Audio Engineering Society

Convention Paper 5584

Presented at the 112th Convention
2002 May 10–13 Munich, Germany

Training of listeners for the evaluation of spatial sound reproduction

Tobias Neher, Francis Rumsey, and Tim Brookes

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

ABSTRACT

This paper presents some preliminary results from an ongoing study into methods for the training of listeners in subjective evaluation of spatial sound reproduction. Exemplary stimuli were created illustrating two spatial attributes: individual source width and source distance. Changes in each of the two attributes were highly controlled in an attempt to allow unidimensional variation of their perceptual effects. The stimuli were validated with the help of an experienced listening panel and then used to instruct naïve listeners. By comparing the listeners' performances at ranking a number of stimuli before and after the training sessions the effectiveness of the adopted method was quantified.

1 INTRODUCTION

With the proliferation of multichannel sound reproduction systems in the recent past there has arisen the need to assess the performance of such systems in terms of their spatial quality. Thus, several studies have been conducted with the aim to isolate orthogonal subjective attributes of spatial impression [e.g. 1, 2, 3]. The identification of these underlying components is advantageous in that they can enable the collection of detailed data. For instance, the audio engineer can employ them for the subjective assessment of a particular multichannel sound system and hence determine the perceptual salience of each of these discrete spatial dimensions. Ultimately, this could allow him to control the spatial quality of the system through appropriate modifications of the corresponding physical factors. Thus, from such parametric assessments very explicit information is potentially available, which is not the case if spatial quality is evaluated globally, e.g. by collecting simple preference or mean opinion score (MOS) data. However, it is self-evident that the complexity of the task and hence the effort on the part of the subjects is greatly increased if a 'direct attribute rating' paradigm is applied instead of a MOS methodology, for example. So in order for such subjective evaluation techniques to produce useful results, suitable precautions have to be taken to circumvent the various problems commonly encountered.

It is widely acknowledged that the treating of 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 is essential to minimize the variability in order to obtain meaningful data on which sound decisions can be made. That is why subjects need to be put in a frame of mind to understand the characteristics they are asked to measure, which can be achieved by controlled practice and training [4]. This is especially important if products 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.

With regard to audio-related applications, there is concrete evidence in the literature concerning the successful training of subjects to improve specific auditory perceptual skills. For instance, trained listeners were shown to perform substantially better than untrained ones in terms of their ability to detect and discriminate between different auditory stimuli in a consistent fashion [5, 6]. As a result, they were more critical and produced data that were statistically more reliable leading to quantifiable improvements in their performances. However, while a lot of such research has dealt with the training of listeners in timbre perception a lack of similar work related to the assessment of spatial sound reproduction is apparent. Therefore, this study was designed to address this shortcoming.

The structure of this report is as follows. It starts with an overview of previous work concerned with the simulation of qualitative factors related to spatial sound reproduction highlighting the benefits of adopting a perceptual as opposed to a physical approach. Next, it is outlined how two sets of reference stimuli to be used for training purposes were generated based on the findings of associated research. This is followed by a detailed discussion of the validation of these sound excerpts with regard to their intended subjective effect. Finally, it is described how the stimuli were employed during a pilot investigation into the training of naïve listeners for spatial sound evaluation purposes.

2 SIMULATION OF PERCEPTUAL FACTORS

2.1 Introduction

When dealing with the simulation of psychoacoustic phenomena, the researcher has two options at his disposal. He can either adopt a physically or a perceptually based modelling strategy. The objective of the former is to emulate the transformation of sound in an enclosure as precisely as possible. Therefore, an extensive simulation of all elementary physical relationships that contribute to the formation of the sensation of interest is required. In the field of acoustics, this usually means that a very elaborate and hence computationally expensive geometrical model has to be constructed, which serves as the framework for any subsequent calculations. Conversely, an approach based on human perception can make use of perceptually valid simplifications warranted by findings from associated psychophysical studies. By deciding on what will be audible and what will be masked, the audio engineer can make substantial savings in terms of signal-processing costs. Further, opting for this method means that a user-interface can easily be designed, which will enable direct manipulation of the sensations of a listener. Such an interface is advantageous in that it saves the user from having to deal with any low-level details of the implementation. Control can be provided with the help of a few buttons and sliders, which regulate several physical parameters simultaneously. Hence, the total number of variables that need to be adjusted for synthesizing or manipulating an auditory scene can be kept to a minimum, thereby making the whole process much more user-friendly and intuitive.

Chowning [7] probably introduced the principle of a perceptually based user-interface to the audio community. The program he implemented gave the user independent control over the lateral position and perceived range of a sound source for the first time. Others followed him along the same lines [e.g. 8, 9, 10], increasing the sophistication and complexity of the simulations as well as adding other subjective dimensions. Arguably, the most comprehensive perceptual spatial sound processor to date is the Spat~, which has been developed at IRCAM [11]. Being primarily aimed at musicians and sound engineers, the Spat~ offers control over several perceptual characteristics of spaces “directly related to the perception of the reproduced sound event by the listener” [12]. Moreover, these attributes can be altered in real-time, thus offering an effective means of familiarizing a listener with the associated qualitative effects.

Although the implementation of a real-time controllable high-level interface is also envisaged as part of the current work, such a tool is not strictly mandatory in order to be able to train listeners in spatial sound perception. It seems reasonable to assume that listeners can be acquainted with changes in certain spatial characteristics with the help of a number of sounds that differ in their intensities with respect to the spatial attributes under investigation. This is the approach currently pursued for this study.

2.2 Aims

When trying to find a way of creating exemplary stimuli, the aim was to devise as simple a processing method as possible so as to enable the straightforward implementation of a corresponding algorithm at a later stage of the work. Another feature strived for was that the processing should work for several types of source material.
In order to achieve these goals, the authors’ attitude was that any means would be justified as long as the perceptual effect of the stimuli would be the intended one. In fact, it was felt that the exaggeration of specific physical and perceptual cues might be helpful in order to acquaint listeners with a particular qualitative phenomenon. Hence, strict adherence to physical reality was deemed unnecessary.

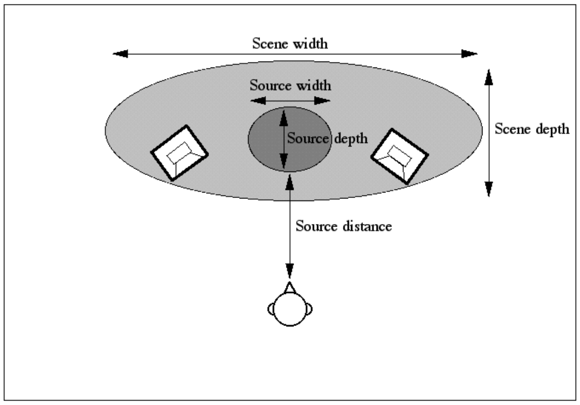
2.3 Which spatial attributes to choose?

Concert hall acousticians have assumed spatial impression to be a complex, multidimensional phenomenon for a long time. More recently, studies have been conducted in the area of multichannel sound reproduction with the objective to identify the individual constituents of spatial quality and their relative perceptual weights. For instance, Berg and Rumsey carried out elicitation experiments that enabled the extraction of several fundamental characteristics from the listeners’ responses using appropriate statistical techniques [1, 2]. The identified spatial attributes included naturalness, localization, source width as well as overall width, source distance, source depth as well as overall depth, and envelopment. Interestingly, similar results were obtained by Koivuniemi *et al.* [3] from another series of tests using a different methodology thus strengthening Berg and Rumsey’s findings.

To be able to determine whether naïve listeners can be trained in spatial sound perception, it seemed sensible to start with purely descriptive constructs, e.g. those related to the geometrical dimensions of sources and spaces. It was expected that the subjects would be more familiar with such concepts compared to more abstract ones that might be less meaningful to them. On these grounds, source distance and source width were selected. Also, the psychoacoustics of source distance and width have been the subjects of considerable research and it was hoped that the associated results would be applicable to or at least offer a good starting point for this study. This is discussed in more detail in the following sections.

The concepts of source distance and source width are illustrated graphically in Fig. 2.1. Although not addressed in this part of the study, source depth, scene depth and sound stage/scene width are also depicted in order to illustrate the differences to the notions of source distance and source width. The distinctions appear to be important as the results from elicitation experiments like the ones mentioned above indicate.

Figure 2.1: Graphical illustration of the concepts of source distance, source depth, source width, scene depth and scene width.



3 DISTANCE STIMULI

3.1 Distance perception

Over the years, a lot of research into distance perception has been carried out. While there appear to be some minor disagreements concerning the importance of particular acoustical and perceptual cues for judging the proximity of a sound source, a consensus seems to have been reached with regard to the most salient ones.

Nielsen [13] investigated the influence of several physical parameters on the perception of distance, frequency content being one of them. In this respect, he claimed that an increase in low-frequency (LF) content can take place for very close sound sources due to the curvature of the sound waves. Blauert [14] acknowledged that spectral distortions occur, but pointed out that their influence is small for distances greater than 25cm. Moreover, the specific spectral attributes evaluated by the auditory system for close sources have not been determined yet. The situation is clearer for a sound source located far away, i.e. 15m and more. In this case, Blauert stated that high-frequency (HF) loss due to air absorption can be distinctly audible above 10kHz.

Being well established, the close relationship between loudness and sound pressure level (SPL) gives no reason for disputes. It is a fact that under free field conditions the SPL will obey the 1/r law implying a drop of 6dB per doubling of distance. This will be less in non-anechoic environments, the exact value depending on the diffusivity of the sound field.

Clearly, the acoustical properties of non-anechoic spaces also determine the relative mixture of the direct and reflected sound, commonly referred to as the direct-to-reverberant sound ratio (d/r ratio). Beyond the critical distance of a room, reflected energy will dominate the sound field and consequently the d/r ratio decreases leading to an increase in perceived source distance. Interestingly, Chowning's [7] 'moving sound source simulator' relied on this parameter to create the illusion of changes in the closeness of a sound image. In addition, the Doppler shift was included in the simulation to supplement the credibility of the movement of a source.

However, the long prevailing view that humans perceive distance almost solely on the basis of the d/r ratio has been questioned by more recent research. Gerzon [15] suggested that each reflection provides the hearing system with an opportunity to deduce information about the proximity of a sound source. Hence, by modelling only a very few early reflections the resultant sense of distance was claimed to be poor in general. With regard to the temporal distribution of reflected sound, Gerzon [16] stated that reflections arriving within the first 50ms after the direct sound supply the predominant cues for distance estimation. Griesinger [17] endorsed the same view and stressed the need for the early reflected energy to arrive from lateral directions. Kendall *et al.* [9] went even further arguing that distance perception is governed by the exact spatial and temporal pattern of early reflections. By simulating reflections, which followed the direct sound by up to 33ms, they were able to successfully simulate changes in the range of a sound image.

Yet, the impact of the directional properties of reflected energy on distance hearing is still open to debate. In fact, there are indications that the spatial distribution of early reflections has only a marginal influence on the perception of the range of a source. For instance, the good sense of distance conveyed by omnidirectional monophonic recordings strongly supports this argument.

To summarize, various physical parameters have been suggested to influence distance perception, the most prominent ones being:

- loudness
- d/r ratio
- finer structure of early reflections
- spectral changes

Based on these findings, informal listening tests were conducted to determine the relative merit of each of these factors to the simulation of source distance. The details and outcome of these tests are presented in the next section.

3.2 Creation of distance stimuli

As mentioned before, the goal of this stage was to produce stimuli that could be used to demonstrate the perceptual effect of changes in source distance to listeners. Ideally, these stimuli would vary along the distance dimension in a unidimensional way, thus requiring all variables to be limited to solely those of interest. That is why recording suitable programme material was deemed inappropriate for this task due to the inability to exert precise control over the temporal and spatial distribution of the reflected energy picked up by the microphones. Thus, it was decided to create the sound excerpts using anechoically recorded monophonic source material as the direct sound signal. Previous research has emphasized the benefits of simple yet representative programme material when conducting subjective listening tests on spatial sound perception [18, 19]. So in order to avoid difficulties caused by asking subjects to judge too complex stimuli, only single stationary sources positioned directly in front of the listener were synthesized. Ecological validity was hoped to be achieved by selecting source material commonly encountered in natural hearing. In addition, it was expected that the choice of familiar sounds would facilitate the training of naïve listeners. The material was taken from the Archimedes CD [20] and comprised a cornet, trumpet, male voice and acoustic guitar. Each audio item was edited to a length ranging from 7s to 12s. Care was taken to ensure that the musical integrity of each extract was preserved so as not to cause annoyance of the subjects during the subsequent listening tests.

The stimuli were created in Studio 3 (ST3) of the University of Surrey's Department of Sound Recording. ST3 is a multichannel surround sound control room, which conforms to many of the design criteria of ITU-R BS 1116 [21] including dimension ratios, noise floor, reverberation time and loudspeaker arrangement. However, it departs from the recommendation with respect to the required attenuation of early reflections (arriving at the listening position within 15ms of the direct sound) due to the presence of a Sony OXF-R3 digital mixing console, a 19" computer monitor and an equipment rack. The reproduction set-up comprises five full-range, active loudspeakers (ATC SCM100As) each being located at a distance of 2.3m from the optimum listening position. A diagram illustrating the layout of ST3 is included in Appendix A.

