one of the associated DS closely. This can be explained with the help of the well-known precedence effect. Reflections arriving within about 50ms after the DS are perceptually combined to allow the human hearing system to localise a source in the direction of the first wavefront. As a result of the sound energy being integrated over this time window, the impression of added loudness arises [26]. Since under natural acoustic circumstances the DS obeys the inverse distance law whereas the combined energy of all ERs decreases less than 6dB for the same (physical) distance change, the desired increase in (subjective) source range was maintained by reducing the early sound energy more or less in parallel with each DS.

The numbers of ERs were chosen so that the resultant listening conditions roughly resembled a real-world situation. For the closest sources, five reflections were reproduced within the distance-salient time window of 15ms to 50ms [27] whereas for the deepest stimulus the inner instruments exhibited 12 reflections. By and large, the temporal distribution of the reflections was uniform. Since during the first part of this study it had been found that listeners are unable to distinguish between ERs panned in accordance with a room model and ones that are reproduced by their nearest loudspeakers [5], the latter approach was applied here for reasons of simplicity. However, overall directional characteristics of ER patterns were still replicated, i.e. the greater the apparent distance of a source, the more reflections were reproduced from in front of the listening position. It is true that no empirical evidence is available, which proves the perceptual salience of this parameter to depth hearing. Nonetheless, since such changes occur under normal listening conditions, their inclusion was expected not to be harmful to this simulation either. Similarly, it was decided to equip each source with a different reflection pattern. Depending on the reflection order the ERs had different frequency spectrums, the two biquad filters in the block diagram shown in Appendix A being used as (second-order) low-pass filters with cut-off frequencies of ~4.5kHz and ~3.5kHz, respectively.

Finally, diffuse reverberation was added, whereby its level and duration ($T_{60} \approx 1.6$s) were adjusted to create a

room size impression that complemented the largest envisaged source distance. Depending on the source material, the chosen reverb levels were slightly different for the individual instruments. In particular, the more sustained cello part needed a few extra dB of reflected energy compared to the other sources with a fairly erratic amplitude envelope (see Figure 4.1) for the same SD effect due to its DS tending to mask the background stream information [28] more. Further, minor intra-source adjustments of the reflected sound level were made as the stimuli got deeper. Although a diffuse sound field is characterised by approximately constant reverberant energy irrespective of measurement position [26], the reverb levels had to be balanced against the corresponding direct sounds in order to maintain adequate localisability.

With the help of the processing platform described in Section 3, nine (see Section 6.1) reference stimuli were created, each illustrating a different degree of ED. The applied processing was highly controlled so as to enable unidimensional changes in the chosen quality.

5. VALIDATION EXPERIMENT I

5.1. Listening panel

Once a set of test stimuli had been generated, a validation experiment was conducted to verify whether the intended unidimensionality had been achieved. For this purpose 12 final-year students and one graduate of the *University of Surrey's* 'Tonmeister' degree course were employed as the listening panel. Four of them were female. As part of their education the students had received considerable training in the detection of small changes and impairments in sound quality. Also, all of them had participated in psychoacoustic tests before. Although the listeners were not checked for normal hearing before doing the test, 10 of them had been screened as part of another research project carried out at the IoSR [29]. The rationale for using such experienced and critical listeners was that if their responses did not contain references to any unwanted differences, the stimuli would almost certainly be free of unwanted artefacts. Hence, unidimensionality of the simulation could be inferred, thereby authorising its use for training programmes. All subjects participated on a voluntary basis, i.e. none of them was remunerated for their time. The whole experiment lasted for a minimum of 35min, but not more than 70min. No information about the nature of the experiment was given to the listeners until the test had been completed.

5.2. Physical set-up

The experiment took place in an ITU-R BS.1116 [30] listening room. Listening test software written by the first author in *MaxMSP* [14] was used to automate the experiment and save the subjects' responses to hard disk. The software was run on an *Apple Macintosh G4* computer equipped with a *Digidesign 001* soundcard whose *ADAT* digital output was connected to a *Yamaha 02R* mixer for D/A conversion. Five *Genelec 1032A* loudspeakers were set up at 0°, ±30°, and ±110° and a distance of 2.1m from the optimal listening position. The loudspeakers were level aligned to within 0.2dBA of each other using a pink noise test signal and a *Brüel & Kjäer 2123* real-time spectrum analyser. The computer monitor was positioned directly in front of the listening position, so that the subjects could control the speed of the listening test and switch between the stimuli at their leisure. To eliminate the influence of any visual cues on the subjects' judgements, the listening room was darkened and an acoustically transparent curtain was hung from the ceiling to conceal the position of the loudspeakers. In addition, subjects were encouraged to listen with their eyes closed. A diagram of the experimental set-up is included in Appendix B.

5.3. Experimental design

As in the earlier studies, MDS was used as the main sensory analysis tool, because it is a relatively neat way of verifying whether an intended qualitative effect has been achieved or not. MDS requires each stimulus in a given group to be compared with every other stimulus of the same group. Since it is an attribute-free technique, subjects do not make comparisons with respect to highly subjective and hence potentially misinterpreted verbal descriptors of a certain quality. Rather, all stimulus pairs are assessed in terms of their overall similarity. This is beneficial in that subjects do not have to try to understand and adopt pre-specified attribute scales. The collected similarity judgements are then transformed into Euclidean distances, which in turn are represented in multidimensional (Euclidean) space [31]. For instance, the rated degree of dissimilarity of stimulus i and j is mathematically modelled as:

$$d_{ij} = \left[\sum_{r=1}^{R} \left(x_{ir} - x_{jr} \right)^2 \right]^{1/2}$$

where x_{ir} is the coordinate of stimulus i on the r-th dimension and R is the number of dimensions in the Euclidean space.

As a result of the virtual absence of semantics and all its inherent ambiguity [32], MDS has a reputation for being a relatively bias-free method for measuring human perception. Nevertheless, MDS techniques have a number of limitations, e.g. the unravelled psychological structure needs to be interpreted. That is to say that MDS cannot provide the meanings or labels for these perceptual dimensions. Instead, they have to be found by other means. What is more, MDS only uncovers those dimensions that are continuous and orthogonal with respect to each other. Put differently, if two subjective effects are directly correlated[3] or a dimension exists that is specific to only one or a few stimuli, these extra characteristics will not show up separately in the MDS space. To give an example, as a group of sources might spread out laterally, some of the individual sources themselves could broaden as well. Thus, a change in perceived 'source width' would occur in tandem with a change in perceived 'ensemble width' (see Figure 2.1). However, these two distinct qualities would only be reflected as a single (continuous) dimension in the MDS analysis. While this aspect of *(multi)collinearity* has been discussed in the context of other statistical techniques (e.g. multiple regression analysis [31]), it appears to have been neglected as far as MDS is concerned.

In order to address the shortcomings described above, it was decided to supplement the MDS results with both verbal and non-verbal responses. Other researchers at the IoSR have looked at the advantages and disadvantages of these two types of data [32]. In this respect, graphical elicitation techniques were found to be especially useful for investigating the spatial attributes of image width, location and skew [33]. For that reason, following the attribute-free comparison required for the MDS analysis, all listeners were asked (1) to verbally express the differences they had perceived between the stimuli and (2) to depict their verbal responses graphically (whenever relevant).

[3] It is assumed that the two factors would also have to be linearly related for this to be true. A, say, linear increase for one and a quadratic increase for the other factor should be disclosed by MDS.

In the case of (1) a questionnaire was provided. Listeners were encouraged to write down words and descriptions that were differential in nature. Moreover, to quantify the perceptual salience of each perceived difference, they had to grade them on a scale from 1 to 10. A '1' was defined to correspond to a subjective effect being *just about audible* while a score of '10' was specified to imply that a particular difference was *the only subjective effect* perceivable between all nine stimuli. Intermediate anchor points were deliberately omitted, since they can cause problems with subject-dependent interpretations and ensuring linear scale increments [30]. Again, the listeners were offered the opportunity to listen to all nine sound excerpts if they felt they had to.

In the case of (2) all participants were given an A4-sized sheet of paper displaying an outline of their surroundings as an indication of scale. The listeners' task was to draw their perception of the changes in ED as faithfully as possible. Due to the fact that this concept cannot be easily described using a few words only (see Section 4.1) it was anticipated that the drawings would help them explicate their verbal responses. Crucially, it was only after the completion of both the MDS and verbal reporting stages that subjects were told that they had to draw what they had heard.

