Audio Engineering Society

Convention Paper 6977

Presented at the 121st Convention
2006 October 5–8 San Francisco, CA, USA

A Comparison between Spatial Audio Listener Training and Repetitive Practice

Rafael Kassier[1], Tim Brookes[2], and Francis Rumsey[3]

Institute of Sound Recording, University of Surrey, Guildford, GU2 7XH, UK
[1] raf@rafaelkassier.com, [2] t.brookes@surrey.ac.uk, [3] f.rumsey@surrey.ac.uk

ABSTRACT

Despite the existence of various timbral ear training systems, relatively little work has been carried out into listener training for spatial audio. Additionally, listener training in published studies has tended to extend only to repetitive practice without feedback.

In order for a generalised training system for spatial audio listening skills to prove effective, it must demonstrate that learned skills are transferable away from the training environment and it must compare favourably with repetitive practice on specific tasks.

A novel study has been conducted to compare a generalised training system with repetitive practice on performance in spatial audio evaluation tasks. Transfer is assessed and practice and training are compared against a control group for tasks involving both near and far transfer.

1. INTRODUCTION, BACKGROUND & MOTIVATION

Interest in spatial audio has increased due to the availability of multichannel reproduction systems for the home and car. Despite various timbral ear training systems having been presented [1-6], relatively little work has been carried out into training in spatial attributes of reproduced sound.

Perhaps the greatest strides in this direction have been taken by Neher [7]. Neher performed a pilot experiment into listener training for spatial audio attributes, and produced a series of unidimensionally varying spatial audio attribute stimulus sets. He argued that in order to train listeners in the perception of spatial audio attributes, one must be able to exemplify changes in specific attributes in the auditory modality. Neher's pilot experiment involved just 5 listeners, but the results indicated that training in detection of differences between, and ranking of, spatial audio attributes can benefit listener performance using the *same* set of stimuli.

It was hypothesised by the current authors that participation in a listener training programme concerned

with the spatial aspects of sound reproduction would also help to create listeners that are more consistent and sensitive when evaluating spatial changes in audio reproduction using a *different* set of stimuli to those used in training (so called *transfer of training*). In order to demonstrate its usefulness outside the context of the stimuli used in training, any training scheme would need to show that learned skills were *transferable*. Transfer of training is frequently classified in terms of *near* and *far* transfer [8, 9]. *Near* transfer is applicable when the trained and target situation and stimuli are similar to one another. *Far* transfer refers to target situations and/or stimuli that are different to the trained situations and stimuli.

In documents such as [10], the terms *training* and *familiarisation* (where the procedures involved in listening tests are explained to, and practised by, the test subjects) are used interchangeably, and in [11] *training* could be better described as *practicing* the task. For this research, *training* refers to a separate process where skills are taught and practised in a context not necessarily identical to the test conditions. The study presented in this paper investigates the difference between *training* and *practice* in spatial audio evaluation tasks.

The goal for the current research project is to work towards the eventual creation of a system to help train listeners in spatial audio evaluation. Such a system could find use in industrial product evaluation (for example automotive multi-channel audio systems or home theatre audio systems), or in ear-training for sound engineers, audiophiles and hobbyists.

According to Shaw and Gaines [12] ambiguities can result when different words are used to describe the same phenomenon, the same words are used to describe different phenomena and different words are used to describe different phenomena. This can result in confusion as to what is meant by one person and understood by another. Care must therefore be taken when selecting appropriate terms to use in the description of spatial audio phenomena.

The first concern addressed by the authors was the need for a spatial audio description language that could be used as the framework within which to base the training system. This description language needed to conform to various criteria, such as the need for unambiguous terms that did not overlap conceptually with one another. The resulting *Simplified Scene-Based Paradigm* was published in [13].

Once this framework had been established, a study was conducted to establish whether or not trained spatial audio listening skills could be transferred from one task and stimulus to another task and stimulus set [14]. Sixteen listeners were tested and placed into two groups of equivalent performance in a spatial audio attribute rating task. One of the groups underwent a formal training programme (a modified implementation of the one used by Neher [7]) which trained the detection and ranking of differences in a spatial audio attribute (Individual Source Width) using a separate set of contrived stimuli (provided by Neher [7]). The other group did not take part in any additional training. There was an established "correct" order in which to rank the items, so it was therefore possible to measure the correctness of each trainee's response. The trained group showed a significant improvement in the way that they ranked the audio stimuli used in the training scheme. Both groups were then retested on the spatial audio attribute rating task. The only transferred training effect observed was in the way the subjects used the 0-100 point scale to rate the items. The trained subjects used a significantly greater range of the scale to express their judgements after training, whereas the non-trained subjects used a significantly smaller range of the scale to express their judgements. No change was seen in either group relating to their consistency or fluency.

The observed lack of transfer of training from the training environment to the task of rating spatial audio attributes (a more ecologically valid task) is a central issue in this research. Issues relating to transfer of training and transfer experiment design were investigated [15] in order to inform further study.

It is possible that the lack of transfer occurred because the rating task was too difficult (and indeed, even experienced listeners struggled to be consistent and sensitive when responding). Another likely factor was a potentially demotivating aspect of the training programme which involved negative feedback being given for incorrect answers in the form of a cartoon character and comic sound effect. Furthermore alternative levels of transfer might have been achieved but had not been investigated.

In order to optimise the current training programme for transfer, as wide a variety of transfer as possible was sought. Near transfer would need to be investigated by including test environments that were identical to the training environment. Far transfer would be

investigated using a modified version of the training environment, and also a completely new scenario (which would test further transfer than a similar environment). In order to encourage transfer by decontextualising the stimuli, analogies and a wide variety of stimuli and tasks would need to be used during training. Encouraging trainees to reflect upon what they have learned is also considered to be beneficial for transfer [16].