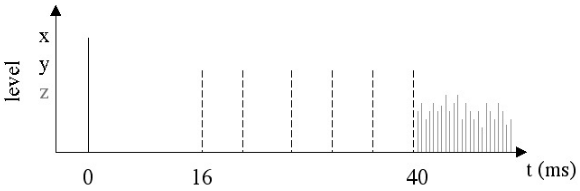
For the synthesis of the sounds, a Lexicon 480L reverberation processor was used in conjunction with the mixing desk. The reverberation unit is equipped with four digital outputs and offers control over up to six discrete reflections, which are individually real-time adjustable in terms of delay and amplitude relative to the input signal. A room preset was selected having a reverberation time (RT) of just under 1s at mid frequencies. Based on the findings of other researchers outlined above, a generic impulse response was designed consisting of three separate time regions, i.e.:

1. Direct Sound: $t = 0$ms
2. Early reflections: 15ms $< t <$ 40ms
3. Reverberation tail: $t >$ 40ms

The choice of 40ms as the upper time limit for the early reflections was meant to prevent them from being spaced too widely along the time axis. Otherwise they might become audible as discrete delays in the case of transient source material, e.g. the acoustic guitar. To avoid conflict with the group of early reflections, the onset of the diffuse reverberation tail was delayed appropriately.

All stimuli were mixed for reproduction over Left (L), Centre (C) and Right (R) only, the direct sound being routed to C. As Griesinger [22] pointed out, medial reflections are generally inaudible in music, but they can lead to colouration if their level is high enough. Hence, it was decided to distribute the six early reflections equally between L and R. All reflections were set to the same level and once a suitable pattern had been found, this was not altered any further. The two channels of decorrelated reverberation were also mixed into L and R, respectively. Left Surround (LS) and Right Surround (RS) were not used because informal listening tests had revealed that they were not required for the simulation of source distance. So in order not to run the risk of exciting any unwanted spatial dimensions by providing reflected energy from the sides and rear of the listening position, they were simply omitted. Fig. 3.1 illustrates the processing applied by means of the reverberator.

Figure 3.1: Generic impulse response of the reverberator used for distance processing. The six early reflections are shown in dashed lines and the reverb tail in grey. The level controls of the three time regions are indicated by x, y and z.



The perceptual effect of changes in the closeness of a source with respect to the listening position was produced by varying the relative gains (indicated as x, y and z in Fig. 3.1) between the three chosen time regions. In particular, the direct sound level was decreased with distance in a monotonic manner for all types of source material whereas the early reflections were slightly increased by a few dB before being reduced in amplitude again. As to the level of the reverberant signal part, it can be assumed that in small spaces there will be little change at the receiver position for a sound source radiating a signal at constant intensity but at varying distances [7]. Yet, to intensify the distance effect, the level of the reverberation tail was raised slightly for those sound examples to be used at the upper end of the distance scale. In addition, a low-pass filter with a cut-off frequency of 10kHz and a slope of 12dB/oct. was inserted into the signal path of the direct sound for the most distant stimuli thereby giving rise to a slight high-frequency attenuation. This was meant to imitate the effect of air absorption. No effort was made to accommodate the LF boost measurable for close sound sources in natural hearing. Blauert's [14] extensive discussion of the spectral distortions occurring for very close sound sources leaves no doubt that its relationship to distance hearing is not clear as yet. Also, the minimum distance of a sound image to be reproduced by the implementation outlined above is currently restricted to the loudspeaker distance, which in this case was 2.3m. Therefore, it was felt that the relative merit of this cue to the simulation was little and could be neglected.
The 'tuning' of the stimuli was based on subjective evaluation and an attempt was made to achieve linear spacing of the sounds to facilitate the detection of differences during the training stage of this study.

A block diagram of the processing structure devised for the distance manipulation is included in Appendix B. The entire processing took place in the digital domain and the sound excerpts were recorded to hard disk as three-channel .aif files with a sampling rate of 44.1kHz and a resolution of 16-bit.

3.3 Some brief remarks

It is self-evident that the adopted approach is a rather crude approximation to the conditions encountered in natural hearing. What is more, it seems to contradict some of the other research findings outlined above. Yet, although being very basic it was found that this method provided sufficient flexibility for creating a number of sounds with different degrees of source proximity that bear a strong qualitative resemblance to what happens in the real world. Therefore, the notion that the spatial and temporal distribution of early reflections is important for the successful simulation of distance is challenged. Instead, it is proposed that while distance hearing depends on the well-known d/r ratio, the temporal distribution of the reverberant energy has a major bearing on it, too. It is likely that the finer structure of the early reflections provides some supportive cues enabling subtler differences to be detected when assessing the proximity of sound sources. However, on the whole, its impact seems to be of lesser importance - a view also endorsed by Michelsen *et al.* [23].

4 WIDTH STIMULI

Awareness of the importance of source width perception emanated in the area of concert hall acoustics where the fundamental psychoacoustic principles behind this subjective phenomenon were established. In reproduced sound, the wish to be able to control the width of a sound image has been around for a long time and various techniques have been developed. By investigating some of these techniques, their suitability for manipulating source width in a unidimensional manner has been evaluated. A brief description of the psychoacoustics of width hearing is presented below, followed by a review of existing width processors and a detailed discussion of the adopted method.

4.1 Perception of apparent source width

The concept of apparent or auditory source width (ASW) stems from findings by concert hall acousticians who were concerned with the identification of the individual components of spatial quality. Beranek [24] provides a good overview of the subject. There is widespread agreement that it is the reflected energy arriving at the listening position within 80ms of the direct sound from lateral directions, which causes the subjective effect of the broadening of a source. Consequently, the sound source appears to fill a larger amount of space than its physical dimensions would suggest.

As Rumsey [25] pointed out, concert hall experiments have shown that listeners prefer larger amounts of ASW (the optimal degree of ASW still having to be determined). However, with regard to reproduced sound, it is unclear whether the same preference for larger ASW exists. While precise 'tight' imaging capabilities of audio systems appear to be important to pop mixing engineers, classical music is often recorded using spaced microphone techniques that are known to give rise to a more 'airy' or 'spacious' aural impression.

In the context of ASW, Rumsey also suggested a differentiation between the overall 'sound stage width' (i.e. the distance perceived between the left and right limits of the stereophonic scene) and 'individual source width' (ISW). Such a distinction is a valuable step towards an unambiguous description of spatial quality that is likely to result in more meaningful responses from listeners. Hence, the notion of ISW was also embraced for this study.

4.2 Available width manipulation systems

A widely known technique for spreading monophonic signals is the so-called 'divergence' feature found in some mixing consoles (e.g. the Sony DMX-R100 or OXF-R3). This technique utilizes simple pairwise intensity panning across a number of speakers (usually L, C and R) and therefore represents an extension to the conventional L-R panpot. Divergence controls the L/C/R panning parameters thereby regulating the proportion of sound mixed into each channel. Undoubtedly, it is far from being psychoacoustically optimized for manipulating the width of a sound image, but manages to score in terms of simplicity.

A more sophisticated approach to width control is stereo shuffling, which is based on Alan Blumlein's M-S concept. Stereo shuffling was revived by Gerzon [26] and can involve frequency-dependent width processing because it allows equalizing the difference and sum signals of 2-channel stereophonic source material in different ways before recovering the original left and right channels. This can be beneficial to widen stereophonic material at LF, e.g. recordings made with two coincident cardioid microphones, which basically produce a monophonic output at LF. Nevertheless, Gerzon stressed that stereo shuffling is not uniformly effective, i.e. there can be unwanted side effects with source material that does not rely entirely on amplitude differences to construct the sound image. For instance, with time-based stereo, frequency-dependent cancellation effects occur when summing/subtracting L and/from R. That is why the sound image will be perceived to be wider at some, but narrower at other frequencies resulting in a "possibly confused and degraded stereo image".

As part of their 'SceneBuilder' program, Corey *et al.* [27] devised a width control to emulate a perceptual effect occurring when a sound source in a real acoustic space moves to a position close to a room boundary. In this case, they claimed that one perceives a widening of the source accompanied by an increase in LF content due to changes in the spatial, temporal and spectral characteristics of the source and the room. To imitate this effect, four "fuzzy" sources were implemented surrounding the direct source and maintaining the same relative distances from it, irrespective of the direct source's position. Essentially, the 'SceneBuilder' uses dynamic level adjustment to ensure that when the direct source is located in the centre of the room, all "fuzzy" sources are attenuated completely, but that two "fuzzy" sources are increased in level when the source moves towards a wall. In addition, a low- and a high-shelving filter are employed to alter the spectrum of both the direct and the "fuzzy" sources. If close to a room boundary, the direct source's spectrum is modified by cutting the HF content and boosting the LF content while both HF and LF content of the "fuzzy" sources are increased by a few dB.

Other approaches to width simulation have been presented in [16] and [28]. But since these systems were designed with somewhat different objectives in mind, they are not discussed any further.

4.3 Creation of ISW stimuli

Since real-time control over discrete physical parameters had proven to be very valuable when generating the subjective impression of changes in source distance, it was decided to tackle the creation of the width samples in a similar manner.

An obvious starting point was to try to emulate the physical characteristics identified to be responsible for changes in perceived source width in natural hearing. Extensive experimentation with the directional and temporal properties of simulated reflections revealed that up to a certain level, reflected energy arriving within 80ms of the direct sound from lateral directions generally caused a spreading of the frontal sound image. Yet, when the level of the reflections exceeded a certain value, the result was a change in several spatial as well as non-spatial characteristics thereby prohibiting the desired unidimensionality of the effect. What is more, the latter's magnitude was certainly not large enough to allow the generation of nine[1] stimuli each exemplifying a different intensity in source width. Hence, it was concluded that the creation of the ISW stimuli could not be achieved by the sole manipulation of early lateral reflections and therefore a new strategy had to be employed. It was planned to accomplish suitable control over the spread of the frontal image with the help of established width processing methods and to use the Lexicon reverberator combined with appropriate mixing techniques to construct a space around the frontal image. The reasoning behind modelling a space was to be able to counteract the occurrence of unwanted perceptual artefacts, e.g. a collapse of the total sound stage into the centre speaker when simulating a narrow source. It was feared that such a change might be interpreted by listeners as a narrowing of the overall scene or room width. Also, reflections arriving from the sides were expected to offer supplementary control over width, i.e. by using a slight in-/decrease in level to broaden/narrow the sound source.

Initial attempts were disappointing since the magnitude of the achieved width effect was still too small. For instance, when making use of the mixing console's built-in divergence control, a spatially very confined image could be easily produced by feeding the monophonic signal to only the centre channel. In contrast, the maximum source broadening achievable with the divergence control was clearly insufficient (even in the presence of early lateral reflections) to be able to generate several other stimuli lying in between the two extremes of the width scale. The opposite was the case with stereo shuffling, i.e. while suitable stereophonic programme material could be made to sound very wide, restricting it to a narrow point source proved to be impossible (phantom mono being too de-focused). Also, due to the fact that stereo shuffling changes the degree of difference between the left and right channels it broadens/narrows the sound scene as a whole rather than individual components contained within it. Although it was found that this could be remedied somewhat with the help of the modelled space, stereo shuffling appears to be really only suitable for controlling scene width.

The attempt to take only one channel of the stereo recording, route it to the centre speaker and construct a space around it turned out to be problematic, too. This was because the reverberant energy contained in the single channel clustered around the sound source since it proved to be impossible to integrate it into the mix. Not only did this sound unnatural, but also spatially very different (i.e. greater distance of the sound source and reduced room width) compared to the stimuli for which both channels of the source material were needed to produce a sound image having ample width. These findings were more or less the same for several microphone techniques (including X-Y, M-S and A-B).

Two conclusions were drawn from the above findings:

1. To enable the creation of both a well-defined, highly localized as well as a very wide, diffuse sound image that would be adequately different from each other to allow several stimuli to be placed in between, three-channel source material was needed.

2. The frontal images of the stimuli would have to be generated using L, C and R in all cases to keep any unwanted spatial and timbral differences between the sounds to a minimum.

While it probably would have been possible to obtain suitable source material with the help of three-channel recording techniques, a viable alternative was to synthesize the centre channel from existing two-channel material. Various matrix-based methods have been developed for this kind of upmixing purpose (e.g. see [25] for more details). Arguably, the best results can be achieved with a technique known as Trifield that was proposed by Gerzon and others as a means of deriving a centre channel. In short, the Trifield technique is a psychoacoustically optimized three-channel panpot that enables a centre channel to be integrated into the frontal image in a truly coherent fashion. In [29], Gerzon provided a detailed insight into the functionality of such a device referring to the frequency-dependent processing that takes into account the different properties of human hearing at different frequencies. He also stressed the importance for such a decoder to be energy-preserving as this leads to far less audible colouration over a wide variety of recording techniques (both time- and intensity-based) compared to decoders with marked variations in energy gain for different components of the input signal. Moreover, an energy-preserving decoder maintains the widest stereo images whereas other systems cause a reduction in total sound stage width.