As was already stressed above, MDS cannot reveal distinct perceived attributes if they (1) vary in parallel along a single continuous dimension or (2) are stimulus-specific. Yet, it is of great importance for the simulation, if a single dimension is identified, that this dimension results from perceived variance in the desired property only. Otherwise the simulation would be flawed and optimal listener training could not be guaranteed. The last 'quality control' step in each listening test therefore was to tell the listeners that the stimuli were intended to vary only in the subjective quality that they had graded highest, and to ask them whether, knowing this, anything "sounded wrong". In this context, it is worth emphasising that throughout the whole experiment the authors' chosen descriptor 'ensemble depth' was never mentioned so as to prevent distortion of the subjects' wordings. For the same reason, care was taken to only use each subject's own terminology when discussing his/her responses for clarification purposes.

Based on the experimental design outlined above, it was hoped that a single strong dimension would be revealed allowing the unidimensionality of the stimuli to be concluded. Since there were nine (see Section 6.1)

stimuli to compare, each subject made a total of 36 gradings. A different order of presentation was created for each listener to minimise carry-over effects. The test was subdivided into three groups of 12 trials. After the completion of each group subjects were offered a short break in an attempt to reduce listener fatigue. The user-interface implemented in the listening test software is shown in Figure 5.1. As can be seen, a scale featuring numeric labels (ranging from 1 to 9) was provided for the subjects to indicate their perceptions, hoping that these visual anchors would help them be consistent in making their verdicts.

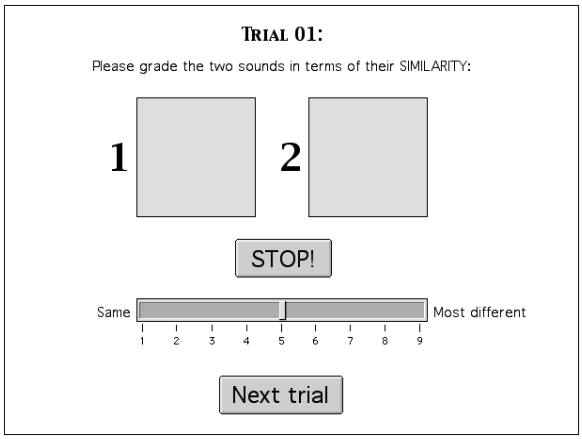


Figure 5.1: User-interface used for MDS experiment

Using written instructions, subjects were informed that they had to make global similarity judgements taking into account any and all detected differences when grading a pair of sounds. The scale provided ranged from 'Same' (or '1') to 'Most different' (or '9'). Listeners were told to give the 'Most different' grading to those sounds that appeared to be the two most dissimilar ones out of the whole group. In order for them to get an idea of the range of differences, the subjects were given the opportunity to listen to all nine sounds before and halfway through each group of 12 trials. Hence, they were able to familiarise themselves and refresh their memories with respect to the maximally possible differences between the stimuli.

In addition, all but two participants were acquainted with the task, the user-interface and the scale by means of a short training session. This was expected to help minimise the error variance in their judgements. The training comprised a comparison of nine stimuli that were different from the ones to be validated afterwards followed by six trials. Before the start of the proper test it was made sure that no questions about the experimental procedure remained.

6. RESULTS

6.1. Analysis of MDS data

The number of stimuli used during the grading phase and the format of the collected similarity data govern what types of statistical analyses are feasible. In a previous paper [3], a detailed account of the relationships between these factors and the chosen MDS layout was given, which is why only the most essential points and considerations will be reiterated here.

With the help of the statistical analysis package *SPSS* [34], non-metric MDS solutions were obtained for the set of data. As a rule of thumb, for a given solution to be stable more than four times as many stimuli as dimensions are required [31]. While this does not prevent the experimenter from executing analyses for higher dimensionalities, violation of this guideline means that results have to be regarded as very tentative until replicated with more stimuli [35]. Since for this experiment listeners had compared a set of nine stimuli, 1- and 2-dimensional, statistically robust solutions could be derived, therefore permitting the unfolding of a second perceptually relevant factor. This was sufficient for the purpose of this study, i.e. to either approve or disapprove the envisaged unidimensionality of the sound excerpts.

To assess dimensionality the 'measures of fit' calculated by the MDS procedure were examined. Measures of fit are non-statistical parameters, which express how well a given model represents a set of raw data. In the case of the *ALSCAL* routine implemented in *SPSS* [36], these are 's-stress' and 'RSQ'. S-stress ranges from 1 (worst possible fit) to 0 (perfect fit). RSQ, the squared correlation index, can be interpreted as the proportion of variance accounted for (VAF) by the MDS model [31]. Although it is desirable to maximise the VAF of a given solution, the maximal number of dimensions taken into account needs to be limited, especially if the increase in explained variance per dimension is less than ~0.05 [37]. This is because dimensions with a low

contribution to the explained variance are difficult to explain and are likely to be associated with noisy data.

In Figure 6.1 a so-called 'scree plot' is shown, displaying s-stress as a function of dimensionality. For the sake of completeness, a second "badness-of-fit" parameter, 'stress', has also been included, which differs from s-stress in that it is reported in terms of linear rather than squared distances.

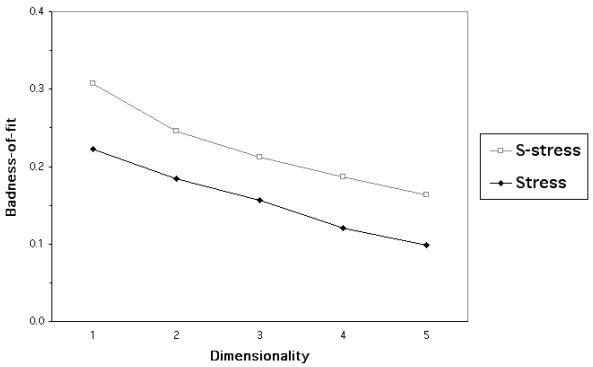


Figure 6.1: Scree plot (experiment 1)

As can be seen, both stress and s-stress decrease monotonically with each additional dimension. However, this comes as no surprise since, as a matter of fact, both parameters will always get smaller if dimensionality is increased, even if the conditions for a stable analysis are not satisfied. In spite of this problem, the scree plot is commonly inspected to see whether a point is apparent beyond which the decrements in the chosen badness-of-fit metric begin to be less pronounced. Several statisticians have argued that such a "knee" corresponds to the dimensionality that should be chosen (e.g. [37, 38]). The reasoning for this choice is that the knee marks the point where MDS uses additional dimensions to essentially only scale the noise in the data, after having succeeded in representing the systematic structure in the given dimensionality [39].

Revisiting Figure 6.1, a slight kink is apparent from the s-stress curve at 2-D whereas no such indication is given by the stress parameter. Nevertheless, since SPSS optimises s-stress and not stress [36], one might tend to decide that a 2-dimensional solution is appropriate to model the structure contained in the data. Yet, as has been argued before [4], the authors believe this statement to be oversimplified. This is because an apparent knee at 2-D does not rule out unidimensionality of the stimuli, because a real knee located at 1-D would not be identifiable as such. So to be precise, an apparent knee at the second dimension on its own cannot resolve whether a set of data contains one or two discrete perceptual characteristics.

Evidently, the situation is not clear, which is why it makes sense to evaluate these results in parallel with the ones obtained for the RSQ measure. In Table i the RSQ values derived for the 1-D and 2-D models are shown. As can be seen, the 1-D solution is characterised by a high RSQ value, which decreases as one goes to a 2-D solution. This indicates that the MDS algorithm struggles to find a systematic pattern in the set of data for a model with more than one dimension.