The study reported in this paper was motivated by two issues. The first was the inconclusive nature of the results from the previous experiment. Whilst the training system showed dramatic results using its own stimuli, transfer to a different situation was severely limited. The current experiment needed to be designed to investigate various forms of transfer in order to be able to discover transferred skills that were potentially hidden in the previous study. The second motivation was the reliance in previous studies [10, 11] on repetitive practice. If a generalised spatial audio attribute training programme is to be shown to be useful it will not only need to be transferable, it should compare favourably with repetitive practice and indeed with no training or practice regime at all.

This report will describe the design of the generalised spatial audio attribute training programme used in the study, before going on to describe the experiment used to evaluate the training programme against repetitive practice and a control group.

2. SPATIAL AUDIO ATTRIBUTE TRAINING SYSTEM

The training system described in this paper is set within the context of the Simplified Scene-Based Paradigm for spatial audio scene description [13], and follows from Tobias Neher's work in the creation of validated multi-channel stimulus sets that each vary in a single perceptual spatial audio attribute [7]. It maintains elements of the training system used in the previous study [14].

Perceptually unidimensionally changing spatial audio stimulus sets are key to the system, as they allow for the demonstration of various levels of each spatial audio attribute in an unambiguous manner. This not only allows the student an opportunity to learn and practice with each stimulus set, but it also allows accurate verification of each subject's perceptual skill. Neher simulated four spatial audio attributes [7] (as defined in [17]). These were Individual Source Distance, Individual Source Width, Ensemble Width and Ensemble Depth. He created and validated a stimulus set for each attribute, and provided multi-channel audio processing platform which can create stimulus sets from mono source recordings.

Ensemble Width and Ensemble Depth (Scene Component Width and Scene Component Depth of multi-source Scene-Components using the nomenclature in [13]) were chosen as the attributes that would be trained in the current study. The inclusion of two attributes in the current study had the advantage of adding variety. This aids in the decontextualisation of the stimuli and hence boost transfer [16], whilst allowing for an expanded range of task difficulty. In order to further increase variety Neher's previously validated processing platform settings were used to create a number of new stimulus sets that, whilst not rigorously validated, were informally evaluated by the authors and found to be suitable simulations. The stimuli were created using acoustically dry recordings of individual instrumentalists playing in a variety of ensembles. Six were chosen to feature in the training system and a further six were chosen to feature in some of the additional tests. Both sets contained programme items with similar musical styles. Each stimulus set featured nine different levels of either Scene Component (SC) Width or Scene Component (SC) Depth of one of six four-source ensembles. It is worth noting that Neher's original Ensemble Depth stimuli also contained four sources, but his Ensemble Width stimuli had five sources. For these experiments SC Width stimuli were generated without a centre source, but this had little effect on the perceptual illusion of the widening of an ensemble of sources. It is worth noting here that feedback received from subjects during the previous training study had called for additional stimuli to be used in training. The use of these varied stimulus sets would address this issue.

The spatial audio attribute training system consists of three main phases:

- Tutorial
- Active learning using the Spatial Audio Toolkit
- Self-administered training drills with feedback.

The training system conforms with Alessi & Trollip's model for successful instruction [18], which has four elements:

- Information presentation
- Learner guidance
- Practice
- Assessment

Presenting information was achieved through an individual tutorial administered by the main author using a computer-based graphical presentation. In the tutorial the need for a universal spatial audio description language was explained and the Simplified Scene Based Paradigm was presented. Visual analogies [9] were used to elicit responses from the trainees and Neher's validated stimulus sets were used as audio examples [19]. During playback of ensemble stimulus sets, trainees were asked to describe how each scene component changed considering what individual sources were doing as well as the ensemble. This *mindful abstraction* is particularly useful for *far transfer* [16]. The tutorial could be administered interactively and exclusively through a self-administered computer package, but at this developmental phase there was more to be gained by the main author in interacting directly with the trainees.

Guiding the learner was performed during the tutorial, and also during the *Spatial Audio Toolkit* phase and the drills phases. The *Spatial Audio Toolkit* is shown in Figure 1 and Figure 2 is in the style of a *constructivist learning environment* [18] and allows the learners to experiment with the various stimuli for SC Width and SC Depth. Individual sources can be muted or soloed and the overall width or depth setting for each ensemble can be selected via a 9-point slider. The *Spatial Audio Toolkit* is actually a simple interface controlling Neher's processing platform. The original single-source files used to create the new stimulus sets are sent through the processing platform and can be muted, soloed or changed at will. Presets for the nine different levels of the attribute could also be selected. Being given control is a powerful motivating force for learners [20, 21], but the main idea behind the *Spatial Audio Toolkit* is to provide a learning environment where the trainees can perform *discovery learning* [18] initially guided by the main author, but then eventually constructing their own knowledge using the toolkit.

The practice and assessment phases are handled via a self-guided test regime. There are two types of test: Discrimination ("are these the same or different?") and Pairwise Ranking ("which of these wider/deeper?"). Both involve the comparison of two items drawn from a randomly selected pool [22]. There are four different difficulty levels for each test. Difficulty level one selects randomly from a pool containing just the most extreme stimuli in the set (stimuli 1 and 9). Difficulty level two adds stimulus 5 (the mid-point) to the pool. Difficulty level three adds stimuli 3 and 7 to the pool, and difficulty level four includes all nine stimuli. In order to train for fluency a "traffic light" system is used to mark the start of each test (green), the half-way point (yellow) and end point (flashing red). During training tests a given trial is marked as incorrect if the user does not respond within 20 seconds (when the red lights flash). Because of the modular design of the system, additional tests can be easily accommodated.

Several motivational devices have been implemented in order to maximise interest and willingness to participate in the tests. Users are given control [20, 21] over the difficulty and the test task as well as the attribute and stimulus set used. Each user has their own board of proficiency indicators (green lights), one for each difficulty level of each element of the tests. Subjects are challenged to complete as many tasks as the can, switching on as many lights as possible in the time available to them. The criteria for completing a task are that at least 20 trials need to have been attempted, and 80% need to have been correctly answered. It is possible for someone who has not achieved the pass mark after the 20^{th} trial to continue the test until they increase their overall score to 80%. The 80% passmark has been carried over from previous studies [7, 14], and can be adjusted if necessary. Progress is tracked with a numerical and graphical display showing the number of trials attempted and the percentage of correct answers given. Upon completion the user is rewarded with a window displaying a smiling face and the corresponding proficiency indicator light is switched on.