For this study, the Trifield algorithm implemented in a Meridian 565 Digital Surround Processor was employed. The Trifield algorithm was fed with source material that was based on the anechoic monophonic sound excerpts used for the creation of the distance stimuli, i.e. the male voice, acoustic guitar and cornet. These were 'spatialized' by simulating an A-B recording (with a microphone spacing of 35cm and a mic-to-source distance of 1m) in CATT Acoustic [30]. Care was taken to ensure that the programme material contained only small amounts of reverberant energy so as not to mask the subsequent processing. Hence, the room modelled in CATT Acoustic had an average RT of approximately 1s. This auralization approach was preferred to making real recordings as it allows one to trace back the reflection patterns picked up by the microphones and it was felt that this might be beneficial at a later stage of the work. A spaced microphone technique was chosen because of the more diffuse, wider sound images such recording systems tend to create compared to coincident ones, for example. Combined with a synthesized centre channel, large differences in source width were achieved. The colouration due to the Trifield processing was found to be negligible.

To emulate a space around the frontal image, the following generic impulse response was designed with the help of the reverberator:

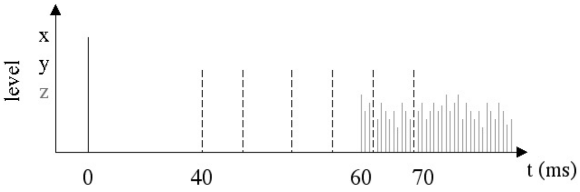
1. Direct sound: $t = 0$ms
2. Early reflections: 40ms $< t < 70$ms
3. Reverberation tail: $t < 60$ms

The reverberator was also fed with the A-B recording. Again, six early reflections were synthesized, which were all set to the same level and once a suitable pattern had been found, this was not altered any further. The time range of 40ms to 70ms turned out to be suitable for controlling scene width and since the output of the Trifield algorithm also contained reflected energy, the relatively large initial time gap of 40ms was not noticeable. The six reflections were divided into two groups of three and amplitude-panned between L and LS or R and RS, respectively, in order for them to arrive at the listening position from lateral directions. The two channels of decorrelated reverberation had a length of 1s

[1] The choice of nine stimuli was due to statistical requirements outlined in Sect. 6.1.

and were routed to L and R. A pre-delay of 60ms was employed to restrict interference with the early reflection set. Fig. 4.1 depicts the generic impulse response of the reverberator.

Figure 4.1: Generic impulse response of the reverberator used for ISW processing. The six early reflections are shown in dashed lines and the reverb tail in grey. The level controls of the three time regions are indicated by x, y and z.



The perceptual effect of changes in the width of a source was produced by (in order of importance):

1. Spreading the synthesized centre channel across L, C and R using divergence;
2. Changing the relative level between the L/R and C outputs of the Trifield by ±1.5dB;
3. Changing the level of the early lateral reflections by up to 1.5dB.

The 'tuning' of the stimuli as well as their loudness alignment was done by ear. Although the processing had to be pushed to its limits in order to achieve some very wide stimuli, care was taken to ensure that the effect of image splitting was avoided. An attempt was made to achieve linear spacing of the sounds to facilitate the detection of differences during the training stage of this study. Since the processing was fairly complex already, no effort was made at this stage to make the width manipulation frequency dependent, as is the case with the stereo shuffler or the 'SceneBuilder'. Rather, it was decided to wait and see whether the desired subjective impression could be achieved without it or whether the processing would have to be modified in some way.

A block diagram of the algorithm deployed for ISW manipulation is included in Appendix B. The entire processing took place in the digital domain and the sound excerpts were recorded to hard disk as five-channel .aif files with a sampling rate of 44.1kHz and a resolution of 16-bit.

4.4 Some brief remarks

In Sect. 3.3, the authors acknowledged that the applied distance processing was only loosely related to physical reality. The simulation is even more artificial in the case of ISW. From the description above, it is obvious that the problems encountered in creating nine width stimuli were manifold and while a technique was found that worked for various types of source material, the resultant differences between the stimuli were still very small. It might have been possible to overcome this problem by spacing the left and right loudspeakers further apart, e.g. ±45° instead of ±30°. However, for reasons of compatibility and transferability of the stimuli it was decided to stay with the standard 3/2-stereo reproduction layout.

5 VALIDATION EXPERIMENT

A validation experiment was conducted to verify whether the intended unidimensionality of the generated stimuli had been achieved. The experiment comprised two listening tests: one for source distance and one for source width. Nine unpaid subjects took part during the listening test on source distance. The subjects were one graduate and eight final year students of the University of Surrey's Sound Recording degree programme who had been chosen for their critical listening skills. The rationale behind using experienced listeners was that if their responses did not reveal any unwanted perceptual artefacts, the stimuli would probably be orthogonal and therefore suitable for training purposes. Five of the eight final year students also completed the test on source width. None of them received any information about the nature of the experiment.

5.1 Physical set-up

The experiment was executed in the Department's ITU-R BS 1116 listening room (LR) using the customized listening test software 'Alex'. The software runs on a Silicon Graphics 02 (SGI) computer whose ADAT digital output is connected to a Yamaha 02R mixer for D/A conversion. Five Genelec 1032A loudspeakers were set up in the standard 3/2-stereo configuration at a distance of 2.2m from the listening position. The loudspeakers were aligned in level to within 0.2dBA of each other using a pink noise generator and a Brüel and Kjäer 2123 real-time spectrum analyzer. 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. The computer monitor was positioned directly in front of the listening position and the subjects could control the speed of the listening test and switch between the stimuli at their leisure. A diagram of the experimental set-up is included in Appendix A.

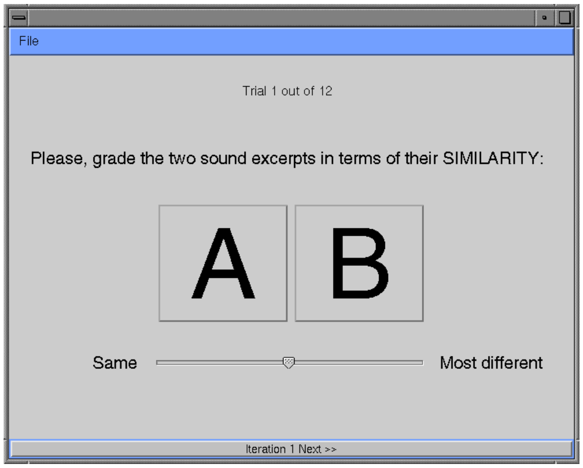
5.2 Experimental design

The experimental design was based on the Multidimensional Scaling (MDS) technique, which can be described as an exploratory, decompositional, attribute-free approach to sensory analysis. It relies on comparing each stimulus in a given group with every other stimulus of the same group. However, the comparisons are not made in terms of highly subjective verbal descriptors of a specific quality. Instead, the stimuli are graded with regard to their degree of overall similarity thereby eliminating the need for instructions concerning pre-specified characteristics on which these judgements are to be made. As a result, any potential bias on the part of the experimenter is reduced because language can be kept to a minimum and different understandings of semantic meanings are avoided. The obtained similarity judgements are then represented in a stimulus space where the 'psychological distances' between the stimuli are illustrated graphically. Those stimuli rated by the subjects to be similar appear as points close to each other whereas those stimuli judged to be dissimilar are distant from one another. Hence, MDS recovers the underlying structure among stimuli that is hidden in the data. Finally, the revealed dimensions need to be interpreted, leading to the problem of having to identify the fundamental perceptual components of the quality under investigation.

So with the help of MDS, it was hoped that a single strong dimension would be revealed for distance, and another for width, allowing the unidimensionality of the stimuli to be concluded. Because of the time-consuming nature of MDS tests, only one type of source material per spatial attribute could be verified. Yet, since the same processing method was used for all distance/ISW stimuli, extrapolation of the findings to the unverified ones was believed to be viable. For both distance and width, the acoustic guitar material was arbitrarily selected. Since there were nine (see Sect. 6.1) stimuli to compare, each subject had to make a total of 36 gradings per spatial attribute. A different order of presentation was created for each listener so as to minimize order effects. Both tests were subdivided into three groups of 12 trials. After the completion of a group, each subject was asked to take a short break in an attempt to reduce listener fatigue. Both listening tests took around one hour.

The user-interface implemented in the listening test software is shown Fig. 5.1. An undifferentiated line scale was provided for the subjects to indicate their perceptions. The subjects were instructed to make global dissimilarity judgements taking into account any and all detected differences when grading a pair of sounds. They were asked 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 the listeners to get an idea of the range of differences, they could listen to all nine sounds before and half way through each group of 12 trials. Thus, it was hoped that they would be able to familiarize themselves and refresh their memories with respect to the magnitude of possible differences between the stimuli and as a consequence grade the stimuli in a consistent manner.

Figure 5.1: User-interface employed for the MDS experiment



Please, grade the two sound excerpts in terms of their SIMILARITY:

To facilitate the interpretation of the dimensions revealed by the MDS analysis, all listeners were given a questionnaire at the end of each test and asked to verbally express the differences they had perceived between the stimuli. They were encouraged to think of words that were differential in nature, which they had to write down in order of priority. Again, they were offered the opportunity to listen to all nine sounds if they felt they had to.

6 RESULTS

6.1 MDS data analysis

To be able to analyze the similarity judgements from the validation experiment, the data were normalized, converted into proximity half-matrices and then entered into SPSS [31]. The chosen layout of the analysis was based on statistical requirements imposed by the format of the collected similarity data.

The number of stimuli used during the data collection phase determines the maximum dimensionality a meaningful MDS solution can portray. For the solution to be stable more than four times as many stimuli as dimensions are required [32, 33]. Since nine stimuli were used for this study, 1- and 2-dimensional solutions could be derived for both source distance and width, therefore permitting the unfolding of a second perceptually relevant factor. Changing the experimental design to allow for a higher dimensionality was unnecessary since the possibility of a 2-dimensional solution was sufficient to allow either confirmation or rejection of the hypothesis of the unidimensionality of the sound excerpts.

The decision of metric vs. nonmetric analysis has to be based on the quality level of the input measures of similarity. As the listeners judged the absolute magnitude of the similarities between the stimuli, the data can be assumed to be at the ratio level of measurement. However, Anderberg [34] recommended against declaring proximity data to be of ratio scale quality. He suggested that in a full array of paired comparisons subjects might not be able to make very precise judgements as to the amount of difference between each pair. Schiffman *et al.* [35] shared this view and proposed that most proximity data are probably ordinal. To the authors this seemed like a reasonable assumption to make. Since neither a reference stimulus nor a marked scale were provided during the trials, each subject had to completely rely on their aural memory. The subjects themselves also raised this point after the tests. Some listeners expressed their concern with respect to their ability to memorize the full range of differences and hence to grade the stimuli in a consistent way.

In order to be able to evaluate how well the raw data fit the MDS model, MDS techniques commonly calculate the stress measure, which ranges from 1 (worst possible fit) to 0 (perfect fit). Yet, for assessing the dimensionality of a solution, this measure can be problematic as it depends on various aspects of the model, e.g. it always gets smaller with increased dimensionality, regardless of whether the conditions for a correct MDS model are satisfied or not. Another index of fit is the squared correlation index (RSQ), which can be interpreted as the proportion of variance accounted for (VAF) by the MDS procedure. Schiffman *et al.* [35] emphasized the value of the RSQ as a measure of fit because it does not suffer from the same problems as stress. Although it is desirable to maximize the VAF of a given solution, the maximal number of dimensions taken into account needs to be limited, in particular if the increase in explained variance per dimension is less than ~0.05 [36]. This is because dimensions with a low contribution to the explained variance are difficult to explain and may be associated with noisy data.

On these grounds, 1- and 2-dimensional nonmetric analyses were carried out for both distance and width. To assess dimensionality both RSQ and stress were scrutinized as a function of dimensionality. The results are displayed in Table 6.1.

Table 6.1: Results from nonmetric MDS analysis for source distance and ISW (LR)

Spatial attribute	Dimensionality	Stress	RSQ
Source distance	1	0.25	0.8
Source distance	2	0.15	0.84
ISW	1	0.46	0.35
ISW	2	0.31	0.35

With regard to the results for distance, two observations can be readily made. Firstly, when going from a 1- to a 2-dimensional solution, there is a reduction in stress, which is not surprising since the more dimensions are available to the model, the easier it finds it to accommodate the data. Secondly, the 1-D solution is characterized by a high RSQ value that increases by only a small amount (i.e. less than 0.05) in the case of the 2-D solution. This indicates that a strong first dimension exists.

A different picture is apparent when one looks at the results for width. The first thing to note is that in the case of the 1-D solution, stress is high and RSQ is small, which suggests a bad fit between the data and the fitted model. Even more striking is the fact that there is no increase in RSQ at all when a 2-D analysis is employed. Thus, it seems that extending the 1-D to a 2-D solution cannot help unfold a pattern contained in the subjects' responses either. This signifies that the listeners either evaluated different perceptual factors or that common perceptual factors could not be completely identified by the MDS procedure due to the noisiness of the data.

The poor fit can be traced back to the transfer of the width stimuli from ST3 to the listening room. When the sound excerpts were played back in the LR, it was apparent that most of the width cues introduced into the sounds had disappeared. This is very bewildering if one takes into account the fact that ST3 and the LR are very similar in terms of their acoustical properties. In fact, the listening conditions in the LR are even more critical than those in ST3 (see Sect. 3.2). This indicates that while the distance stimuli appear to be not affected by the different reproduction set-up in the LR, the ISW sounds do not travel easily from one listening environment to another.
Feedback given by the subjects confirmed these findings. They reported that while they could hear clear changes in terms of distance, the detectable differences between the width stimuli were very small indeed.