Dimensionality	RSQ
1	0.82
2	0.79

Table i: RSQ results obtained from non-metric MDS analysis (experiment 1)

Thus, by conjointly examining the different measures of fit, one is tempted to deduce that the panel identified and employed a single perceptual factor when comparing the sounds. Support for this impression is also provided by the MDS model's representation of the stimuli's psychological structure. In Figure 6.2 the 'stimulus space' is shown, which is the result of aggregating the subjects' dissimilarity judgements and depicting them graphically as the 'psychological distances' between the stimuli. Those stimuli that the subjects rated to be similar appear as points close to each other whereas those stimuli judged to be dissimilar are distant from one another. Knowing that 'a' was created to be the flattest and 'i' the deepest stimulus, it can be seen that the sounds were perceived in the order intended. The spacings between each pair of stimuli are not constant, which might be due to inaccuracies during the generation stage and/or the subjects' inability to be consistent in their judgements. However, all sounds appear to have a different intensity with regard to a particular quality, thereby enabling the listeners to rank them correctly.

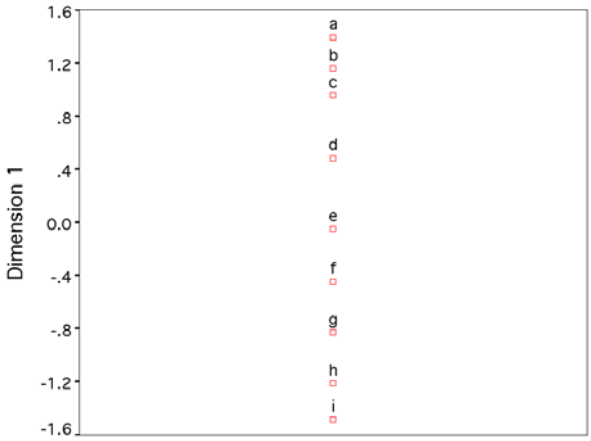


Figure 6.2: 1-dimensional MDS stimulus space (experiment 1)

On the whole then, the results seem to suggest that the MDS analysis managed to successfully uncover the envisaged unidimensional, perceptual organisation of the stimuli from the panel's responses.

6.2. Analysis of verbal data

Seeking confirmation for the apparently unidimensional structure of the ED stimuli, the verbal responses were examined. As was the case with the MDS analysis, the method for scrutinising the verbal data was similar to the one used previously [3]. The reader may recall that the motivation for collecting additional verbal (and graphical) data was to establish a basis on which to identify the fundamental dimension(s) used for evaluating the stimuli. Apart from encouraging them to think of differential terms, the subjects had not been instructed to comply with any particular response format in order not to bias their responses. As a result of this free verbalisation approach, the data were fairly diverse and hence needed to be structured. For this purpose a form of Verbal Protocol Analysis (VPA) [40] was employed. Put simply, VPA is a methodology that allows classification of verbal descriptors of certain properties into different groups.

At the first level of analysis, the data were separated into two categories, one for holistic and one for analytical terms. The distinction was based on whether the subjects' responses were directly related to a perceptual phenomenon as a whole (i.e. high-level descriptors) or whether they described more specialised and/or technical aspects related to signal properties influencing the formation of a certain perception (i.e. low-level descriptors). As the experimenter has to interpret the meaning of the subjects' verbalised perceptions, there is a risk of biasing the outcome by condensing the data into fewer and fewer groups. By limiting the classification process to two stages, an attempt was made to restrict distortion of the meanings of the responses as much as possible. To obtain an index of perceptual importance for each group of terms, a total weight factor was calculated. Since the subjects had graded the perceived differences on a scale from 1 to 10, the scores of all verbal descriptors within each group were added up. The result was then divided by 130, giving a maximum possible weight factor of 1.

To the authors' disappointment the results did not reflect the intended subjective effect. With regard to the analytical terms, five groups were identified, which are shown in Table ii. The first group ('Relative level of sources') constitutes the strongest and hence subjectively most significant one out of all the categories that were found for this experiment. Ten participants identified and used this difference for making their similarity judgements, four of which rated it as the perceptually dominant effect, resulting in an overall weight factor of 0.48. Incidentally, this value is fairly low compared to the highest ones obtained from previous studies for which unidimensionality was accomplished (e.g. 0.82 [4]). It is also worth pointing out that six listeners perceived the level variations of the two pairs as being antipodal, while two subjects specified hearing balance changes only for the outer instruments and the other two exclusively for the cello.

Analytical groups	Occurrences	Weight factor
Relative level of sources	10	0.48
Audibility of reverb	4	0.15
Attack/punchiness/edginess of notes	3	0.15
Reverb level of sources	2	0.05
Overall level	1	0.02

Table ii: Groups of analytical terms and their relative weights (experiment 1)

The second group is similar to the first one in that it also seems to delineate the applied processing, i.e. it contains references to the audibility of the reverberation. However, it is characterised by a small weight factor of 0.15, as is the third group, which comprises terms and phrases describing properties of individual notes. It is surmised that this subjective effect goes hand-in-hand with the first one, i.e. an increase in source level leads to more 'punchy' or 'edgy' notes. The last two categories ('Reverb level of sources' and 'Overall level') are directly related to the stimulus creation process again, but because of their very low weight factors (0.05 and 0.02, respectively) this is no reason to be cheerful.

The results from systematising the holistic terms are displayed in Table iii. Generally speaking, references to changes in the width of the ensemble were made most often. Three out of the seven listeners who noticed this change used these very wordings to describe it, the remaining ones preferring the terms "Lateral position of instruments", "Stereo distribution of instruments", "Widening of stereo image: violin 2 and viola move outwards" and "Width of overall image". Although the last response could also describe a scene- rather than an ensemble-related auditory aspect, the associated graphical data (see Section 6.3) indicates left-right variations in the source positions, thus justifying its inclusion in this category. Five listeners perceived this to be the dominant subjective difference between the stimuli. Nonetheless, a weight factor of 0.34 does not indicate an unequivocal perceptual effect either.

Holistic groups	Occurrences	Weight factor
Ensemble width	7	0.34
Relative source distance	5	0.23
Ensemble distance	1	0.06
Room size	1	0.05
Source width	2	0.04

Table iii: Groups of holistic terms and their relative weights (experiment 1)

The second holistic category finally concurs with the envisaged qualitative change a bit more. Five listeners detected changes in the apparent distance of the instruments. Curiously, for one such listener "Relative distance of instruments" was the strongest auditory effect. What is more, he also perceived the converse movement in almost the intended way, as evident from his graphical response (see below). Yet, with a total weight factor of 0.23 the 'Relative source distance' group can hardly be called salient. Moreover, two other participants reported hearing this type of difference only for the cello or mainly for the central instruments, respectively. Parenthetically, only one listener used the descriptor "depth" in this context to delineate her perception of the stimuli. The other three categories are composed of terms that were either mentioned only once or that were given low 'magnitude of audibility' scores. One listener effectively stated that the ensemble as a whole appeared to get closer or further away ("Distance of quartet"), thus failing to work out the diametric movement of the two instrument dyads. Regarding the other groups, it is speculated that 'Room size' concerns an effect similar to the one denoted by the (analytic) 'Audibility of reverb' group, hence being related to the applied reverb processing. The latter may also be blameable for the 'Source width' category, because a decrease in the D/R will tend to de-focus a source and vice versa [5].

To summarise the findings of this section, as the analysis of the verbal data clearly showed the generated stimuli failed to impart the intended quality to the panel. First and foremost, it is striking that there is no single strong category even though this is exactly what one would have expected as a result of the MDS solution. Instead, the above discussion leads to the following inferences:

1. Subjects picked out and concentrated on different characteristics of the sound excerpts – most notably level and direction of the DS – when forming their verdicts, but none of the detected effects really prevailed.

2. The various ED 'ingredients' identified by the listeners were correlated, which is why a unidimensional MDS representation of the stimuli's auditory structure was possible.

3. The additional, verbally elicited information was fundamental in laying open the perceptual deficiencies of the simulation. Without this type of data, an undistorted reading of the MDS results would not have been possible[4].

[4] This aspect has been largely ignored by perceptual researchers who have utilised MDS techniques for their studies.

On the positive side of things, this experiment showed that the employed validation strategy is sensitive to unwanted dimensions; if they 'sneak into' the simulation, they will be revealed. Besides, it is satisfying to see that four listeners made use of the words "ensemble" or "quartet" when verbalising their perceptions, denoting that the choice of musical programme material was helpful in moving away from a 'several discrete sources' perception.