The *Spatial Audio Toolkit* and the self-tests were implemented using the Max/MSP programming language. Each user is assigned a unique number along with their first name. Their progress is saved every time they complete a task. In addition most interactions that they make with the software are logged for subsequent analysis.

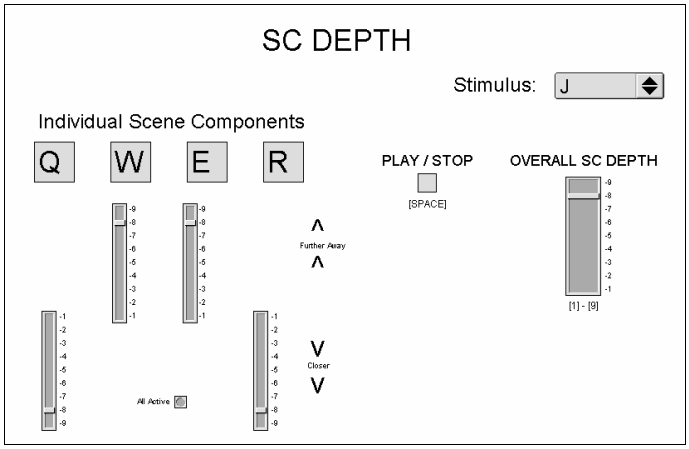

Figure 1: Spatial Audio Toolkit showing SC Depth.

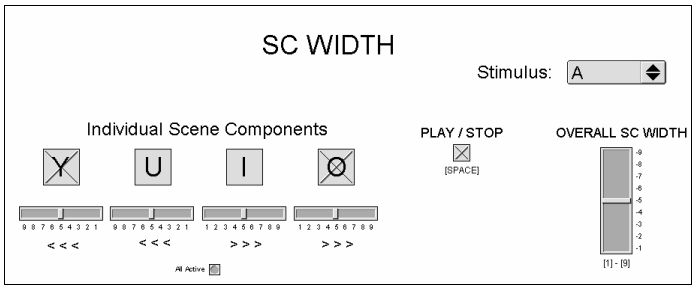

Figure 2: Spatial Audio Toolkit showing SC Width.

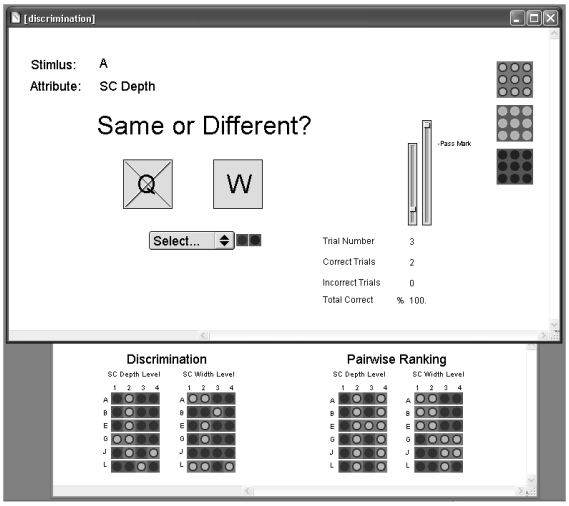

Figure 3: Spatial Audio Training Tests, showing a discrimination test in progress. Note the motivational light board below the main window, and the traffic light timer on the top right.

3. TRANSFER INVESTIGATION – OUTLINE

In order to gauge the effectiveness of a spatial audio attribute training regime, it was necessary to compare it against two control groups, one that repetitively practiced the task and one that did no additional training or practice. Comparison with non-trained subjects allowed the overall training effect to be quantified. Comparison with repetitive practice (as described in [10] and [11]) allowed the training system to be gauged against a previously established method.

Because potential transfer effects, especially *near* transfer effects, could have been missed during the previous experiment [14], a range of transfer tests were devised in order to evaluate the training system.

The task of rank-ordering the contrived stimuli was found to be an effective way of evaluating training during Neher's pilot experiment [7] and the previous experiment conducted by the authors [14]. A rank-ordering task therefore formed the basis of the current study. The stimuli used in the training were the six stimulus sets used in the training system.

Subjects were pre-tested using a rank-ordering task and their resulting performance used to separate them into three groups of approximately equal skill. One of these groups was trained using the spatial audio attribute training system, another repetitively practiced the initial task and the third did no additional practice or training. Thereafter the three groups were tested once again using the ranking task in order to compare their performance in *near* transfer.

To test for *far* transfer, two different transfer scenarios were used. Firstly the post-test task was repeated using a different set of stimuli (the other six stimulus sets not used in the training system). Examining the performance (between groups rather than pre-post) in these tasks would therefore indicate how effectively training and practice would transfer to stimuli other than those practiced on. Secondly, the stimuli were reproduced in a different manner than the original (contrived) stimuli to make them more ecologically valid. This resulted in stimuli where many different attributes of the sound reproduction changed. The ability of subjects to discern and describe a particular sensory characteristic in a "sea" or "fog" of other sensory impressions is more important than sensory acuity [23]. If training or practice were shown to improve performance with such stimuli (whether using the *near* or *far* transfer test paradigms) then this would be powerful evidence for their wider usefulness.

Therefore the following hypotheses were tested:

- Both the trained and practice groups will show improved performance in the *near* transfer test and *far* transfer tasks over the untrained group, and over their previous performance. (Because practice and *near*-transfer training will aid the initial test).

- The practice group will show improved performance over the trained group for the *near* test and stimuli, because they practiced on a task and stimuli closer to the initial task.

- The trained group will show improved performance over the practice group for the other transfer tests. Because more decontextualised training and varied examples will lead to greater *far* transfer.