This is also reflected in the obtained stimulus spaces. Regarding the 1-D solution for distance (Fig. 6.1), the stimuli appear in the correct order ('a' being the closest and 'i' the most distant stimulus). The spacings are not linear, which might be due to inaccuracies during the generation stage and/or the inability of the subjects to be consistent in their judgements. However, all sounds appear to have a different intensity with regard to distance thereby enabling the listeners to rank them correctly.

For width, on the other hand (Fig. 6.2), the stimuli almost cluster into two groups ('a' being the narrowest and 'i' the widest stimulus). This can be explained by the small differences between the sounds, which led the subjects to judge the sounds in each cluster to be highly similar or even the same. Also, stimuli 'f' and 'g' are shown in reversed order, which was probably due to the listeners' inability to distinguish between sounds close to each other along the width scale in a reliable fashion.

Figure 6.1: 1-D solution for distance

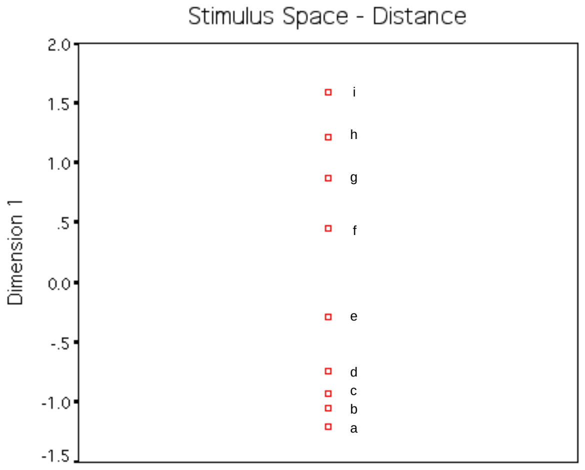
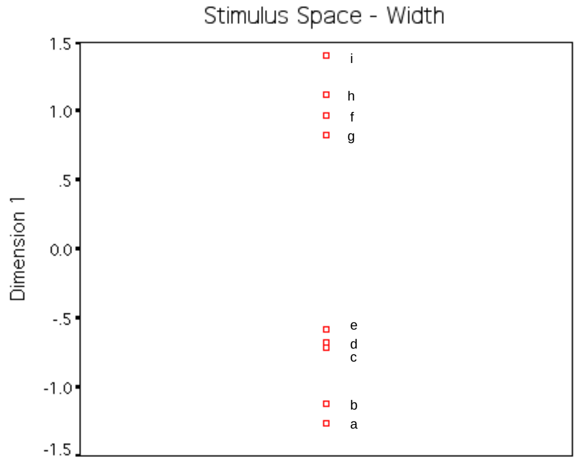


Figure 6.2: 1-D solution for ISW (LR)



The 2-D stimulus spaces are shown in Fig. 6.3 and 6.4. In both cases, an almost circular shape can be clearly seen. As was the case for the 1-D stimulus space for distance, the stimuli appear in the correct order, although this time they form a circle. For the width stimuli, on the other hand, no clear structure can be detected if one traces the stimuli's locations in alphabetical order. This is not too much of a surprise since it appears to confirm the trend already noted before, i.e. that the listeners' judgements are inconsistent. However, the circular pattern is rather perplexing in the case of distance since it seems to imply that a second important dimension exists. This opposes the trend insinuated by the RSQ values, namely that beyond the 1-D solution, the explained variance increases by only a marginal amount and thus a second meaningful dimension does not exist. At first glance, the cause for this pattern appears to be incomprehensible. Yet, a plausible explanation was given by Hair *et al.* [32] who pointed out that circular patterns are a strong indication of a degenerate solution, which is caused by inconsistencies in the data or the inability of the MDS algorithm to reach a stable solution. In both cases, the computer program is unable to differentiate among the objects for some reason. Degeneracy of MDS solutions was also addressed by Schiffman *et al.* [35]. They warned against overinterpretation of such solutions, e.g. by trying to give precise meaning to the stimulus positions. Rather, they advised trying to make sense of the stimulus spaces by considering the known properties of the stimuli.

To verify the meaningfulness of the second dimension revealed by the MDS analyses in more detail, the verbal data were examined. From this, it was hoped that an insight could be gained into the possibility of the degeneracy of the 2-D solution. The results are presented in the next section.

Figure 6.3: 2-D solution for distance

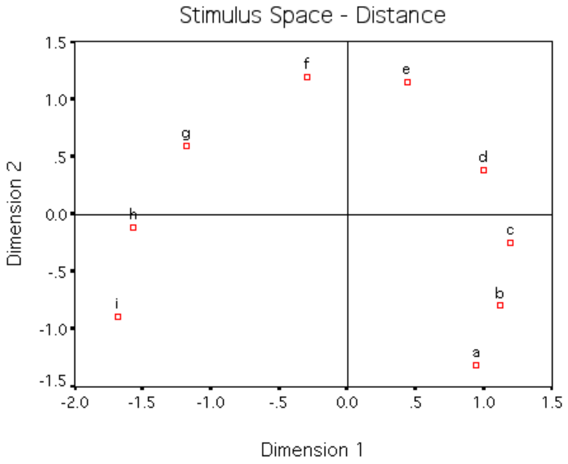
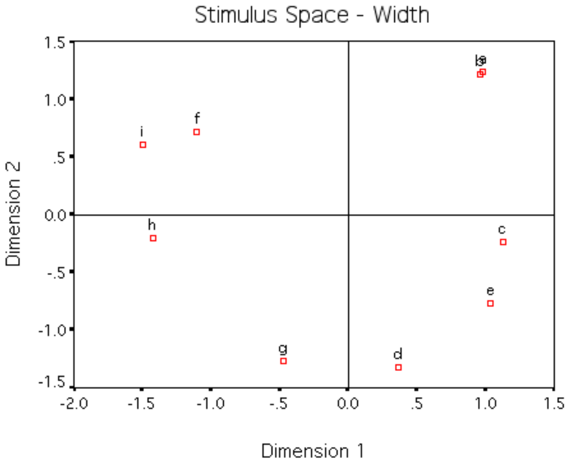


Figure 6.4: 2-D solution for ISW (LR)



6.3 Analysis of verbal data – source distance

The rationale behind the collection of verbal data was to facilitate the interpretation of the perceptual structure uncovered by the MDS analyses. For the verbal reporting stage, the subjects had been asked to express in their own words the differences they had perceived between the stimuli. They had not been instructed to comply with any particular response format in order not to bias their responses in any way. As a result of this free verbalisation approach, the obtained data were fairly diverse and needed to be structured. For this purpose Verbal Protocol Analysis [37] was employed, which is a methodology that enables 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 referred to a perceptual phenomenon as a whole or whether they described discrete physical factors contributing to the formation of a specific qualitative impression related to spatial sound reproduction. Because the experimenter has to interpret the meaning of the subjects' verbalised perceptions, there is a risk of biasing the outcome by subdividing the data into more and more 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 importance for each group of terms to the perception of the listeners, an overall weight factor was calculated. Since the subjects had been instructed to write down the perceived differences in order of priority, each term was weighted according to its position in its list. For example, the first verbal descriptor in a given list was assigned the value 1, the second 1/2, the third 1/3 and so forth. The values of all verbal descriptors within each group were then added up and the result was divided by the number of subjects that had taken part in the corresponding listening test.

With regard to the holistic terms, it was found that for both distance and width the first classification stage (i.e. holistic vs. analytical) structured the data in a logical manner. For distance, four groups were identified, the details of which are shown in Table 6.2. The first group contains terms strongly related to source distance such as 'proximity', 'distance' or 'closeness of the sound source'. This perceptual effect was noted by all nine listeners who participated in this experiment and seven of them wrote it down in first place resulting in a strong weight factor of 0.87. References were also made with regard to room size ("size of space" and "size of recreated room ambience"), source width ("focus/width of the source") and "source depth". Yet, for none of them was the agreement as uniform between the subjects as for source distance and hence their weight factors are a fraction of the one obtained for source distance.

Table 6.2: Groups of holistic terms and their relative weights obtained from listening test on source distance

Holistic groups	occurrences	weight factor
1. Source distance	9	0.87
2. Room size	2	0.11
3. Source width	2	0.06
4. Source depth	1	0.06

The results from the classification of the analytical terms are depicted in Table 6.3. Verbal descriptors related to timbre perception were stated most often. Four out of the seven terms in the timbre category addressed changes in HF content (e.g. "loss of HF", "sharpness" or "brightness"), one concerned overall timbre, and two LF content ("loss of LF", "relative volume of bass line to upper parts"). The relative weight factors of these three subgroups are 0.14, 0.06, and 0.04 denoting that changes in the HF part of the spectrum were most prominent. Thus, the intended effect of simulating air absorption appears to have been detectable.

Furthermore, three listeners made specific references to the d/r ratio, which is known to be closely related to distance perception. Hence, this provides further evidence that changes in source proximity were present.

The 'Clarity' category is comprised of the terms 'distinctiveness' (one mentioning) and 'loss of attack' (three mentionings). This can be explained by the fact that a lower d/r ratio tends to decrease the clarity of the sound image.

The last group contains terms that were only brought up once and that did not fit into any of the other categories. It includes terms such as "loudness" (another parameter varied for distance manipulation), "ringing sounds in rear speakers" and "delay of sounds coming from the rear relative to the front". The latter two descriptors are somewhat peculiar since the rear speakers were not used for the simulation of source distance.

Table 6.3: Groups of analytical terms and their relative weights obtained from listening test on source distance

Analytical groups	occurrences	weight factor
1. Timbre	7	0.24
2. D/R ratio	3	0.22
3. Clarity	4	0.17
4. Other	5	0.20

Finally, it is worth noting that although all listeners had received thorough training in analytical listening skills, seven out of nine expressed their perception of changes in distance in the form of a holistic descriptor first before resorting to analytical terms. And although most of the analytical groups contain terms directly related to distance hearing, their weight factors are much lower compared to the holistic source distance category. This can be interpreted as strong evidence for the successful simulation of the perceptual phenomenon of source distance.

6.4 Analysis of verbal data – ISW (LR)

Table 6.4 presents the results from analyzing the responses given by the five listeners who completed the experiment on source width. It is evident that there is no single strong group of holistic terms. Interestingly though, two references were made regarding changes along the lateral plane, i.e. image shift and source width. However, there is no agreement between several subjects with respect to the two groups nor do they have large weight factors. This implies that their perceptual relevance is low, which makes sense because the differences between the width stimuli were negligible when auditioned in the LR. It is also worth noting that two listeners detected changes in the distance of the sound source. Other researchers [e.g. 2, 38] have highlighted the difficulty associated with defining source width, where there is a risk of confusion between narrower image width, increased source distance and less spread of LF content. This might be an explanation for the two references to source distance.

Table 6.4: Groups of holistic terms and their relative weights obtained from listening test on ISW (LR)

Holistic groups	occurrences	weight factor
1. Image shift	1	0.20
2. Source distance	2	0.13
3. Source width	1	0.07
4. Envelopment	1	0.05

The groups of analytical terms are shown in Table 6.5. It has to be pointed out that the only meaningful subgroup that could be identified for this data set was 'Timbre'. Overall, this category has a large weight factor, but this is due to the aggregation of all descriptors related to spectral changes. More precisely, three references were made to the HF content ("brightness", "sparkle", "loss of HF"), two to the LF content ("lightness", "fullness") and one to the "loss of mid frequencies". The relative weight factors of these three subgroups are 0.25, 0.25, and 0.20. Thus, on the whole, the range of the weight factors for both the holistic and analytical ISW group is small, 0.05 being the lowest and 0.25 the highest value. Hence, it can be inferred that because of the lack of discernible differences between the stimuli, each listener might have started to focus on specific details of the reproduced sound field, giving rise to the diversity of terms that is apparent. Thus, there is hardly any inter-listener concordance again and the obvious conclusion is that the data is noisy and therefore meaningless.

Table 6.5: Groups of analytical terms and their relative weights obtained from listening test on ISW (LR)

Analytical groups	occurrences	weight factor
1. Timbre	6	0.70
2. Clarity	1	0.20
3. Early reflection time	1	0.20
4. Slight phase shift	1	0.10
5. Differences in overall sound	1	0.10
6. Length of reverb tail	1	0.10
7. EQ of reverb tail	1	0.07
8. Some guitar resonances seemed stronger in certain extracts	1	0.07

7 VALIDATION EXPERIMENT II

Since the results obtained from the validation experiment had proven to be inconclusive with regard to the dimensionality of the width stimuli, it was decided to repeat the MDS experiment in ST3 where the stimuli had been generated.

In order for the results to be directly comparable, it was ensured that the experimental design was as similar as possible to the listening test carried out in the LR. Nonetheless, a new listening test software had to be designed as the SGI computer could not be interfaced with the reproduction equipment in ST3. The program was implemented in the object-oriented programming language Max/MSP [39] and run on a Macintosh G3 computer. Care was taken to ensure that the user-interface resembled the one depicted in Fig. 5.1 as much as possible. Unfortunately, it was not feasible to position the computer monitor in front of the subjects due to the mixing console installed in ST3. However, all subjects were instructed to restrict their head movements and to make this easier for them, the computer keyboard could be used to control the playback of the stimuli. Also, due to the loudspeakers being larger than the ones in the LR and the presence of the mixing desk, only the three loudspeakers in front of the listening position could be concealed with the help of the acoustically transparent curtain (see Appendix A for a diagram of the experimental set-up). The mixing desk was configured in such a way that the channels in use were not displayed and all level meters were covered so that no visual cues were available to the subjects. The loudspeakers were level aligned to within 0.2dBA of each other using pink noise and care was taken to ensure that the absolute sound pressure level at the listening position matched the one measured in the LR.