6.3. Analysis of graphical data

Since five listeners either had effectively stated that spatial differences had been secondary to their verdicts or had not noticed any spatial changes at all, they were not asked to depict their perceptions visually. The eight graphical responses that were obtained were scrutinised by means of visual inspection. Specifically, it was checked whether the spatial information contained in the drawings agreed with the implied meanings of the verbal data. Curiously, those five listeners who had graded ensemble width changes most strongly, drew the ensemble in such a way that the sources spread out on an arc maintaining approximately constant distance to the listening position. In other words, they subconsciously included depth changes in their graphical depictions. In Figure 6.3 one example of such a response is shown. When questioned about this, some subjects replied that they would expect a string quartet to be arranged like this in a performance situation. The possibility of larger source spacings along the front-back axis, on the other hand, was ruled out in this respect. This suggests that consideration of cognitive aspects can also be harmful if the aim is to stimulate a specific impression. Due to their pre-conditioning, subjects might unintentionally suppress any potentially noticeable deviations from their internal portrayal of a particular percept. Furthermore, a sixth listener also included a clear front-back separation of the instruments in his sketch, which was absent from his verbal responses. This example (once again) illustrates the usefulness of graphical elicitation, because it can help subjects explicate their perceptions, especially when being presented with a fairly complicated set of stimuli like the one used in this case.

Yet, even though the above observations make the results look a bit more promising, the variations in ED evident from the drawings were admittedly smaller than the ones in perceived ensemble width.

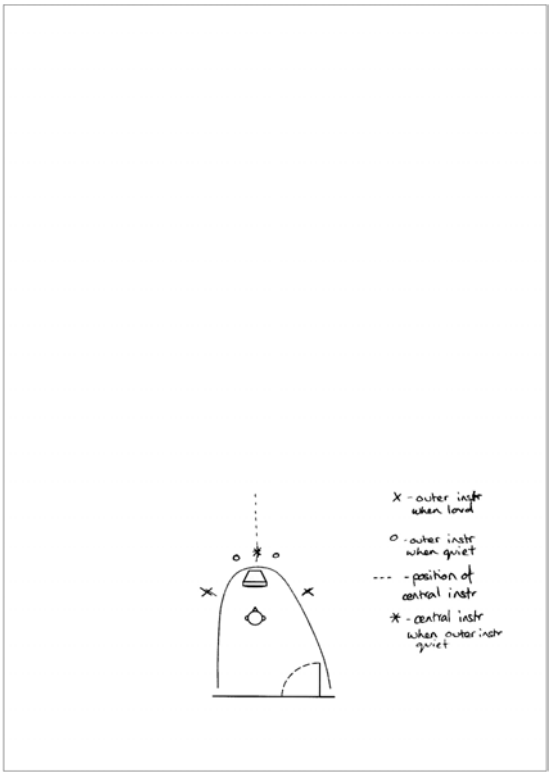


Figure 6.3: Graphical response sheet displaying the visible surroundings and a typical listener response (experiment 1)

6.4. … but does it sound right?

As mentioned in Section 5.3, for the final step of the verification methodology it had been planned to inform all listeners that the stimuli were intended to vary only in the subjective feature that they had perceived as being the most prominent difference. Next, they were meant to reflect upon their aural experience and to decide whether alteration of the sounds was required to ensure total compliance with their *own* understanding of the concept. It is self-evident that this step had to be modified, because considerable discrepancies between the envisaged and perceived result were apparent from the verbal data of most listeners. Therefore, they were questioned about their listening strategies so as to find out what had handicapped depth detection on their behalf. As it turns out, several listeners had based their similarity judgements solely on the behaviour of the two outer instruments and not paid attention to the centre of the sound image. A possible explanation for this may be that the four string instruments were too similar in terms

of their physical as well as musical characteristics (see Figure 4.1), thus making their perceptual segregation more difficult. Since the outer sources were generally louder, they probably became the listeners' reference points within each auditory scene. For the same reasons it is assumed that subjects did not hear the reverberation enough, causing them to deduce level rather than distance changes. This would explain the predominance of verbal references to changes in the direct sounds (i.e. direction and level) over responses taking into account reflected sound energy.

Having obtained all this feedback, it was decided to create new stimuli. This was imperative, because the 'string quartet' examples had turned out to be totally unsuitable for conveying the ED attribute to a group of experienced listeners in an unambiguous fashion.

7. CREATION OF STIMULI II

Rather than starting from scratch, the experience gained from the first experiment was capitalised on by refining the ED simulation. This basically involved two steps: finding better source material and modifying the applied processing. As for the simulation strategy, it was decided to maintain the contrary SD changes together with the source groupings. Despite the first simulation effort being a failure, this feature was still believed to be beneficial for helping listeners identify the depth of the band being the variable for the same reasons as outlined in Section 4.2. The rationale for changing the source material and modifying the processing as well as the actual refinements are described below.

7.1. Source material

Regarding the choice of suitable source signals, it was decided to stick with musical programme material as the string quartet stimuli had made the subjects listen in a perceptually more integrative manner compared to the speech sounds used for a previous simulation [4]. Yet, in an attempt to improve the localisability of the individual sources and therefore depth changes, different types of instruments were picked this time to increase the sources' dissimilarity with respect to their physical properties. In particular, an 8-bar recording (~14s in length) of a funk band featuring a tenor saxophone, a double bass and two electric guitars was employed. Again, the instruments had been recorded separately and contained very little reflected sound. By selecting a fairly unusual combination of instruments it was hoped that subjects would be less prejudiced in terms of the expected instrumental layout, thereby perhaps being more tolerant towards an unnaturally large front-back spacing. Special care was also taken that the musical arrangement had sufficient 'space', i.e. the instruments were not all constantly active to guarantee the audibility of reverberant decays. Figure 7.1 shows a screenshot of the 'funk band' recording.

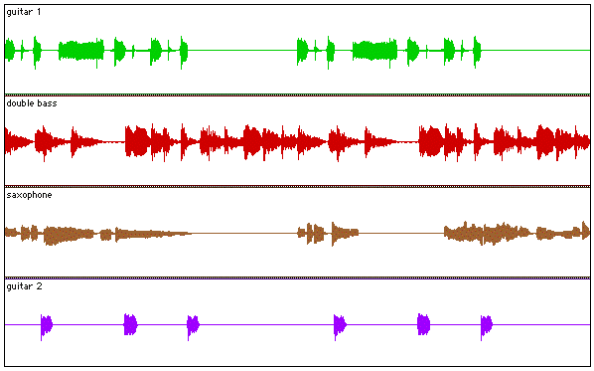


Figure 7.1: Screenshot of *funk band* source material

7.2. Applied processing

The procedure for generating the new sound examples was generally very similar to the one outlined in Section 4.3. A welcome side-effect of choosing miscellaneous instruments was that it allowed an increase in 'dynamic range' by at least 3dB per instrument. Consequently, the step sizes in the DS level (and hence ED) could be made a bit bigger. Since the double bass and saxophone had the busiest parts, they were placed in the centre of the ensemble so as to draw the listeners' attention to the distance changes whenever there would be a 'gap'. Also, the DS step sizes of the guitars were slightly reduced for the three deepest stimuli, i.e. when the guitars came closest. This was meant to hinder them from being perceptually too prominent. It was feared that otherwise the loudness changes would dictate the listeners' perceptions again and hence impede the detection of a depth dimension. Likewise, since subjects had turned out to be very receptive to lateral source position changes, differences in the direction of the direct sounds were minimal this time.

To further raise the panel's awareness with respect to the variations in source proximity, the reverberation time (RT) was increased to 2s. As a result, reflected energy was longer audible and hence more difficult to 'miss'. Additionally, the reverb level of the two guitars was gradually raised by ~2dB as they appeared to get closer. Otherwise, their direct sounds would have been psychologically overwhelming, causing the guitars' estrangement with respect to the otherwise reverberant conditions. It is supposed that this was the case for the first experiment during which subjects had not noticed any range but only loudness changes for the string instruments moving towards the listening position.

Finally, the stimuli were fine-tuned by the authors and another expert listener. Previously, this had been done by just two of the authors, but the problematic results from the first experiment and the large spread of responses commonly reported for absolute distance estimation studies [41] seemed to demand this measure. Hence, to take no chances the various sound field parameters described above were adjusted until a common ground had been reached with regard to the subjective functioning of the envisaged ED notion.