3.1. Experimental Set-Up

The listening tests, practice and training all took place in the Listening room at the University of Surrey. This room conforms to ITU-R recommendation BS. 1116 [10] and features five active loudspeakers (Genelec 1032A). The loudspeakers were placed 2.2m from the listening position in the 3/2 stereo configuration [24]. The tests, practice and training were administered via an Apple Macintosh G4 computer running Max/MSP 4.5 from Cycling '74. The computer was situated in an adjacent room connected to a Universal Serial Bus (USB) keyboard, mouse and 17" video monitor in the listening room via extended cables. During the training phase, a notebook computer was connected to the 17" monitor in order to display the tutorial presentation to the trainees.

3.2. Selection of Subjects

Subjects for this experiment were recruited from the first year undergraduates on the University of Surrey's Music and Sound Recording (Tonmeister) course. Tonmeister students are expected to be part of the target group for the spatial audio attribute training system, as they can be expected to be motivated to improve their

listening skills. Initially 18 students signed up for the experiments (15 male, three female), but three requested to *not* be considered for the training phase due to work pressure.

All 18 subjects were asked to partake in the pre-test. Because of the limited availability of test subjects, the 3 subjects who had expressed a wish *not* to take part in the additional training phases were included at this stage, as it was assumed that they could be used in the non-trained control group in order to maximise the sample sizes.

The performance of the subjects in the pre-test was examined by evaluating how their rank-ordering data matched the expected order, and how long they took to complete the tasks. The intention was to separate the subjects into three groups, each of which had approximately equal performance characteristics.

In order to investigate how accurately the stimuli were ranked before and after the training system, the sum of the squares of the Euclidean distances (SSED) between the correct rank order and those provided by the subjects was calculated. For this, the difference in the rank order number between the expected and subjective results was calculated (called the Euclidean distance). The Euclidean distance was then squared (to make any differences occur in magnitude only, not direction), and then summed across the nine available stimuli. This gives a sum of the squared Euclidean distances (SSED) for the particular rank order provided by the subject. For rank ordering of nine stimuli, the maximum SSED (and hence the most incorrect rank order possible) would be 240. Over six pages, the maximum possible SSED would be 1440, and the summed total across both attributes would be 2880.

SSED and total time taken was calculated for each of the tasks (six pages of ED and six pages of EW) for each of the subjects. The summed totals of both SSED and timing was used as overall performance measures for each subject (shown in Table 1).

Groups were created by first attempting to balance total SSED then total time taken of three groups of six subjects. Because three of the subjects had agreed to take part in the experiment on condition that they were not required to do the additional training phase, they would need to be placed into the control group together. The three subjects in question were subjects 1, 2 and 16. As can be see from Table 1, these three subjects actually displayed the worst performance in terms of total SSED (subject 2 had nearly four times the average SSED, subjects 1 and 2 had nearly twice the average SSED score). Because one of the groups of six subjects would need to incorporate the three worst-performing subjects, it was found to be impossible to create three balanced groups of six subjects. For this reason, and in the interests of correct experimental technique (allowing a random assignment of subjects to experimental groups), subjects 1, 2 and 16 were excluded from the analysis. In terms of their participation in the experiment, they were treated as being part of the control group (ie: they were required to attend the post-training transfer tests), but their results were not used in the analysis. In actual fact Subject 1 did not attend the real-world transfer tests, and Subject 16 pulled out of all tests. Subject 2 attended all transfer tests, but his data was not included in the analysis because he was not part of the selected subjects.

With the excluded subjects eliminated from the subject pool, it was found to be possible to create three very closely balanced groups of subjects based upon SSED. The secondary issue of time taken was addressed by swapping subjects between groups such that the SSED remained similar, but the time taken for the tasks was more closely aligned. One subject (15) proved to be difficult to accommodate, as his time taken was 66% of the average time taken for all subjects (the next fastest subject- 6, was only 86% of the average time taken for all subjects). This meant that whichever group Subject 15 was assigned to would have a lower total time taken than the others.

Once the 3 groups were created, they were randomly assigned to be either the *training* group, the *practice* group or the *control* group. The resulting groups are shown in Table 1.

As can be seen from Table 2, group totals of SSED were able to be balanced to within ± 2 and the total time for the groups, whilst being influenced by Subject 15, were nevertheless very close.

Table 1: Subjects' Pre-Test Performance (* indicates excluded subjects)

Subject Number	Total SSED	Total Time Taken (s)
1*	442	1432
2*	896	1308
3	226	1416
4	218	1609
5	374	1392
6	168	1203
7	284	1361
8	174	1515
9	210	1314
10	70	1270
11	188	1548
12	122	1627
13	128	1371
14	244	1658
15	250	926
16*	418	1352
17	176	1539
18	114	1226

Table 2: Subject group performance

Training			Practice			Control		
Subject	SSED	Time	Subject	SSED	Time	Subject	SSED	Time
5	374	1392	7	284	1361	3	226	1416
6	168	1203	8	174	1515	4	218	1609
11	188	1548	9	210	1314	15	250	926
12	122	1627	10	70	1270	17	176	1539
13	128	1371	14	244	1658	18	114	1226
Total	**980**	**7141**	**Total**	**982**	**7118**	**Total**	**984**	**6716**

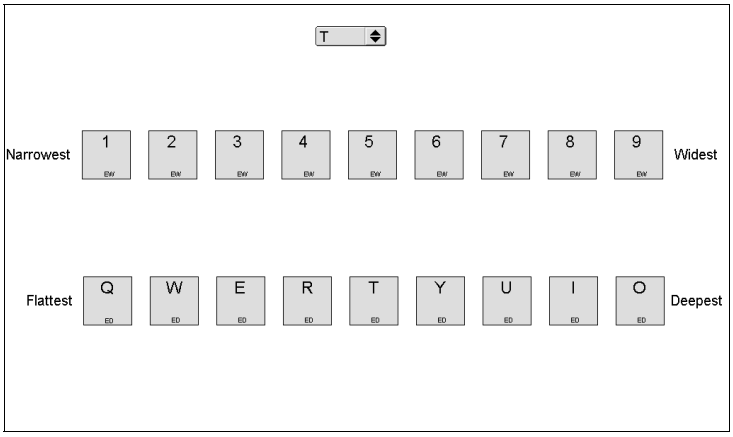

Figure 4: Familiarisation Screenshot. Programme items are changed with the drop-down menu.