7.1 MDS data analysis – ISW (ST3)

The five listeners who had participated in the first validation experiment on source width also completed the test in ST3. Their responses were treated in the same way as before and then analyzed using a 1- and 2-dimensional nonmetric MDS model. The results for stress and RSQ as a function of dimensionality are presented in Table 7.1.

Table 7.1: Results from nonmetric MDS analysis for ISW (ST3)

Spatial attribute	Dimensionality	Stress	RSQ
ISW	1	0.35	0.62
ISW	2	0.24	0.65

Evidently, the RSQ value of the 1-D solution obtained from the second experiment (0.62) is much higher than the one obtained from the first one (0.35). While it is clearly less than the one obtained for source distance, it is beyond the 0.60 mark and can therefore be considered as acceptable [32]. Moreover, the increase in VAF achieved by the 2-D solution is minimal again (i.e. less than 0.05), which appears to indicate that only one major dimension is present in the width samples.

An inspection of the 1-D stimulus space (Fig. 7.1) shows that the stimuli do not cluster as much into two groups as was the case for the stimuli from the first listening test on source width (Fig. 6.2). Again, two stimuli ('g' and 'h') are shown in reversed order, but this is comprehensible if one acknowledges that the differences between adjacent stimuli were small. Once more, the 2-dimensional stimulus space (Fig. 7.2) has strong characteristics of a degenerate solution and therefore should not be given too much weight.

Figure 7.1: 1-D solution for width (ST3)

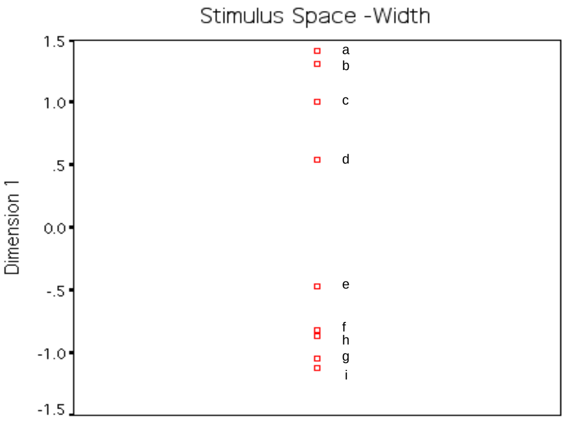
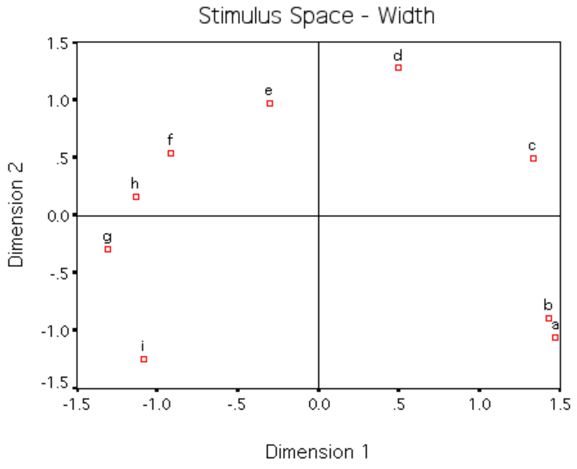


Figure 7.2: 2-D solution for width (ST3)



7.2 Analysis of verbal data – ISW (ST3)

The strategy employed for systematizing the verbal data collected after the second listening test on ISW was equivalent to the one used before. The subjects' responses were divided into holistic and analytical terms and identical or highly similar descriptors were grouped together.

The holistic groups of verbal descriptors are summarized in Table 7.2. At first glance, the results of the classification process appear to contradict the explanation found for the MDS solutions. Two holistic groups with the same weight factor have emerged from the analysis of the verbal data connoting that two equally strong dimensions exist, i.e. source width and source position. Regarding the former, four out of five listeners noticed changes in the focus or width of the sound image, which compares favourably with only one mentioning of the same subjective phenomenon after the first listening test. Furthermore, three listeners also perceived changes in the stereo position (i.e. between the left and right loudspeaker) of the sound source. A similar perceptual effect was brought up by the analysis of the verbal data from the first experiment, but in this case only one listener indicated the perception of an "image shift". The remaining three holistic groups (source distance, scene width and source depth) are due to a single mentioning in each case and as a result have significantly lower weight factors than source width and source position. So while there is some common ground between the sets of data obtained from the two listening tests on ISW, the listeners seem to be much more in agreement as to the differences they could detect during the listening test run in ST3.

Table 7.2: Groups of holistic terms and their relative weights obtained from listening test on ISW (ST3)

Holistic groups	occurrences	weight factor
1. Source width	4	0.60
2. Source position	3	0.60
3. Source distance	1	0.10
4. Scene width	1	0.07
5. Source depth	1	0.04

The structuring of the analytical terms revealed the groups shown in Table 7.3. The most noticeable difference is that the total number of analytical descriptors is almost half that collected after the first experiment resulting in fewer categories. Again, there is a group containing timbral descriptors, which comprises one reference to LF content ("loss of LF"), one to HF content ("fullness of high notes") and one to timbre in general. The relative weight factors of these three subgroups are 0.10, 0.07 and 0.07 proving that neither of them was perceptually important to any of the subjects. The same statement can be made about the other four categories, which contain terms that were elicited only once. All in all, the range of the weight factors of the analytical ISW groups is very small indeed, 0.05 being the lowest and 0.10 the highest value. Taking into account this finding and the ones for the holistic groups, it can be deduced that this time the differences between the width stimuli were far less ambiguous compared to the first listening test on ISW. Accordingly, a clear dominance of two holistic descriptors, a reduction in the total number of terms elicited and much stronger inter-listener agreement can be certified.

Table 7.3: Groups of analytical terms and their relative weights obtained from listening test on ISW (ST3)

Analytical groups	occurrences	weight factor
1. Timbre	3	0.24
2. Front-back first echo time	1	0.10
3. Reverb time	1	0.07
4. Phase shift	1	0.07
5. Loudness	1	0.05

Nevertheless, the fact that the RSQ values of the 1-D and 2-D solutions imply a reasonably strong single dimension whereas the verbal data indicate that two dimensions exist, poses the question of how to interpret the results. A possible explanation might be that the subjects mistook the changes in width of a sound image as changes in its stereo position. This could be explained by the much more difficult nature of the notion of ISW compared to source distance. In everyday life, humans make sense of their environment by constantly evaluating the closeness of the sources distributed around them. Hence, source distance is a very familiar concept and therefore listeners can easily relate to it. Conversely, ISW is an unfamiliar perceptual construct that is likely to be less meaningful to listeners. It is probably a fair assertion to make that most humans hardly ever consciously assess the width of a real single source based on auditory information only because of the dominance of visual cues.

It is also possible that the subjects' head movements might have induced a perceived movement in the position of the source within the stereo image. Although an attempt was made to reduce the influence of this variable on the subjects' verdicts as much as possible, the likelihood that it could be eliminated completely is small.

Because of the obvious uncertainties, it was decided to conduct another experiment that was meant to help answer the remaining question of the true dimensionality of the ISW sounds.

8 CONFIRMATION EXPERIMENT

This experiment consisted of two parts. For the first part, a classification approach was adopted because of the confirmatory nature of this experiment. In particular, the subjects were asked to make paired comparisons of a selected number of width stimuli. They then had to choose a spatial attribute out of a given list that could be used to describe the dominant difference which they could detect between each pair of stimuli. The stimuli were taken from the acoustic guitar and cornet source material and for both instruments four sounds positioned at the ends of the ISW scale were selected. Including all width stimuli was deemed unnecessary because the 1-D MDS solution had placed them in the correct order (except for one adjacent pair of sounds) thus indicating that all the sounds differed in terms of one perceptual factor. So to limit the duration of the listening test, only extreme stimuli were picked, as they were most suitable for evoking a width sensation from the listeners. The spatial attributes offered to the subjects were adopted from the holistic categories contained in the verbal data from the second ISW listening test (see Table 7.2). To restrict the number of attributes to a reasonable minimum, the three groups with the highest weight factors were chosen, i.e. source width (0.60), source position (0.60) and source distance (0.10). Also, by including an 'Other' category the subjects were given the option to indicate a different sensation. If a subject selected this category, a pop-up menu appeared asking the listener to specify their perception by entering a short description. Hence, 'no difference' verdicts were permitted as well as the possibility to denote a completely different perception in case the listeners found that none of the three spatial attributes could successfully describe the dominant difference they perceived.

The avoid misunderstandings, all participants were given the following definitions of the three selected spatial attributes:

1. *Source position*: Does the apparent spatial location of the reproduced sound source appear to change, e.g. do you perceive a shift of the source from a central position towards the left or right or vice versa?

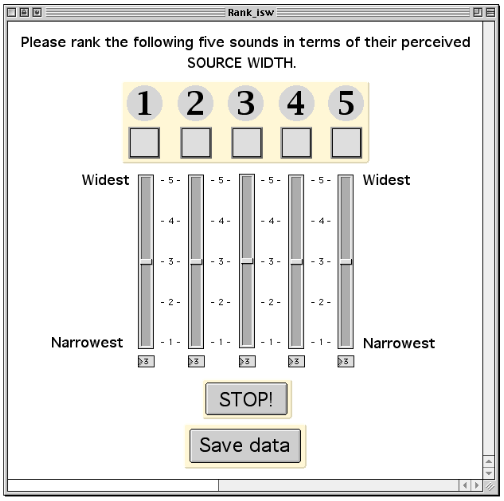
2. *Source width*: Does the width of the reproduced sound source appear to change, e.g. does one of the two sounds resemble a narrow, well-defined, focused point source whereas the other is wider, ill-defined, more diffuse?

3. *Source distance*: Does the perceived range between you, the listener, and the reproduced sound source appear to increase/decrease, i.e. does one of the two sources seem to be closer to/further away from you than the other?

During the second part of this experiment, all listeners had to rank a number of sounds that illustrated different intensities of ISW. As the sounds had proven to be very difficult to discern if all nine stimuli were presented, only five cornet stimuli were used (i.e. stimuli a, c, e, g and i). That is why the differences between adjacent sounds were slightly bigger. The user-interface (Fig. 8.1) comprised five sliders, which could be set to values ranging from 1 to 5 and the listeners were told to assign each value only once.

The objective of this ranking exercise was to verify whether the listeners found the concept of ISW to be meaningful or not if they were instructed to apply it to the judgement of a number of suitable stimuli.

Figure 8.1: User-interface of ISW ranking experiment



Please rank the following five sounds in terms of their perceived SOURCE WIDTH.

The listening test took place in ST3 again. The experimental set-up was identical to the one of the second validation experiment on ISW, except that this time no effort was made to disguise the positions of the loudspeakers. As the previous listeners had been exposed to some of the chosen sound excerpts before, they were likely to be pre-conditioned with regard to their judgements of the sounds' dissimilarities. That is why four different listeners were asked to take part in this confirmatory study. Two of them were first year and the other two second year Sound Recording students. All listeners were instructed to face forward when switching between the sounds.

8.1 Analysis of classification results

The stimulus pairs presented to the subjects during the classification test as well as the corresponding responses are shown in Table 8.1. Evidently, the listeners perceived changes in source width for more than half of all pairs of sounds and therefore this formed the strongest category. However, a significant proportion of all judgements was made in favour of source position, which came up as the next strongest group reflecting the finding of the second ISW listening test. Variations in source distance were perceived only a few times and the single 'Other' verdict was due to a listener perceiving two sounds to be the same. Hence, while these results allow the deduction that the changes in the created ISW stimuli are generally not mistaken for a distance sensation, confusion between source width and source position is still apparent. Although ISW turned out to be the dominant spatial attribute during this test, the results cannot prove the unidimensionality of the sound excerpts. In fact, it is likely that the inclusion of smaller differences between the ISW sounds would have spread out the responses more evenly.

Table 8.1: Accumulated spatial attribute judgements from four listeners from classification experiment

Stimulus pairs	Position	Width	Distance	Other
guitar A-H		3	1	
guitar A-I	2	2		
guitar B-H	1	2	1	
guitar I-B	2	2		
cornet A-H	1	3		
cornet I-B	1	2	1	
cornet I-A		3	1	
cornet H-B	1	2		1
Total	8	19	4	1

8.2 Analysis of ranking results

For this part of the confirmation experiment, the listeners were asked to rank five ISW stimuli in terms of their perceived source widths. The definition of ISW employed for the classification test was presented to the subjects again.

To allow the evaluation of the 'degree of wrongness' of the ranking sequences relative to the correct sequence, the squared Euclidean distance (SED) was applied. The SED is the sum of the squared differences over all of the variables, i.e.:

$$SED_{ij} = \sum_{a}^{r} (x_{ia} - y_{ja})^2$$

where x_{ia} specifies the position of point i on dimension a, y_{ja} specifies the position of point j on the same dimension and r equals the maximal dimensionality. Hence, a ranking sequence that is 'close' to the correct response will yield a small SED while a sequence that is 'far away' gives rise to a large value. In the case of this particular example, a response with a SED of 40 corresponds to the 'worst case scenario', i.e. the exact reversal of the correct answer. A value of 2 implies that an adjacent pair of sounds has been inverted and an SED of 4 denotes the reversal of two pairs of sounds.