7.3. Listening panel, physical set-up and experimental design

Seven final-year Tonmeister students, five members of the IoSR and one professional audio engineer – all experienced in assessing reproduced sound – served as the listening panel for this test. Eight of these 13 subjects had previously undergone an audiometric examination [29] and ten were trained with respect to the experimental procedure. Due to the fact that five listeners had also taken part in the first validation experiment, one might argue that they were biased. However, it has to be borne in mind that after the completion of the first experiment listeners were not told about the aim to achieve unidimensionality, because considerable differences between the perceived and intended effect had been found (see Section 6.4). Also, as the results of the first validation study clearly showed, the dominant effect detected by the panel was width- rather than depth-related. Therefore, in terms of subject preconditioning it is assumed that in the worst case listeners may have expected the presentation of spatial sound stimuli.

The physical set-up and experimental design were identical to the ones described in Section 5.2 and 5.3, respectively. Depending on the listener, this test took between about 40min and 70min to complete.

8. RESULTS II

8.1. Analysis of MDS data

The similarity data were treated and analysed in exactly the same way as before. In Figure 8.1 the resultant scree plot is shown. The first thing to note is that the s-stress curve also exhibits a knee at 2-D, but this time it is more pronounced than the one evident in Figure 6.1. Secondly, the stress measure seems to be in better agreement with the s-stress metric, i.e. it also appears to hint at a 2-dimensional solution being the appropriate one for the new set of responses. However, the knee is still slight and, as was pointed out in Section 6.1, does not exclude a sharper but invisible knee at 1-D. At any rate, evaluated with the previous results in mind, these observations imply that, for the same dimensionality, the MDS algorithm managed to find a clearer structure in these data compared to the ones from the first experiment.

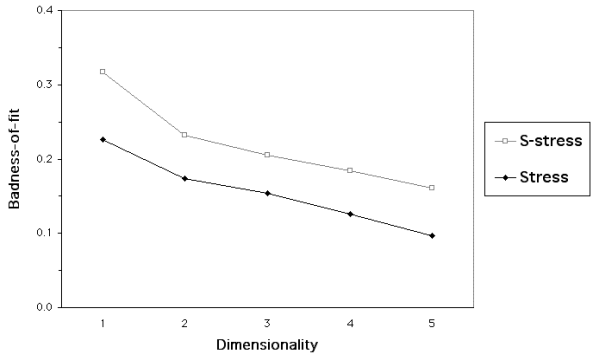


Figure 8.1: Scree plot (experiment 2)

The results for the explained variance can be seen in Table iv. Regarding the 1-D solution, there is hardly any difference compared to the RSQ value obtained for the first experiment, i.e. it has an almost equally large magnitude of 0.81. This time, though, augmenting the MDS model by a second dimension does not cause the RSQ to decline, but since there is no increase either no incentive is given for exploring its origin.

Dimensionality	RSQ
1	0.81
2	0.81

Table iv: RSQ results obtained from non-metric MDS analysis (experiment 2)

The outcome of visualising the sound excerpts' psychological distances in the form of the 1-dimensional stimulus space (Figure 8.2) reveals that the stimuli are all sufficiently different from each other in terms of (at least) one subjective effect, because they appear in the correct order again. Interestingly, the sound examples seem to be compressed at the flat (stimulus 'a') end of the simulation. Going back to Figure 6.2, the same tendency is apparent, which suggests that subjects found it generally harder to discern stimuli when all sources had roughly the same D/R. Further work would have to be done in order to trace back the root of this phenomenon.

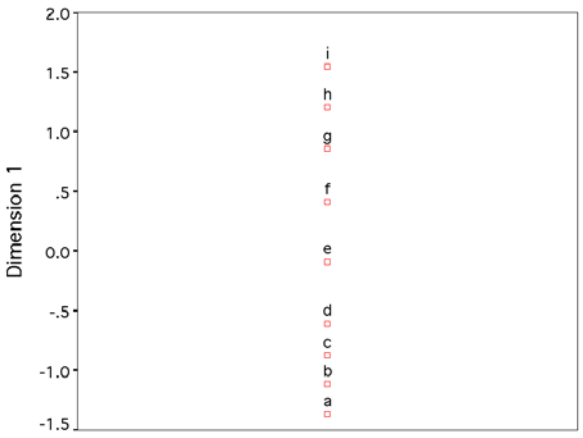


Figure 8.2: 1-dimensional MDS stimulus space (experiment 2)

All in all, the findings of the MDS analyses of the two sets of similarity judgements are very much alike, except that this time the badness-of-fit measures attest a more obvious rating pattern, as apparent from applying the 'knee criterion'.

8.2. Analysis of verbal data

Classification of the verbal responses according to the VPA procedure employed previously turned out to be more difficult than before. This was because listeners sometimes reported the same qualitative phenomenon for different (groups of) instruments. To give an example from the analytical terms, both intra- as well as inter-source level changes were detected again, but this time some listeners described level differences for the (outer) guitars and the (inner) saxophone separately. Consequently, the associated 'Relative level of sources' category (see Table v) can be considered somewhat inflated. Nonetheless, both the occurrences and the weight factor are lower compared to the first experiment. Inter-listener disparities were also found again in terms of the instruments that were perceived to be influenced by this type of difference; while most listeners did not specify any particular instrument(s), one subject noticed it in the case of the saxophone and three others solely for the guitars.

Analytical groups	Occurrences	Weight factor
Relative level of sources	8	0.38
Spectral changes	6	0.18
Reverb level of sources	3	0.09
Overall level	1	0.03

Table v: Groups of analytical terms and their relative weights (experiment 2)

The second analytical bin comprises all terms concerning spectral modifications. Generally speaking, listeners either stated that the guitars were affected or they did not particularise whether these differences related to the source, ensemble or scene level. In terms of perceived frequency range, the descriptors "Brightness of individual instruments", "Timbre: dull vs. bright", "Bassiness of guitars" and "Close sounds were harsher" were elicited, suggesting that HF changes were most audible. As was the case with the first analytical group, the third category ('Reverb level of sources') had also been encountered already. In terms of perceptual salience, it is slightly more important this time (presumably because it was more audible due to the longer RT and the new source material), as evident

from its marginally higher weight factor. The same applies to the last group ('Overall level'), too.

The holistic groups are displayed in Table vi. Satisfactorily, the first group contains descriptors reflecting the desired ED construct. What is even better, it is also characterised by a large weight factor of 0.84. As was the case with the (analytic) 'Relative level of sources' category, some subjects deliberately distinguished between the distance/proximity of inner/outer (groups of) instruments when asked to enumerate the perceptual differences they had used in forming their similarity verdicts[5]. This is likely to be a side-effect of the chosen simulation strategy, i.e. the opposite movement of the two source dyads. All but one listener noticed changes pertinent to this category, 10 of them giving them the highest score out of all their verbalisations. Four listeners (who perhaps perceived the stimuli in a more unitary way) used descriptors such as "Sense of perspective" or "Depth", which are clearly preferable in terms of the aims of this study.

Holistic groups	Occurrences	Weight factor
Ensemble depth	14	0.84
Ensemble distance	2	0.13
Source direction	4	0.07
Togetherness vs. disjoint-ness of ensemble	1	0.03
Elevation of guitars	1	0.03
Envelopment	1	0.02

Table vi: Groups of holistic terms and their relative weights (experiment 2)

The next category has a familiar look to it as well. The single listener who had not perceived any relative source distance alterations, heard changes in the proximity of the whole ensemble. The other mentioning of 'Ensemble distance' stems from a participant who noticed this effect in addition to "Individual source distance" variations. In terms of physical sound field manipulation, the 'Source direction' group is clearly related to 'Ensemble width', which represented the strongest holistic group of the first experiment. It is assumed that because the panning was substantially reduced and hence less noticeable, listeners verbalised these changes as being source- rather than ensemble-related. Trailing in fourth place, the 'Togetherness vs. disjointness of ensemble' group is a bit awkward to categorise, because of its more profound nature. However, it seems to circumscribe the envisaged effect in a holistic manner, which is why it was included here. The last two groups are not easily accounted for, but on the grounds of their very small weight factors they can be dismissed as being perceptually insignificant.