3.3. Familiarisation & Initial Practice

Before beginning the pre-test procedure, the subjects had a chance to read through a description of the experiment including definitions of the terms "Ensemble Width" and "Ensemble Depth" that would be used in the tests (these terms were used instead of SC Width and SC Depth during the pre- and post-tests because not all subjects would be trained to use the Simplified Scene-Based Paradigm). They also had a chance to listen to the nine labelled gradations of Ensemble Width and Ensemble Depth for each of the six programme items, as well as Neher's validated stimulus sets (see Figure 4). Once they were familiar with the items they had a chance to practice ranking five of the nine items (stimuli 1, 3, 5, 7 and 9) against the clock, and with feedback. There was a notional time-limit of a minute (where-after the red "traffic lights" would flash). However they were not forced to progress at this point, but rather requested to finish the rank ordering as soon as possible, and continue to the next page. Feedback was provided to give the subjects an idea of how they were doing, and to give them the confidence to attempt the main task. The programme items used in the pre-test were used in the five-stimulus ranking practice. The last stage of the initial practice phase was a "test-conditions" practice of ranking all nine versions of Neher's validated Ensemble Width and Ensemble Depth stimuli (on separate pages) without feedback, and with a notional limit of two minutes. This allowed each subject to experience the time-pressure and complexity of the main task using validated stimuli, but not to pre-bias any particular subject to any particular test programme item. This proved to be a very valuable phase, as it allowed the clarification of the test procedure to at least one subject who had become confused and had not spotted that they had ranked two items in the same position. Once familiarisation and initial practice had been completed subjects progressed to the pre-test rank ordering exercise.

3.4. Rank-Ordering (Pre-Test, Near Transfer and Far Transfer Tasks)

The pre-test and two of the post-tests used a similar test procedure, which involved the rank ordering of a number of pages of nine stimuli according to their SC Width or SC Depth. For the pre-test and near transfer test, there were six programme items, each containing two sets of nine stimuli (with varying levels of SC Depth and SC Width respectively). These were the same stimuli that were used in the training system. For the far transfer test, six different programme items were used, but otherwise the tests were identical to the pre-test.

Each test began with a chance to use the familiarisation page in order to re-acquaint the listeners with the stimuli, and give them a chance to prepare themselves for the task (their so-called *set* [25]). During the tests, subjects were asked to rank-order six pages of Ensemble Width and six pages of Ensemble Depth items. They were given the choice of which order to attempt the attributes in order to make them more comfortable during the experiment.

Subjects were asked to complete the ranking of each page (Figure 5) of nine stimuli within two minutes, but there was no automatic progression. Stimuli were auditioned by clicking on-screen buttons with the computer mouse, or by pressing a key on the computer keyboard. Subjects then used nine 9-position sliders to assign rank positions to each of the stimuli on the page. Once each of the nine ranks had been assigned, the computer allowed the subject to move to the next page. Once six pages of one of the attributes had been completed, the subjects could move onto the other attribute. When both attributes had been completed, the session was over. Ranking and timing information and the human-computer interaction log were saved along with the subject's profile by the computer. Each session lasted about 30 minutes.

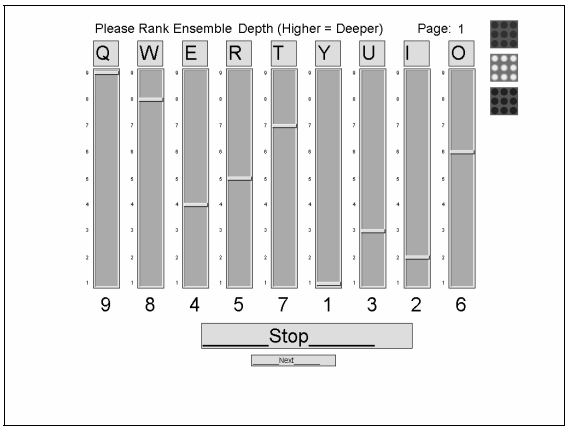

Figure 5: Ranking task screenshot.

3.5. Training & Practice Phase

The training and practice groups took part in 6 additional sessions each. This was done to balance the amount of additional time that each group had spent in the listening room. The sessions were scheduled during one week. This allowed for about six half-hour sessions for each of the subjects in the training and practice groups. The practice group performed six additional iterations of the pre-test.

For the training group, the first additional session consisted of an explanation of the training system, the tutorial and a guided practice session with the *Spatial Audio Toolkit*. The next four sessions consisted of self-paced practice with the *Spatial Audio Toolkit* and self-administered discrimination and pairwise ranking tests. The final session consisted of a repetition of the pre-test in order to gauge progress against the performance of the practice group's final session.

The control group were informed that they had not been randomly assigned to receive additional training, and were asked to report back for the transfer tests during the week following the training. After the training and practice phase, all three groups took part in the two ranking transfer tasks and the far transfer task (attribute rating).

3.6. Spatial Audio Attribute Grading (Far Transfer Task)

In order to investigate far transfer away from the task and stimuli used in the training and practice sessions, a completely new scenario was devised. The requirements were that there needed to be a spatial audio attribute evaluation task that would involve complex, more ecologically valid stimuli in order to test transfer from the training and practice environments to a more ecologically valid task.

The experimental paradigm decided upon was the grading of one spatial audio attribute on a 0-100 point scale over three iterations. Consistency and sensitivity in the grading data could be evaluated over the three iterations for the three experimental groups and the individual subjects in order to compare their relative performances.