The SED values calculated for each of the four listeners are presented in Table 8.2.

Table 8.2: Ranking sequences given by the listeners and the associated SEDs

Listener	Rank sequence	SED (max. = 40)
1	e, a, c, g, i	6
2	a, c, e, i, g	2
3	c, a, e, g, i	2
4	c, a, e, g, i	2

As can be seen, three out of the four listeners came very close to the correct response reversing only two adjacent sounds. This is encouraging, because it indicates that i) the listeners could apply the notion of ISW to the set of stimuli and hence rank them almost correctly, and ii) the stimuli varied along a common perceptual dimension that was pre-defined to be ISW. It is interesting to note that listener 1, who performed worse than the others, was one of the two first year Tonmeister students who took part in the test. It is possible that because of his limited experience in critical listening exercises, he found it harder to detect and evaluate the changes in ISW.

To summarize, there is no doubt that the results obtained for the ISW stimuli are not as convincing as the ones obtained for source distance regarding the unidimensionality of the intended perceptual effect. Nevertheless, there are strong indications that changes in source width dominated the listeners' perceptions of the ISW stimuli. The difficulties encountered with validating these sound excerpts might be due to the simulation of the effect being only partially successful. Another possible explanation is that because of the listeners' inexperience with evaluating the width of single sources, the perceptual construct of source position arose. This could have been at least partly caused by changes in the sound images due to head movements, which cannot be ignored completely.

The fact that subjects were capable of adopting a definition of ISW and applying it during a ranking task after they had been instructed to do so, strongly suggests that the ambiguous results were due to listener unfamiliarity with ISW. That is why on the whole, the results were deemed acceptable to allow the stimuli to be used for training purposes. This conclusion was based on the notion that ISW is a complex perceptual construct that listeners have confused before in other studies. Therefore, it appears that ISW needs to be pre-defined by expert listeners in order that non-experts can make sense of it. Thus, it was decided to proceed to the training stage of this study.

9 TRAINING OF LISTENERS

The idea to train humans for sensory evaluation purposes is not new and has been applied in a wide range of disciplines. Watson [40] gave an overview with regard to auditory perception and showed that detection and discrimination skills can be learned. As to audio-related applications, various researchers have successfully dealt with the training of listeners in timbre perception [e.g. 5, 41]. However, in the relatively new field of spatial sound reproduction, a lack of such studies is apparent. Hence, as part of this work a preliminary investigation into the training of listeners for the evaluation of spatial sound reproduction was conducted.

9.1 Experimental design

For the training part of this study, a paid listening panel was set up, which consisted of five music students of the University of Surrey. All listeners were queried regarding their experience and skills in listening to sound in an analytical way, but none of them had received any form of training in this respect, which is why all of them were considered to be naïve listeners. Nevertheless, two subjects indicated an interest in music technology and sound recording, one of which had participated in a listening test before.

To be able to quantify the effect of the adopted training method, a pre-/post-test methodology was employed. All five listeners completed the pre- and post-test. However, only three participants proceeded to the training stage that took place in between the two tests. The other two listeners did not receive any training thereby acting as a control group. Hence, a comparison of the intra- and inter-listener performances and thus an assessment of the true training effect was made possible. The pre- and post-test were identical for each subject but varied across the listeners to balance any interaction effects related to the order of the presentation. The listeners' task was to rank five sounds in terms of their perceived ISW and nine sounds in terms of their perceived source distance. Hence, the candidates' abilities to discriminate several levels for each of the two spatial attributes were verified. The results of the pre- and post-test were then compared in terms of the correctness of the response and the time taken to finish each ranking task.

Due to the preliminary nature of this investigation it was decided not to take the training any further than the ranking exercises outlined above. Yet, ideally, the training of a listening panel goes beyond simple detection of differences between sounds and the ordering of stimuli that differ in their intensities with respect to a specific property. In particular, an expert panel would also be able to quantify perceived differences in absolute and not just relative terms. Further, the ability to discern and describe or categorize a particular auditory feature in the presence of various other subjective impressions is a highly desirable characteristic of a good panel. However, it has to be borne in mind that the objective of this pilot study was simply to get an indication of the suitability of the adopted method for training a listening panel in spatial sound perception. Therefore, to economize the procedure, it was decided to address scaling and categorization at a later stage.

An essential aspect of any training programme is to provide a structured framework for learning in order to allow listeners to develop both skills and confidence [4]. In this case, the programme was split into two parts. While the first half focussed on source distance, the second one addressed ISW. The design of the training was very similar for both spatial attributes. Each training session was scheduled to last no longer than 30 minutes because it was felt that a longer duration would have increased the likelihood of listener fatigue.

During the first training session, the listeners had to listen to two sounds taken from the extreme ends of the attribute scale and to verbally describe the differences they perceived. The intention was to elicit verbal descriptors that were suitable to describe the perceptions of the individual listeners rather than to impose on them potentially non-meaningful terms. This was found to be especially useful in the case of ISW where it turned out that listeners preferred to use terms such as "out of focus/focused" or "fuzzy/defined" to describe their perceptions rather than "wide/narrow". The second step of the training was to introduce the listeners to the physical principles governing the perception of each attribute so as to provide them with a firm background in the underlying modality. By drawing their attention to specific features of the sound that changed due to the applied processing, they were taught what to listen for when making their judgements. During the familiarization phase of the programme, all listeners were exposed to a large array of exemplary stimuli that could serve as a frame of reference. Depending on the spatial attribute, stimulus sets were used that contained two to three (ISW) or three to five (source distance) different intensity levels. Once the listeners felt that they had a good grasp of a given spatial attribute, they were given a couple of training exercises where their ability to detect a difference between two sounds was tested. In general, sounds that represented large differences in intensity were selected so that the panel could gain confidence as well as learn the basic listening skills. The familiarization phase was then repeated until the subjects had listened to the full range of intensities of each attribute and successfully completed the associated exercises.

The true training phase built on the knowledge that the listeners had acquired during the previous stages. This time they completed various exercises that were based on the following tasks:

1. Discrimination
2. Pairwise ranking
3. Multi-stimulus ranking

The exercises usually consisted of 10 trials. If a listener achieved 80% or more correct answers in a given exercise, he/she was moved up one level in the training hierarchy. In case a listener gave four wrong answers, the exercise was terminated and the listener had to go back and start again. If a subject failed to complete a particular exercise twice, he/she had to go back a complete stage until listening skills and confidence were restored. Exercises were made progressively harder by i) choosing a more difficult task, ii) making the differences in intensities between the stimuli smaller and iii) increasing the number of stimuli presented to the subjects in case of the multi-stimulus ranking exercises. Once a subject had successfully completed the final multi-stimulus ranking exercise, i.e. the one with the largest possible number of stimuli (five for ISW and nine for source distance), the training programme was considered to be finished.

Tables 9.1 and 9.2 present some details with respect to the number of sessions each listener needed as well as the total time period over which the training took place.

Table 9.1: Details of source distance training

Listener	No. of 30min sessions	Time period (weeks)
1	4	1
2	3	0.5
3	3	0.5

Table 9.2: Details of ISW training

Listener	No. of 30min sessions	Time period (weeks)
1	7	3.5
2	3	1.5
3	6	1.5

The whole training programme was implemented in software using Max/MSP. The software automatically collects and saves the details of each exercise for further analysis purposes. This includes all interactions with the program as well as the total time spent on completing the various tasks. A 'performance' window is included in the implementation, which displays whether the listeners' responses are correct or not so that learning can take place. To strengthen the learning effect and to make the program more fun to use, cartoon characters were chosen for providing the user with instant aural and visual feedback. The 'performance' window also includes a summary of the listeners' achievements during a particular exercise so that subjects know "how they are doing". During the familiarization stage of the training programme, the auditory spatial changes were also depicted in graphical form in the 'viewer' window to supplement the training effect. Screenshots of the various components of the program are included in Appendix C.

All pre- and post-tests (as well as training sessions) were run in ST3 and the experimental preparations were identical to the ones of the confirmation experiment. The user-interfaces were based on the one shown in Fig. 8.1. All listeners were thoroughly instructed in how to indicate their perceptions and how to control playback of the stimuli using the computer keyboard before the start of the tests. The definitions for source distance and source width used during the confirmation experiment were shown to the subjects and all of them were informed of the importance of restricting head movements to a minimum. No time limit regarding the maximum duration of the tests was set.

9.2 Analysis of results from training programme

To recap, two groups of listeners completed the pre- and post-tests: those that took part in the training programme (i.e. the trained group comprising listeners 1 to 3) and those that did not take part in the training (i.e. the untrained group comprising listeners 4 and 5). Quantification of the success of the training was to be achieved by comparing the results from the pre- and post-tests on an inter- as well intra-listener level. As before, the SED was used to estimate the 'degree of wrongness' of each ranking sequence relative to the correct sequence[2]. In addition, response time was to be taken into account when assessing the suitability of the chosen methodology. To allow comparability of the results, an effort was made to schedule the last training session and the post-test of each trained listener in such a way that the intermediate time period was about equal to the one between the pre- and post-tests for the untrained listeners. The time gaps ranged from three to five days.

The SED results for distance from the trained and untrained group are shown in Fig. 9.1 and 9.2, respectively. It is obvious that for both trained and untrained listeners, SED was lower during the post-test except for listener 2 who ranked all nine sounds correctly during both tests. However, the untrained subjects obtained much higher SEDs than the trained subjects. Admittedly, both listener 1 and listener 4 reduced their SEDs by the same amount (i.e. 6), but listener 1 managed to put the nine sounds into the correct order resulting in a SED of 0 whereas listener 4 was still far away from that goal. Generally speaking, the trained group came very close to achieving a perfect match between the responses and the correct answer during the post-test.

Figure 9.1: Wrongness of rank responses – Source distance, trained group

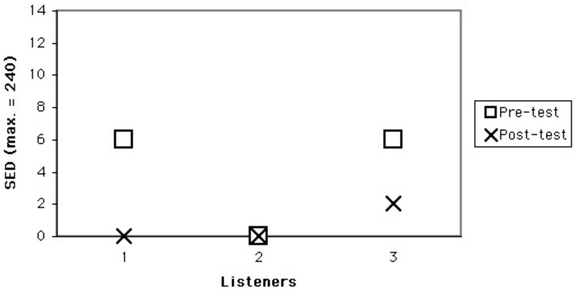
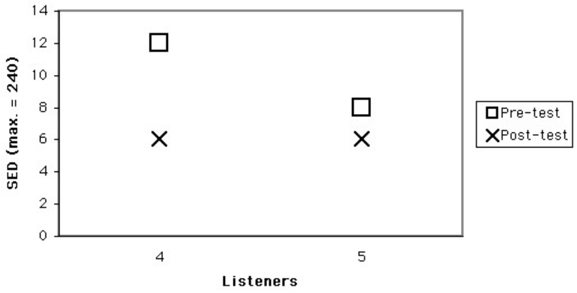


Figure 9.2: Wrongness of rank responses – Source distance, untrained group



If one looks at the listeners' response times from the distance rank test (Fig. 9.3 and 9.4), a significant decrease in the time taken to complete the task is apparent for the trained listeners. For each of them response time dropped by more than 50%, whereas no such trend is detectable for the results from the untrained subjects. More precisely, listener 5 almost had identical response times and listener 4 even needed slightly longer for the post-test. Interestingly, listener 2, who ranked all sounds correctly from the start, took considerably longer than all other subjects to complete the pre-test and was still the slowest trained subject after completion of the training programme.

It may be argued that because of continued usage of the software during the training sessions, the trained group became more familiar with operational aspects and therefore managed to complete the post-test faster. However, the magnitude of the reduction in response time seems to be too big to justify this merely on the basis of learning how to use the software on behalf of the listeners.

Figure 9.3: Response times – Source distance, trained group

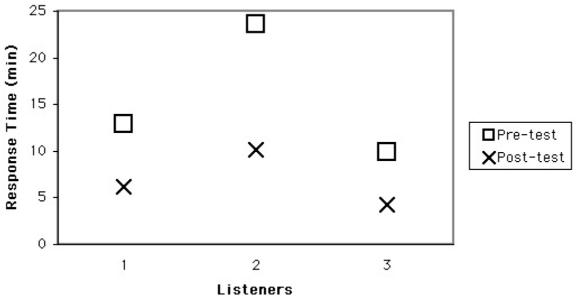
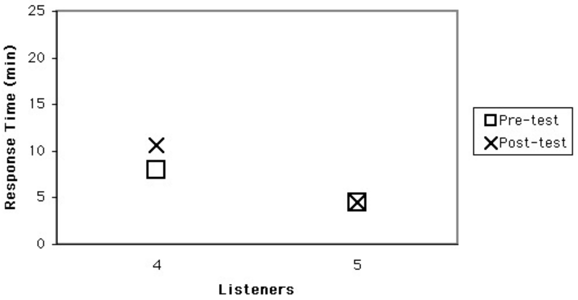


Figure 9.4: Response times – Source distance, untrained group



Looking at Fig. 9.5 and Fig. 9.6, the results from the ranking tests on ISW display similar trends to the ones identified for distance. Once more, listener 2 achieved a remarkably low SED during the pre-test, which remained the same for the post-test. Hence, no training effect can be inferred from these results. However, with regard to listeners 1 and 3, a clear drop in SED can be seen in their post-tests, which points towards an effective training programme. This impression is reinforced by the fact that the

[2] Perhaps a word of caution is appropriate at this point. The SED has the disadvantage that the results from the distance rank test are not directly comparable to the results from the ISW rank test. This is because the former has a maximum SED of 240 whereas in the case of the latter the maximal SED is 40. While normalizing the two sets of data could have easily rectified this problem, this was considered not beneficial because the original SED values directly reflect the closeness between the subjects' responses and the correct rank sequence. For instance, a SED of 2 corresponds to the reversal of an adjacent pair of sounds and so forth.

results from the untrained group do not reveal any similar tendencies. For instance, listener 2 (trained) and listener 4 (untrained) obtained the same large SED during their pre-tests, but while listener 1 performed much better during the post-test, listener 4 managed to improve only slightly. Conversely, listener 5 achieved the second lowest pre-test SED of all five listeners, but obtained a worse result during the post-test.