From the point of view of having to argue the case for unidimensionality, it is also important to point out that the holistic descriptors "sense of perspective" and "depth" were elicited four times during the second experiment. This compares favourably with the one mentioning of "depth" for the first test (see Section 6.2). Besides, two listeners clarified their responses by stating that they detected distance or depth changes for those pairs of stimuli sufficiently different from each other. In contrast, for two very similar sound examples these types of (holistic) variations were not perceptible as such. Instead, the listeners relied on changes in the directions of the guitars and the reverb level of the saxophone to determine if the stimuli were identical. This seems to imply that at least some of the analytical terms may have arisen due to the small in/decrements in ED between adjacent stimuli. Put differently, there may be some kind of a perceptibility threshold for an (ostensibly) unidimensional percept like ED.

8.3. Analysis of graphical data

Even though the visible surroundings had been delineated on the response sheet to help the subjects draw to scale, clear inter-listener differences in overall depth are apparent from the graphical responses. As a rule of thumb, the magnitude of the source distance variations was drawn to be unexpectedly small[6], albeit much bigger than any other spatial attribute. Notwithstanding, this finding should not be given too much weight, as humans are notoriously bad at judging absolute source range [41]. More importantly, no additional systematic change is evident from the

[5] This explains the 14 occurrences in this category, even though there were only 13 listeners.

[6] Bearing in mind that an RT of ~2s and a ~20dB difference in DS level were used, the authors themselves perceived (and hence would have also anticipated) a maximal source distance well beyond the physical constraints of the listening room.

responses, thus substantiating the apparently unidimensional structure of the similarity and verbal data. According to their graphical depictions, some listeners perceived the inner sources as being closer to the listening position for the (supposedly) flat extreme, while for other subjects it was the other way round. Similarly, some responses showed more variation in the apparent range of the saxophone and (occasionally) the bass compared to the guitars, whereas for others it was the exact opposite. An example of such a graphical response is given in Figure 8.3.

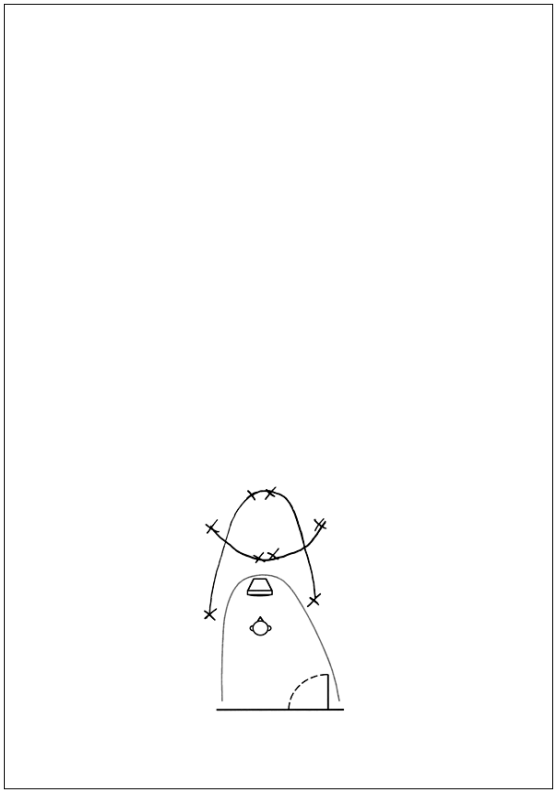


Figure 8.3: Graphical response sheet displaying visible surroundings and a listener response showing converse movement of the inner and outer instrument pair (experiment 2)

Further analysis of the sketches showed that, in terms of source locatedness, the double bass was problematic. To be more precise, two listeners did not perceive any changes in its position, while two others depicted its subjective location in form of a large, fuzzy region. In the case of one response the bass is missing entirely. In retrospect, these results are not really surprising. Since the transient sound information (e.g. the attack portion of a plucked or rattling string) was likely to be masked when the other instruments were playing (see Figure 7.1), listeners only heard the low-frequency content of the bass for most of the time. As a result, an impoverished set of localisation cues may have been available to the panel for this instrument.

8.4. … but does it sound right?

Having completed the similarity rating, verbal and graphical elicitation stages, the ED simulation was subjected to the final 'quality control' step. When asked to specify any deficiencies inherent in the set of stimuli as well as how the simulation would have to be improved, three listeners complained about the sporadic nature of the individual musical parts, stating that it made intra-source comparisons difficult. Another subject suggested that detection of the depth changes would have been easier if only one source had been moved at a time. Both of these problems had been anticipated when creating the sound excerpts, but it was decided to put up with them for reasons outlined in Section 4.1 and 7.1, respectively. In any case, they do not exclude the achievement of unidimensionality. Two other comments implied that the bass was difficult to localise when it was distant and that its movement was minimal. Even though congruence in the movement of the bass and saxophone had been aimed for during the generation process, it is not essential for being able to demonstrate the concept of ED. Since all but one listener perceived the saxophone and the guitars in the intended way, ED changes could still be conveyed. Hence, it can be concluded that, by and large, the stimuli "sounded right".

8.5. INDSCAL analysis

To gather further evidence for the nonexistence of a second meaningful dimension the data were submitted to an Individual Differences Scaling (INDSCAL) analysis. INDSCAL can be considered a derivative of MDS, because it also calculates the co-ordinates of a group of stimuli on a number of perceptual dimensions common to a set of similarity judgements. The result is then displayed in the stimulus space (e.g. see Figure 8.2). However, in contrast to MDS, INDSCAL acknowledges that subjects may differ in how they form

their verdicts and therefore tries to take such individual differences into account. More precisely, it models inter-subject agreement *as well as* disagreement, separating those factors common to a group of subjects from those in which the subjects differ [39]. That is why INDSCAL can provide a quantitative characterisation of the individual differences that exist within a panel, which are captured as subject-specific weights placed upon each of the INDSCAL dimensions [37]. These weights are commonly portrayed in the 'subject space'. Mathematically speaking, INDSCAL is very similar to MDS, the distance between stimuli i and j for subject n being defined as follows:

$$d_{ijn} = \left[\sum_{r=1}^{R} w_{nr} \left(x_{ir} - x_{jr} \right)^2 \right]^{1/2}$$

where x_{ir} is the co-ordinate of stimulus i on dimension r and w_{nr} is the weight (required to be non-negative) for dimension r associated with subject n.

In Figure 8.4 the 2-dimensional subject space of a non-metric INDSCAL analysis executed on the listeners' dissimilarity judgements is displayed. In interpreting this (or any other) subject space it is important to note that the origin of this space is not arbitrary, but has a fixed meaning [42]. The *distance* of a subject from the origin corresponds, at least roughly, to the variance accounted for in the data from that subject. This means that if a subject's point is precisely at the origin, no variance at all is accounted for. The *direction* of a subject from the origin relates to the pattern contained in the data from that subject. Therefore, two subjects who lie on the same straight line issuing from the origin would have identical configurations except for a single overall scale factor. One subject's being closer to the origin on that line would indicate simply that less of the variance in his/her data is accounted for by that common configuration, either because his/her data are noisier or because additional dimensions are needed to explain the subject's judgements fully.

Inspecting the SD subject space, one can see that the weight placed upon Dimension 1 is greater for all listeners, denoting that it was perceptually more important to them than Dimension 2. Thus, this is in line with the previous finding that the participants identified and used the same main difference for grading the stimuli.

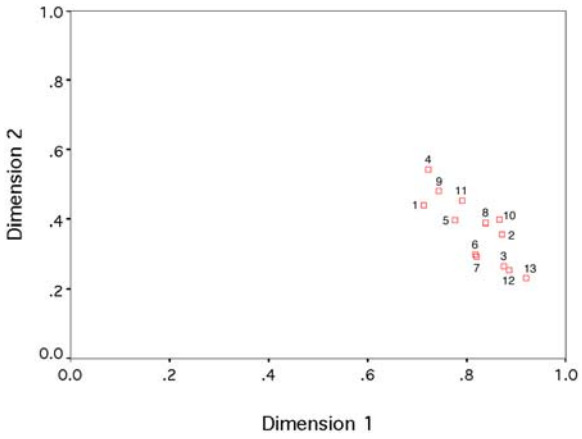


Figure 8.4: 2-dimensional INDSCAL subject space (experiment 2)

As to the second dimension, the previous analyses gave the impression that no meaningful, additional perceptual factor is contained in the data. To ascertain whether this is true or not, one can search for correlations between the listeners' verbal responses and their positions along Dimension 2. Basically, the verbal data are mapped onto the subject space to see if any inter-listener agreement exists. Fortunately, this is not the case. An example for the meaninglessness of Dimension 2 is evident from Subject 9. According to his verbal responses, the listener experienced just a single audible difference, which he verbalised as "Relative distance of instruments: guitars vs. sax and bass". However, his weight along the second dimension is larger than the one of almost all of the other panellists. This is a strong sign that, at least in the case of Subject 9, Dimension 2 constitutes noise. Thus, in effect, the subject space has the potential to act as some kind of a noise gate that can 'mute' those signal parts unrelated to the input to the auditory system, i.e. those dimensions, which are the result of listeners making inconsistent similarity judgements. However, another signal is required that can be used to set the 'gating threshold' correctly as well as 'trigger' the gate – a verbal response.