Alterations in the relative positions of elements in an ensemble give rise to naturally occurring changes in ensemble width and ensemble depth. Changes in microphone technique and configuration also result in changes in spatial audio attributes (as demonstrated in [26]).

Simultaneous multiple microphone recordings had proved to be a convenient way of creating a series of switchable and complex multichannel stimuli in previous experiments [14, 26]. This method was therefore employed to create the varying stimuli needed for the experiment.

However, if elements of an ensemble needed to be recorded in various positions, a highly repeatable performance is essential. Any small changes in the timing or feeling of the performance would be recognisable when switching between stimuli recorded at different times. Because more control was needed, it was decided that a repeatable acoustic playback system would be to provide the sound sources. In order to provide a degree of continuity with the previous training and practice sessions, the original mono source recordings used to create the stimuli for the training system were used.

The experimental recording session took place in Studio 1 at the University of Surrey's Department of Music and Sound Recording. The studio is 14.5m wide, 17m long and is approximately 6.5m high. It is primarily used for the recording of classical music.

Figure 9 shows the layout of the recording session. The sound source stimuli were replayed via four loudspeakers (Genelec 1032A) arranged in various configurations toward the front of the studio. The recording set-up consisted of three triplets of microphones positioned facing towards the front of the studio. The three techniques were chosen from the techniques already used in [26]. This consisted of a Fukada triplet [27] (using AKG C451 cardioid microphones), an OCT-inspired technique [26, 28] (using an AKG C414 B-ULS cardioid as the centre microphone and two AKG C414 B-XLS hypercardioid microphones as the side microphones), and an INA-3 technique [29] (using AKG C414 B-ULS cardioid microphones). The three triplets were mounted on a bespoke microphone stand that centred all triplets, and were raised to a height of 220cm from the ground. A spaced cardioid technique was used to capture ambience, and was implemented using two B&K 4011 microphones at a height of 3.4m from the ground.

These were positioned towards the rear corners of the studio, facing the corners (to reject as much direct sound as possible). All microphones were level-aligned using a known source held a fixed distance from the capsule of each microphone. All microphones were connected to a digital audio workstation (DAW) via similar microphone preamplifiers. Four outputs of the DAW were connected to the four loudspeakers in order to replay the source files. This enabled the DAW to simultaneously record the 11 microphone channels whilst replaying the four source files, allowing for a very repeatable procedure for each recording pass.

The loudspeakers were repositioned to create different physical widths and depths of ensemble between recording passes. Each of the twelve programme items used during the ranking tasks were recorded in six different loudspeaker configurations. Because each triplet was used twice (once in the forward position, once in the back position) this created twelve unique 5-channel recordings for each of the twelve programme items.

Figure 6, Figure 7 and Figure 8 show photographs of the experimental recording session.

Figure 6: Photograph of front triplets in the "front" position.

Figure 7: Photograph of the recording studio, taken from behind the loudspeakers. Note the frontal triplets in the centre of the picture (actually in the "back" position, next to the staging), and the two spaced cardioids extended on either side of the studio.

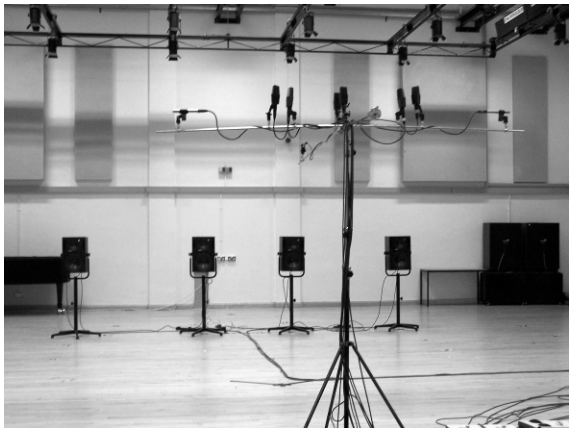

Figure 8: Photograph of the recording session, taken from behind the frontal triplets and showing the four loudspeakers in one of the configurations.

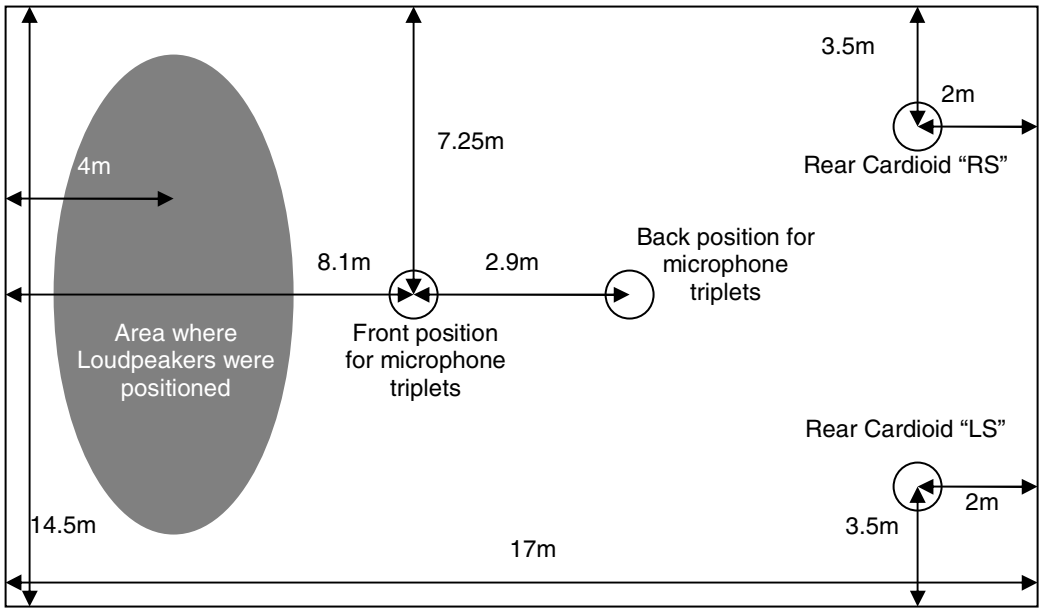

Figure 9: Plan View of the Experimental Recording Session.