Lastly, it is worth pointing out that no listener achieved a SED of 0, even though only five ISW sounds were used compared to nine stimuli for distance. Thus, it can be inferred that because of different degrees of stimulus uncertainty subjects found ranking the ISW samples much harder than the distance sounds.

Figure 9.5: Wrongness of rank responses – ISW, trained listeners

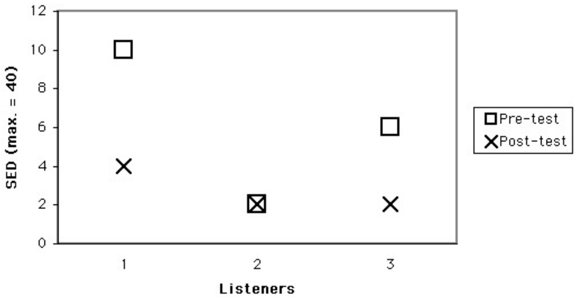
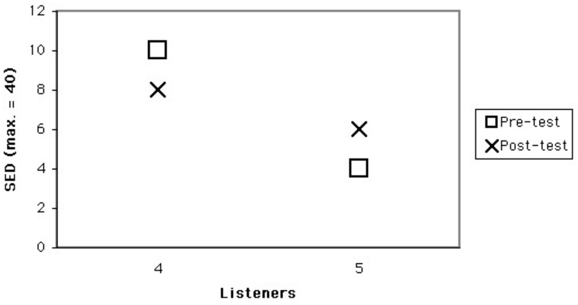


Figure 9.6: Wrongness of rank responses – ISW, untrained listeners



When looking at the response times from the ISW rank tests as a whole (Fig. 9.7 and 9.8), it can be seen that generally they were lower across all subjects compared to the tests on distance, which was to be expected since there were fewer stimuli to compare. However, while all trained listeners improved considerably in terms of the time needed to rank the distance stimuli, this was not the case for ISW. It is true that listeners 2 and 3 managed to complete the post-test slightly faster, but not so listener 1 who needed slightly longer. Likewise, listener 4 was slower during the post-test as well, but in her case response time increased by about 3min compared to only 1min for listener 1. Similar to the test on distance ranking, listener 5 achieved almost the same response time for the two tests on ISW. It is also worth noting that listener 2 took longest again out of all trained listeners to complete both pre- and post-test.

Figure 9.7: Response times – ISW, trained group

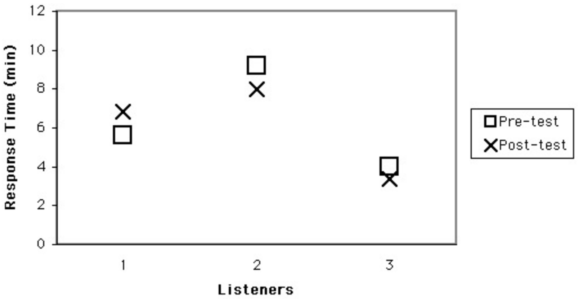
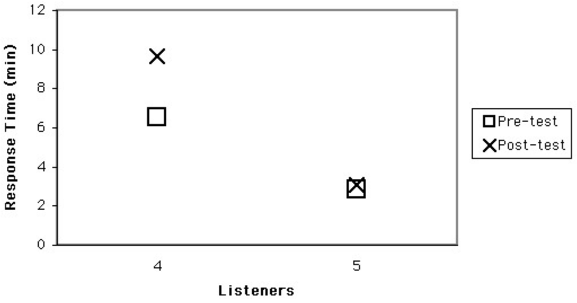


Figure 9.8: Response times – ISW, untrained group



So overall, slight decreases in the time taken to complete the task can be seen for the trained but not for the untrained group. Hence, the results indicate that while the training had a definite positive effect on the correctness of the subjects' responses, little improvements were made regarding speed. This seems to be in line with the previous finding that the changes in ISW are harder to detect and judge than the changes in source distance, which probably led to the minor improvements in the subjects' response times even after several training sessions.

A final look at the training details presented in Tables 9.1 and 9.2 provides further evidence. Both listeners 1 and 3 needed (almost) twice as many training sessions for ISW to reach a similar level of performance compared to source distance[3]. Other researchers [40, 43] addressed the aspect of duration of training programmes in more detail and showed that the rate of learning of subjects is highly dependent on the complexity of the perceptual task whereby an increase in the latter results in a prolonged training period.

Hence, the above supports the conclusion drawn from the results of the validation and confirmation experiment, namely that ISW is an unfamiliar perceptual construct, which is much more difficult to assess than source distance. Nonetheless, as the results of the training programme show, it is possible to train listeners in that respect.

10 SUMMARY AND CONCLUSIONS

The primary objective of this study was to create exemplary source distance and ISW stimuli that would vary in a unidimensional way to allow them being used for the training of naïve listeners in spatial sound evaluation. Therefore, the

[3] The fact that listener 2 only needed three sessions for both distance and ISW to achieve the desired level of performance warrants further mentioning. During a conversation it turned out that in addition to his interest in music technology, he also had nearly perfect pitch. Thus, it might be speculated that listener 2 had a stronger 'perceptual awareness' leading to superior performances during the pre-tests. On the other hand, the fact that he needed significantly longer to complete the tasks somewhat weakens this argument. At the least, his results show that subjects can differ considerably in terms of their learning rates, an aspect discussed in great detail by Bech [42].

development of appropriate processing methods formed a major part of this work. The development was based on perceptual considerations rather than on physical models for reasons of simplicity as well as user-friendliness.

As the above discussion clearly showed, the generated distance stimuli could illustrate the intended effect of changes in the proximity of a sound source to experienced listeners in an unequivocal manner. Therefore, the claim can be made that the intended unidimensionality of the sound excerpts exemplifying changes in source distance was achieved.

The verification of the source width samples turned out to be more problematic, which was mainly attributed to the subjects' unfamiliarity with this concept. It is proposed that this caused listener confusion resulting in the appearance of the perceptual construct of source position. However, rigid definition of the notion of ISW by expert listeners showed that subjects could adapt and successfully apply it to a number of ISW stimuli after they had been instructed to do so.

In this context, it also has to be said that a relatively large number of references were made with regard to changes in timbre. Yet, there did not seem to be any consensus between the subjects as to which frequency region was most affected, which is why their perceptual relevance turned out to be very low.

So on the whole, the results of the various experiments demonstrated that the predominant differences between the ISW stimuli were related to changes in the width of the sound source. It is acknowledged that the results are not 100% conclusive, but while there may be a weak second dimension, there is no clear evidence for it and it would not appear to be orthogonal to the first one. Based on this finding, it was decided that the source width stimuli were acceptable for training purposes and did not need to be modified any further at this stage.

As a consequence, both source distance and ISW stimuli were employed during an initial training programme aimed at teaching a group of naïve listeners in the evaluation of spatial attributes of reproduced sound. The results obtained from the devised method left no doubt that the subjects improved in terms of the correctness of their responses for both distance and ISW. Moreover, response time reduced significantly in the case of distance and slightly for source width. This difference was traced back to the higher level of stimulus uncertainty, which had been identified during the validation stage of the sound examples.

11 FURTHER WORK

This investigation formed just the starting point of an ongoing study. The two spatial attributes that were addressed were taken out of a pool containing several more. Hence, the expansion of the spatial training toolkit with respect to other qualitative impressions of spatial sound reproduction is one aspect to be addressed.

As was also mentioned before, improvements regarding the training software itself are envisaged, e.g. in the area of real-time controllability. This capability would offer the advantage of allowing for active training tasks to be incorporated into the programme, thus letting the users manipulate the various dimensions themselves rather than pre-specifying the changes for them. This could facilitate the training of listeners with respect to the more difficult scaling and categorization exercises that are to be included in the training programme.

Another main concern is to determine what exactly caused the ISW changes to disappear when moving the stimuli to the LR. Informal listening tests conducted so far have shown that this was mainly due to the different loudspeakers used in the two listening environments. Therefore, the identification of the responsible physical parameters is planned in order to be able to give full particulars regarding the transferability of the sound excerpts.

12 ACKNOWLEDGEMENTS

The authors would like to thank Rhonda Wilson of Meridian Audio Ltd. for her valuable contributions during various discussions as well as Meridian Audio Ltd. and the AES Educational Foundation for financial support.

13 REFERENCES

[1] Berg, J., Rumsey, F. 1999: 'Spatial attribute identification and scaling by Repertory Grid Technique and other methods', *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction*, 10-12 April, pp. 51-66

[2] Berg, J., Rumsey, F. 2001: 'Verification and correlation of attributes used for describing the spatial quality of reproduced sound', *Proceedings of the AES 19th International Conference on Surround Sound*, Schloss Elmau, Germany, June 21-24, pp. 233-251

[3] 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. 5425

[4] Meilgaard, M., Civille, G. V., Carr, B. T. 1991: *Sensory evaluation techniques*, CRC Press, Inc.

[5] Olive, S. E. 1995: 'A method for training listeners and selecting program material for listening tests', *Audio Engineering Society Preprint*, 97th Convention, preprint no. 3893

[6] Olive, S. E. 2001: 'A new listener training software application', *Audio Engineering Society Preprint*, 110th Convention, preprint no. 5384

[7] Chowning, J. M. 1971: 'The simulation of moving sound sources', *Journal of the Audio Engineering Society*, vol. 19, no. 1, pp. 2-6

[8] Moore, F. R. 1983: 'A general model for spatial processing of sounds', *Computer Music Journal*, vol. 7, no. 6, pp. 6-15

[9] Kendall, G. S., Martens, W. L. 1984: 'Simulating the cues of spatial hearing in natural environments', *Proceedings of the International Computer Music Conference*, Paris, pp. 111-125

[10] Martin, G., Corey, J., Woszczyk, W., Quesnel, R. 2001: 'A computer system for investigating and building synthetic auditory spaces - part 2', *Proceedings of the AES 19th International Conference on Surround Sound*, Schloss Elmau, Germany, June 21-24, pp. 75-83

[11] Jullien, J.-P., Kahle, E., Marin, M., Warusfel, O. 1993: 'Spatializer: A perceptual approach', *Audio Engineering Society Preprint*, 94th Convention, preprint no. 3465

[12] Jot, J.-M. 1997: 'Real-time spatial processing of sounds for music, multimedia, and interactive human-computer interfaces', *ACM Multimedia Systems Journal*, special issue 'Audio and Multimedia' (February)

[13] Nielsen, S. H. 1991: 'Depth perception - Finding a design goal for sound reproduction systems', *Audio Engineering Society Preprint*, 90th Convention, preprint no. 3069

[14] Blauert, J. 1997: *Spatial hearing*, The MIT Press, Cambridge Massachusetts, London England

[15] Gerzon, M. A. 1992: 'The design of distance panpots', *Audio Engineering Society Preprint*, 92nd Convention, preprint no. 3308

[16] Gerzon, M. A. 1992: 'Signal processing for simulating realistic stereo images', *Audio Engineering Society Preprint*, 93rd Convention, preprint no. 3423

[17] Griesinger, D. 2000: 'The theory and practice of perceptual modeling – How to use electronic reverberation to add depth and envelopment without reducing clarity', *Preprint 21st Tonmeister Conference*, Hannover, Germany, Nov. 24-27

[18] Rumsey, F. 1998: 'Subjective assessment of the spatial attributes of reproduced sound', *Proceedings of the AES 15th International Conference on Small Room Acoustics*, Copenhagen, Denmark, Oct. 31-Nov. 2, pp. 122-135

[19] 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

[20] Hansen, V., Munch, G. 1991: 'Making recordings for simulation tests in the Archimedes project', *Journal of the Audio Engineering Society*, vol. 39, no. 10, pp. 768

[21] ITU-R BS 1116. 1994: 'Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems', International Telecommmunications Union, Recommendations ITU-R BS 1116, pp. 267-297

[22] Griesinger, D. 1992: 'Measures of spatial impression and reverberance based on the physiology of human hearing', *Proceedings of the AES 11th International Conference*, Portland, Oregon, pp. 114-145

[23] Michelsen, J., Rubak, P. 1997: 'Parameters of distance perception in stereo loudspeaker scenario', *Audio Engineering Society Preprint*, 102nd Convention, preprint no. 4472

[24] Beranek, L. 1996: *Concert and opera halls - how they sound*, Woodbury, NY: Acoustical Society of America

[25] Rumsey, F. 2001: *Spatial Audio*, Focal Press

[26] Gerzon, M. A. 1986: 'Stereo shuffling: New approach – old technique', *Studio Sound*, vol. 28, no. 7, pp. 122-130

[27] Corey, J., Woszczyk, W., Martin, G., Quesnel, R. 2001: 'An integrated multidimensional controller of auditory perspective in a multichannel soundfield', *Audio Engineering Society Preprint*, 111th Convention, preprint no. 5417

[28] Sibbald, A.: *ZoomFX for 3-D sound*, Sensaura White Paper

[29] Gerzon, M. 1992: 'Optimum Reproduction Matrices for Multispeaker Stereo', *Journal of the Audio Engineering Society*, vol. 40, no. 7/8, pp. 571-589

[30] See http://www.catt.se/ for details.