In contrast to Subject 9, the other participants verbalised between two and six differences each, many of which are directly related to the intended spatial percept or at least the applied processing. Enough evidence is therefore available to deduce that the panel as a whole did not identify a second perceptual dimension in the ED stimuli.

8.6. Concluding remarks

From the above it should be apparent that no type of collected data can disclose the dimensionality of the stimuli if examined in isolation. Due to the fact that the subjects were not in any way restricted when verbalising their perceptions, these responses are particularly problematic to interpret, albeit crucial to the outcome of this experiment. It is true that, like the dissimilarity judgements and graphical responses, the semantic data potentially document inter-listener similarities and differences. Yet, as was demonstrated in Section 8.2, 8.3 and 8.5, the various elicited wordings appear to have a common underlying meaning with respect to the one perceptual difference identified by almost all listeners.

The problem of eliciting and interpreting verbal information from subjects for the sake of attribute identification was also discussed by Berg and Rumsey [13]. Referring to Shaw and Gaines' work [43], they stated that subjects may share only parts of their terminology and conceptual systems. Thus, listeners might use the same term for different concepts, different terms for the same concept, the same term for the same concept, or use different terms and have different concepts. These four possible scenarios are summed up in Figure 8.5. Evidently, the 'Correspondence' quadrant would be pertinent to the findings of this study. However, it is surmised that, following training with well-defined reference samples, subjects might move closer to 'Consensus', owing to a clear definition of verbal terminology and its relation to the stimuli.

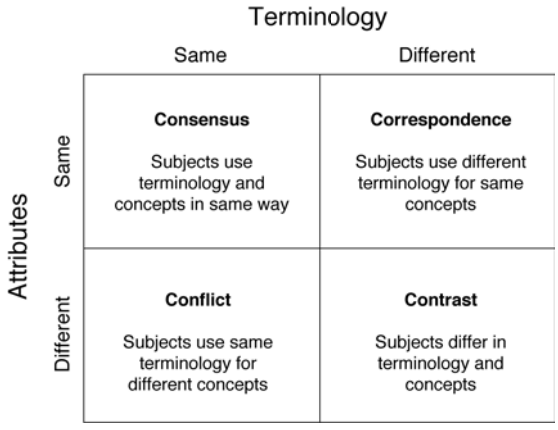


Figure 8.5: Relationships between terminology and attributes (after [43])

9. SUMMARY AND CONCLUSIONS

The primary objective of this study was to create exemplary ensemble depth stimuli that would vary in a unidimensional way to allow them to be used for training naïve listeners in spatial sound assessment. The development of an appropriate simulation method therefore formed a major part of this work. In this respect, the choice of source material was found to be crucial to the perceptibility of a depth component. More precisely, syllabicity on both an intra- as well as inter-source level was needed to ensure the audibility of each instrument's reflected sound. In contrast, controlled manipulation of the finer temporal and spatial structure of reflections turned out not to be critical to enable listeners to perceive depth changes.

As the above discussion clearly showed, MDS on its own cannot guarantee that all perceptual factors contained in a set of similarity ratings are revealed. Some form of verbal elicitation is needed in order to be able (1) to interpret and label the uncovered continuous dimensions and (2) to discover qualitative factors varying in parallel to these. For this purpose, the collection of additional graphical data may also be beneficial, especially if a complex psychological phenomenon is investigated that is difficult to describe in words.

Conjoint evaluation of the dissimilarity judgements and the verbal and non-verbal data could overcome these deficiencies of MDS. In particular, it was found that the generated stimuli illustrated the intended ED effect to experienced listeners in an unambiguous manner. Importantly, no additional qualitative factors were detected by the panel as a whole. Thus, it can be concluded that unidimensionality of the sounds was achieved, thereby suggesting them to be suitable for training purposes.

10. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Rhonda Wilson of *Meridian Audio Ltd.* for her valuable contributions to this paper, *Meridian Audio Ltd.* for financial support and all the listeners for their time and patience.

11. REFERENCES

[1] Mason, R., Rumsey, F. (2001): 'Interaural time difference fluctuations: Their measurement, subjective perceptual effect, and application in sound reproduction', *Proceedings of the AES 19th International Conference on Surround Sound*, Schloss Elmau, Germany, June 21-24, pp. 252-271

[2] Meilgaard, M., Civille, G. V., Carr, B. T. (1991): *Sensory Evaluation Techniques*, Boca Raton, Florida: CRC Press

[3] Neher, T., Rumsey, F., Brookes, T. (2002): 'Training of listeners for the evaluation of spatial sound reproduction', *Audio Engineering Society Preprint*, 112th Convention, preprint no. 5584

[4] Neher, T., Rumsey, F., Brookes, T., Craven, P. (2003): 'Unidimensional simulation of the spatial attribute 'ensemble width' for training purposes', *Audio Engineering Society Preprint*, 114th Convention, preprint no. 5769

[5] Neher, T., Brookes, T., Rumsey, F. (2003): 'Unidimensional simulation of the spatial attribute 'ensemble depth' for training purposes – Part 1: Pilot study into early reflection pattern characteristics', *Proceedings of the AES 24th International Conference on Multichannel Audio*, Banff, Canada, June 26-28, pp. 123-137

[6] Rumsey, F. (2002): 'Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm', *Journal of the Audio Engineering Society*, vol. 50, no. 9, pp. 651-666

[7] Barron, M., Marshall, A. H. (1981): 'Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure', *Journal of Sound and Vibration*, vol. 77, no. 2, pp. 211-232

[8] Beranek, L. (1996): *Concert and Opera Halls: How They Sound*, Woodbury, NY: Acoustical Society of America

[9] Schroeder, M. R., Gottlob, D., Siebrasse, K. F. (1974): 'Comparative study of European concert halls: Correlation of subjective preference with geometric and acoustic parameters', *Journal of the Acoustical Society of America*, vol. 56, no. 4, pp. 1195-1201

[10] Gabrielsson, A., Sjögren, H. (1979): 'Perceived sound quality of sound-reproducing systems', *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 1019-1033

[11] Berg, J., Rumsey, F. (2000): 'In search of the spatial dimensions of reproduced sound: Verbal Protocol Analysis and Cluster Analysis of scaled verbal descriptors', *Audio Engineering Society Preprint*, 108th Convention, preprint no. 5139

[12] Koivuniemi, K., Zacharov, N. (2001): 'Unravelling the perception of spatial sound reproduction: Language development, Verbal Protocol Analysis and listener training', *Audio Engineering Society Preprint*, 111th Convention, preprint no. 5424

[13] Berg, J., Rumsey, F. (2000): 'Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction', *Audio Engineering Society Preprint*, 109th Convention, preprint no. 5206

[14] See http://www.cycling74.com/ for details.