Informal listening by the main author conducted in the listening room determined that, although ensemble depth changes were apparent, these were not as obvious as the ensemble width changes. A wide variety of ensemble width changes were found across the various stimuli, and it was therefore selected as the attribute to be used in the grading experiment. Out of the entire collection of recordings three configurations of loudspeakers were selected (with the microphone triplets in the "frontal" position), and all three microphone techniques were used. This gave three configurations and three microphone techniques for each programme item, resulting in nine items to grade per page. To allow for three iterations of the test during a single 30 minute session, four programme items were selected (allowing for two minutes per page). Two were taken from the training stimuli, and two were taken from the far-transfer ranking stimuli. All four featured distinct musical styles.

Each subject took part in one far-transfer rating test. After reading through the test instructions, the subjects were given the opportunity to familiarise themselves with the stimuli and begin to place rate them using the 0-100 point scale. Each of the four programme items could be selected using a drop-down menu, subjects could use the nine sliders to assign a grading to each of the stimuli on the page. The familiar "traffic-light" timer reset every time a new programme item (labelled "stimulus") was selected. The familiarisation and practice screen is shown in Figure 10.

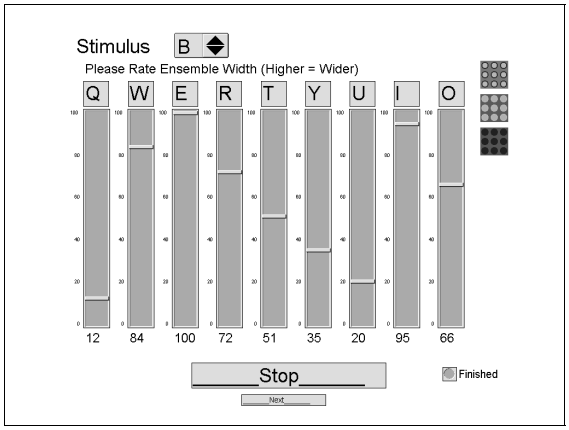

Figure 10: Screenshot of the rating task familiarisation and practice screen.

Once subjects were happy to move on, they began the ranking task. This consisted of twelve pages of nine stimuli to rank. Within the trial there were three blocks of four pages, where all four programme items would be evaluated. Subjects were warned that the same programme item could appear on subsequent pages, and instructed to pay attention to the incrementing page number. The computer randomised the presentation for each subject, and once all twelve pages were completed, it saved their grading data with most of the interactions that they made with the interface for subsequent analysis. Figure 11 shows the rating task test screen. The "next" button would appear to allow progression to the next page once every slider had been moved in some way (which guarded against accidental progression).

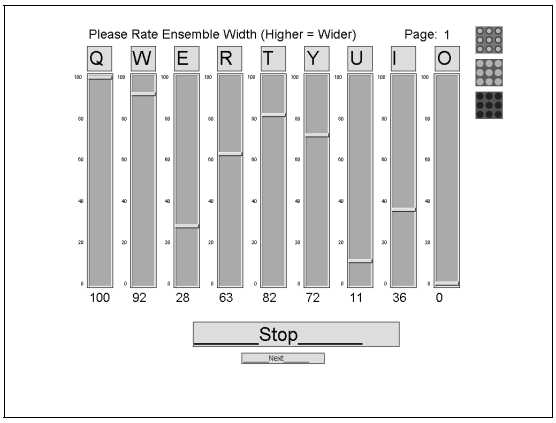

Figure 11: Screenshot of the rating task test screen.

4. RESULTS

Referring back to the three hypotheses of interest:

- Both the trained and practice groups will show improved performance in the *near* transfer test and *far* transfer tasks over the untrained group, and over their previous performance. (Because practice and *near*-transfer training will aid the initial test).

- The practice group will show improved performance over the trained group for the *near* test and stimuli, because they practiced on a task and stimuli closer to the initial task.

- The trained group will show improved performance over the practice group for the other transfer tests. Because more decontextualised training and varied examples will lead to greater *far* transfer.

These were tested using the data obtained from the pre-test, near-transfer ranking test, far-transfer ranking test and far-transfer post-test.

4.1. SSED Ranking Data

Ranking data (SSED and time taken) for each group was examined using dependant and independent non-parametric tests (due to small sample sizes).

Taking the near- and far-transfer ranking tests to begin with, Figure 12 shows SSED data and Figure 13 shows time taken for the three experimental groups.

From Figure 12, it is likely that significant decreases in wrongness-of-rank occurred for the practice and training groups between the pre-test and either of the far-transfer tests. It looks as though the practice group either got worse or stayed the same between the two tests.

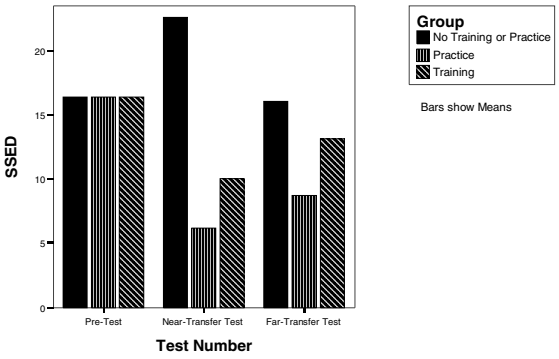

Figure 12: Mean SSED data for the three groups over the three test phases.

These observations were formally verified with the use of non-parametric Wilcoxon Signed Ranks tests [30]. No significant change in SSED was evident for the control group between the pre-test score and either of the transfer tests, meaning that the subjects got no better or no worse but did not improve over their previous scores.

The practice group's SSED did reduce significantly between the pre-test and near-transfer tests ($z = -4.21$, $p < 0.01$, $r = -0.38$).

The training group's SSED also reduced significantly between the pre-test and near-transfer tests (z = - 3.18, p < 0.01, r = - 0.29).