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

[32] Hair, J. F., Anderson, R. E., Tatham, R. L., Black, W. C. 1995: *Multivariate data analysis: with readings*, Prentice-Hall, New Jersey

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

[34] Anderberg, M. R. 1973: *Cluster analysis for applications*, Academic Press, New York

[35] Schiffman, S. S., Reynolds, M. L., Young, F. W. 1981: *Introduction to Multidimensional Scaling*, New York: Academic Press

[36] 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

[37] Ericsson, K. A., Simon, H. A. 1984: *Protocol Analysis - Verbal reports as data*, MIT Press

[38] Martin, G., Woszczyk, W., Corey, J., Quesnel, R. 1999: 'Controlling phantom image focus in a multichannel reproduction system', *Audio Engineering Society Preprint*, 107th Convention, preprint no. 4996

[39] See http://www.cycling74.com/products/maxmsp.html for details.

[40] Watson, C. S. 1980: 'Time course of auditory perceptual learning', *Annals of Ontology, Rhinology and Laryngology*, vol. 89, pp. 96-102

[41] Quesnel, R., Woszczyk, W. 1994: 'A computer-aided system for timbral ear-training', *Audio Engineering Society Preprint*, 96th Convention, preprint no. 3856

[42] Bech, S. 1992: 'Selection and training of subjects for listening tests on sound-reproducing equipment', *Journal of the Audio Engineering Society*, vol. 40, no. 7/8, pp. 590-610

[43] Bech, S. 1993: 'Training of subjects for auditory experiments', *Acta Acustica*, vol. 1, no. 3-4, June/August, pp.89-99

APPENDIX A: EXPERIMENTAL SET-UPS

Figure 1: Graphical illustration of Studio 3 set-up

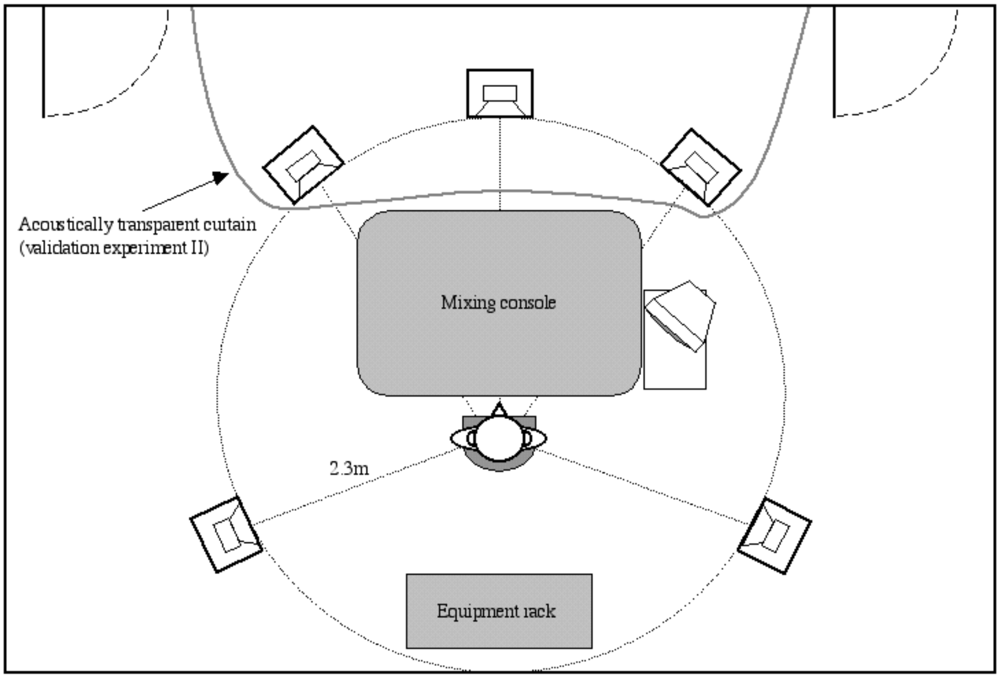
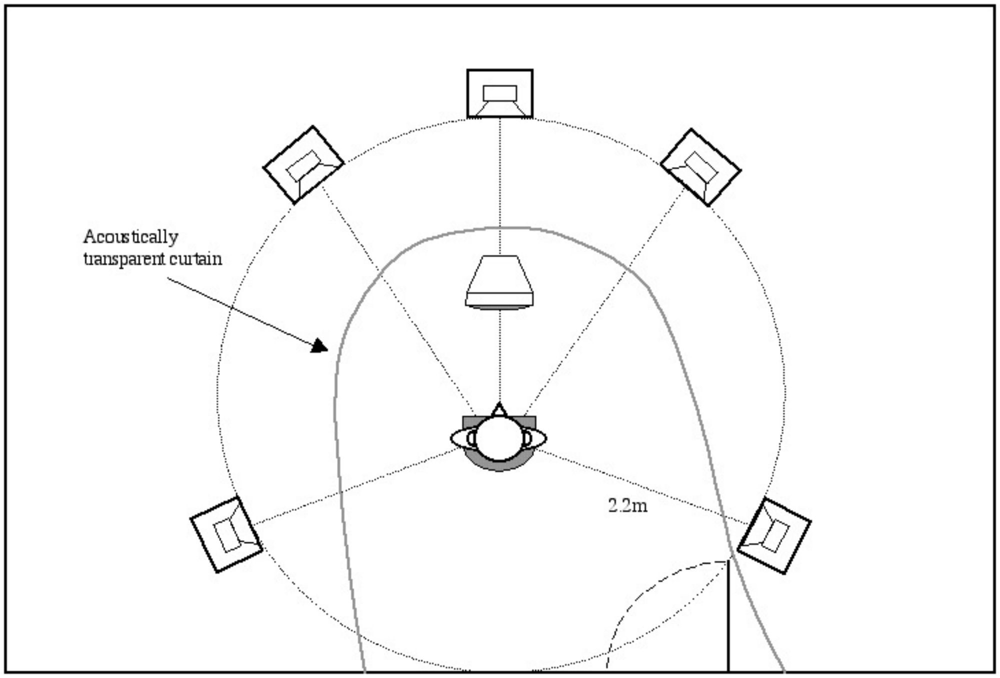


Figure 2: Graphical illustration of listening room set-up



APPENDIX B: BLOCK DIAGRAMS OF PROCESSING METHODS

Figure 1: Distance processing

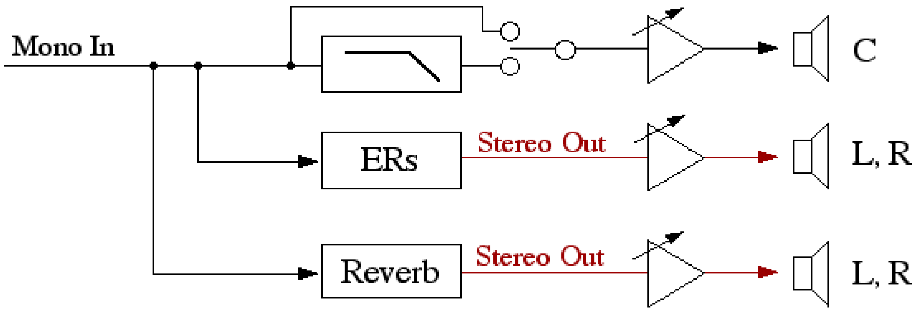
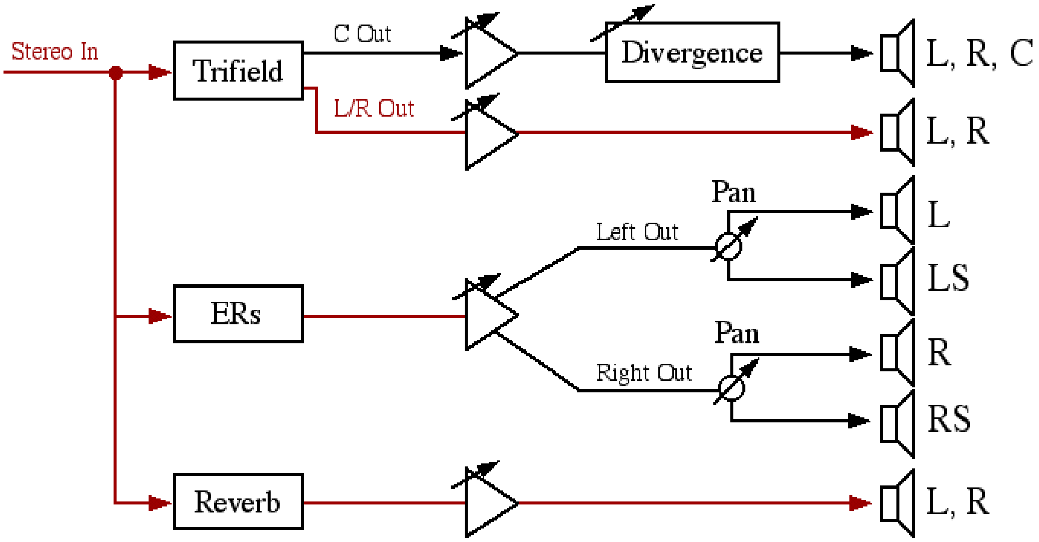
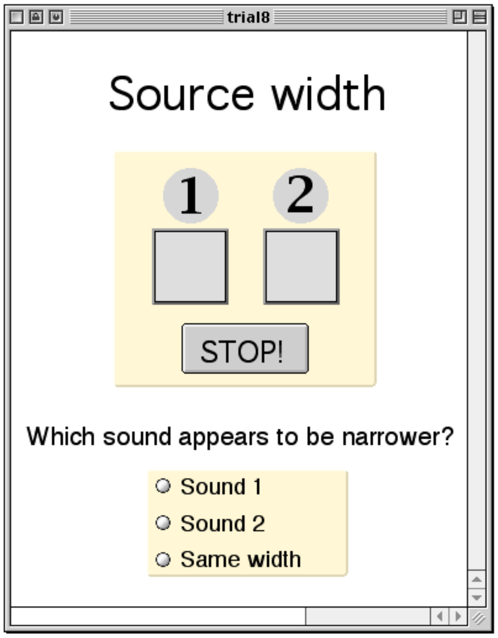


Figure 2: ISW processing



APPENDIX C: SCREENSHOTS OF TRAINING SOFTWARE

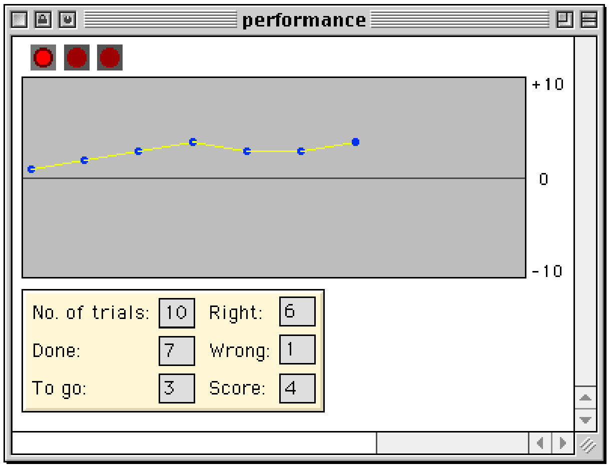
Figure 1: Typical trial window



Which sound appears to be narrower?

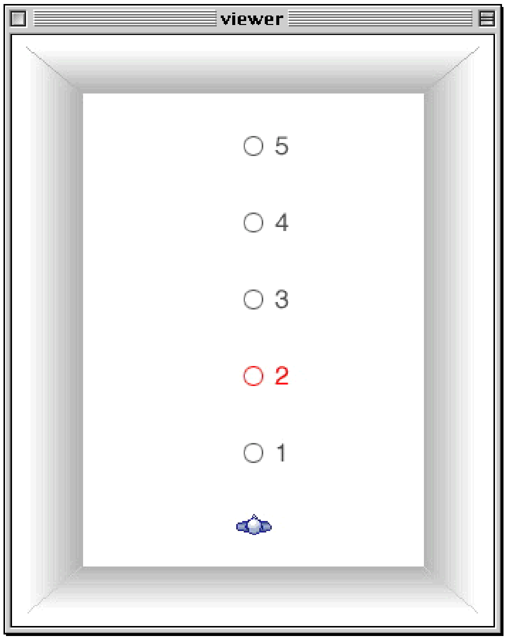
- ○ Sound 1
- ○ Sound 2
- ○ Same width

Figure 2: 'Performance' window



No. of trials:	10	Right:	6
Done:	7	Wrong:	1
To go:	3	Score:	4

Figure 3: 'Viewer' window showing graphical representation of source distance


viewer
○ 5
○ 4
○ 3
○ 2
○ 1