[15] Martin, G. (2001): 'A hybrid model for simulating diffuse first reflections in two-dimensional synthetic acoustic environments', *PhD Thesis*, Faculty of Music, McGill University

[16] Jot, J.-M. (1997): 'Efficient models for reverberation and distance rendering in computer music and virtual audio reality', *Proceedings of the International Computer Music Conference*, Thessaloniki, Greece, pp. 236-243

[17] Craven, P. G. (2003): 'Continuous surround panning for 5-speaker reproduction', *Proceedings of the AES 24th International Conference on Multichannel Audio*, Banff, Canada, June 26-28, pp. 261-266

[18] ITU-R BS.775 (1994): 'Multichannel stereophonic sound system with and without accompanying picture', *International Telecommunication Union – Radiocommunication*, Geneva, Switzerland

[19] Begault, D. R. (1987): 'Control of auditory distance', *PhD Thesis*, University of California, San Diego

[20] Griesinger, D. (2001): 'The psychoacoustics of listening area, depth and envelopment in surround recordings and their relationship to microphone technique', *Proceedings of the AES 19th International Conference on Surround Sound*, Schloss Elmau, Germany, June 21-24, pp. 182-200

[21] Pellegrini, R. S. (2002): 'A virtual reference listening room as an application of auditory virtual

environments', *PhD thesis*, Institute of Communication Acoustics, Ruhr-University of Bochum, Verlag im Internet GmbH, Berlin

[22] Krumhansl, C. L. (1989): 'Why is musical timbre so hard to understand?', In: Nielzén, S. and Olsson, O., eds., *Structure and Perception of Electroacoustic Music*, Amsterdam: Elsevier, ISBN: 0444811052, pp. 43-53

[23] Von Békésy, G. (1949): 'The moon illusion and similar auditory phenomena', *American Journal of Psychology*, vol. 62, pp. 540-552

[24] Shepard, R. (2001): 'Cognitive psychology and music', In: Cook, P. R., ed., *Music, Cognition and Computerized Sound: An Introduction to Psychoacoustics*, Cambridge, MA: The MIT press, ISBN: 0262531909, pp. 21-35

[25] Savioja, L., Huopaniemi, J., Lokki, T., Väänänen, R. (1999): 'Creating interactive virtual acoustic environments', *Journal of the Audio Engineering Society*, vol. 47, no. 9, pp. 675-705

[26] Everest, F. A. (2001): *Master Handbook of Acoustics*, New York: McGraw-Hill

[27] Theile, G. (2001): 'Natural 5.1 music recording based on psychoacoustic principles', *Proceedings of the AES 19th International Conference on Surround Sound*, Schloss Elmau, Germany, June 21-24, pp. 201-229

[28] Griesinger, D. (1997): 'The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces', *Acustica*, vol. 83, no. 4, pp. 721-731

[29] Zielinski, S., Rumsey, F., Bech, S. (2003): 'Comparison of quality degradation effects caused by limitation of bandwidth and by down-mix algorithms in consumer multichannel audio delivery systems', to be presented at the *114th Audio Engineering Society Convention*, Amsterdam, March 22-25

[30] ITU-R BS.1116 (1997): 'Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems', *International Telecommunication Union – Radiocommunication*, Geneva, Switzerland

[31] Hair, J. F., Anderson, R. E., Tatham, R. L., Black, W. C. (1998): *Multivariate Data Analysis*, Upper Saddle River, New Jersey: Prentice-Hall

[32] Mason, R., Ford, N., Rumsey, F., de Bruyn, B. (2001): 'Verbal and nonverbal elicitation techniques in the subjective assessment of spatial sound reproduction', *Journal of the Audio Engineering Society*, vol. 49, no. 5, pp. 366-384

[33] Ford, N., Rumsey, F., de Bruyn, B. (2001): 'Graphical elicitation techniques for subjective assessment of the spatial attributes of loudspeaker reproduction – A pilot investigation', *Audio Engineering Society Preprint*, 110th Convention, preprint no. 5388

[34] See http://www.spss.com/ for details.

[35] Kruskal, J. B., Wish, M. (1978): *Multidimensional Scaling*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-011, Beverly Hills: Sage Pubns.

[36] Norûsis, M. J. (1994): *SPSS Professional Statistics™ 6.1*, Chicago, Illinois: SPSS Inc.

[37] Martens, W. L., Zacharov, N. (2000): 'Multidimensional Perceptual Unfolding of Spatially Processed Speech I: Deriving Stimulus Space Using INDSCAL', *Audio Engineering Society Preprint*, 109th Convention, preprint no. 5224

[38] Schiffman, S. S., Reynolds, M. L., Young, F. W. (1981): *Introduction to Multidimensional Scaling: Theory, Methods and Applications*, New York: Academic Press

[39] Borg, I., Groenen, P. (1997): *Modern Multidimensional Scaling: Theory and Applications*, New York: Springer-Verlag

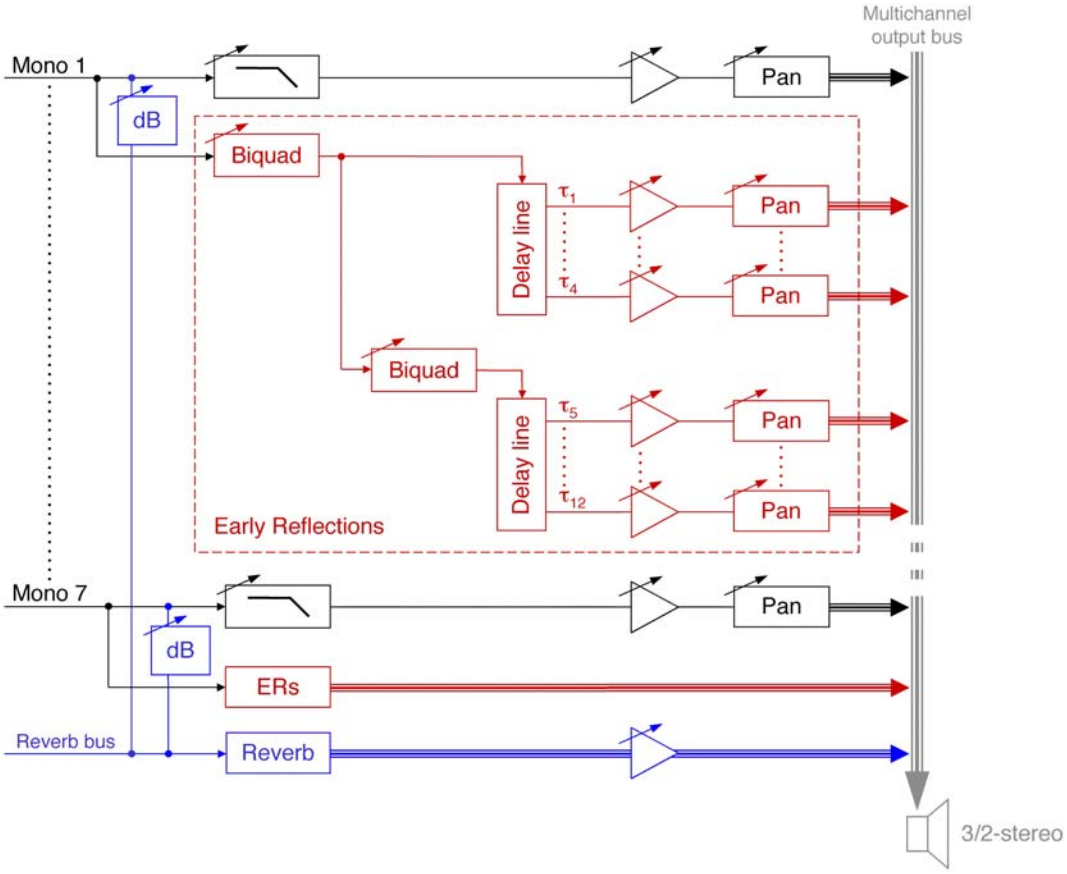
[40] Ericsson, K. A., Simon, H. A. (1993): *Protocol Analysis: Verbal Reports as Data,* Cambridge, MA: The MIT Press

[41] Nielsen, S. H. (1993): 'Auditory distance perception in different rooms', *Journal of the Audio Engineering Society*, vol. 41, no. 10, pp. 755-770

[42] Carroll, J. D., Chang, J.-J. (1970): 'Analysis of individual differences in Multidimensional Scaling via an n-way generalization of "Eckart-Young" decomposition', *Psychometrika*, vol. 35, no. 3, pp. 283-319

[43] Shaw, M., Gaines, B. (1995): *Comparing Conceptual Structures: Consensus, Conflict, Correspondence and Contrast*, Knowledge Science Institute, University of Calgary

APPENDIX A: BLOCK DIAGRAM OF PROCESSING PLATFORM



APPENDIX B: EXPERIMENTAL SET-UP