There was no significant difference found between the practice and training group's SSED in the near transfer test, meaning that their near-transfer performance in wrongess-of-rank data was similar after their respective regimes.

In addition, practice SSED was found to be better (lower) than control SSED (U = 1089.5, p < 0.001, r = -0.34), and training SSED was found to be better (lower) than control SSED (U = 1292.50, p < 0.01, r = -0.24) during the near-transfer tests.

The only significant reduction in SSED that occurred between the pre-test and far-transfer test was in the practice group (U = 1160.00, p < 0.001, r = -0.31), however a break-down showed that this was due to "ensemble depth" test scores, rather than "ensemble width" scores (which did not change). Interestingly the training, practice and control SSED scores in the far-transfer tests showed no significant differences between each other. This was most probably due to the small sample size available and the fact that the stimuli were different between the two tests, which meant that independent non-parametric tests needed to be used which further reduce the statistical power of the analysis (Mann-Whitney tests [30]). It is important to note that although the practice group's SSED was shown to reduce significantly, there were no significant differences between the overall performances during the far-transfer task between any of the groups.

As far as SSED data is concerned, the hypotheses examined in this study are only partially proven. Both trained and practice groups showed improved performance in their near transfer test, but only the practice group showed improvement in the far-transfer ranking task.

The practice group did indeed improve more than the other groups in the near transfer task. However, they also showed improved performance over the other groups in the far-transfer ranking task, the training group did not show a significant improvement here.

4.2. Timing Data of Ranking Tests

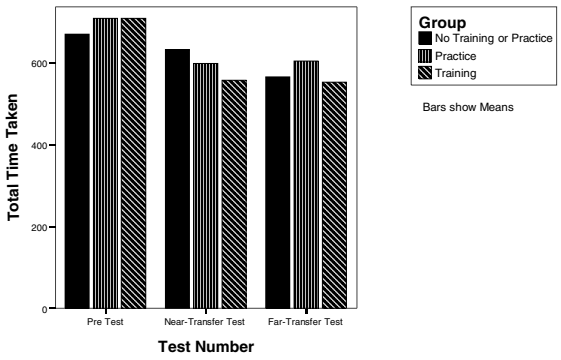

Figure 13: Mean total time taken data for the three groups over the three test phases.

Means of total time taken for the ranking tests were plotted in Figure 13. Non-parametric tests were used to discover any significant changes within groups and across groups. It was found that the control group's time taken did not reduce significantly between the pre-test and either transfer tests. However, the total time taken for the practice group decreased significantly in the near-transfer test (z = -2.49, p < 0.05, r = -0.56) and far-transfer test (U = 12, p < 0.01, r = -0.64). The total time taken for the trained group also decreased significantly in the near-transfer test (z = -2.80, p < 0.01, r = -0.62) and far-transfer test (U = 13, p < 0.01, r = -0.62). Times were not, however significantly different between the groups (again, probably because of the need for independent data analysis methods.

Returning once again to the hypotheses, analysis of the timing data has shown that the trained and practice groups did show an improvement in their performance, however these far-transfer performances were not significantly different to the control group. It is worth noting that Subject 15 was in the control group, they had completed the pre-task much quicker than the others. There were no significant timing differences between any of the groups in the final transfer test.

4.3. Far-Transfer Rating Data

The rating data was used to establish measures of consistency and sensitivity by running an analysis of variance (ANOVA) on the grades provided for each individual stimulus (each version of each programme

item) by each subject individually. The resulting sum of squares is a measure of the inconsistency with which that subject graded the particular stimulus over the three iterations. The estimate of effect size "Partial Eta Squared" was also used as a measure of how sensitive the subjects were to differences between the individual stimuli. Time taken was also analysed.

Sum of Squares (consistency of grading), Partial Eta Squared (sensitivity) and time taken for the test (fluency) data were examined using a Kruskal-Wallis test in order to find significant differences in any of these three measures across the three groups. None were found, implying that there were no significant differences in the way in which the three groups rated the various stimuli.

This is likely to be due to the small size of the groups, and it is hoped that further planned experimentation and analysis can help shed light on any far transfer that may have been obscured. It is also possible that the complexity of consistently rating nine stimili resulted in an experiment that was too difficult for the subjects regardless of their prior training.

5. CONCLUSIONS

The first striking finding in this experiment was that the three least motivated subjects produced the three worst performances during the pre-test. There is evidence in the literature that motivation assists learning and transfer (in [19] for example), but this experiment has shown how lack of motivation can negatively affect subjects' performance.

In answer to the three hypotheses to be covered by this study, current data analysis has revealed answers to part of them.

Both trained and practice groups showed improved performance regarding SSED and time taken during the near-transfer ranking tasks. In addition, time-taken was seen to improve for the trained and practice groups during the far-transfer ranking task. Far-transfer to other scenarios and stimuli was not shown to exist by the measures and number of subjects employed here. The performance of the practice group did indeed show improved performance over the trained group for the *near* transfer tasks, supporting hypothesis number 2. The performance of the trained group has not been shown to be superior to the practice group for far-transfer tasks.

Due to the constrained number listeners available and the amount of time available in the listening room, it was only possible to run the entire experiment over the course of three weeks, one for pre-testing, one for training and one for the post-training tests. This has reduced the statistical power available. Planned further work should allow for a further 15-30 subjects to be either added to the pool, or tested using a similar experiment to that described here.

Additional limitations from using such a small sample set include the need to use non-parametric data analysis and the susceptibility to any temporary influences (such as mood swings or late nights) that can occur in each subject from time to time.

From the data analysis performed to-date, it is possible to say that the training system performed as well as repetitive practice without feedback. Both systems performed were beneficial when compared with a control group for similar stimuli. For different stimuli, repetitive practice has helped more than the training system, although both were beneficial. For different stimuli and situations, only the group drilled with repetitive practice showed any sign of positive transfer. Further analysis and study will be necessary to draw firm conclusions regarding the comparative merits of the spatial audio training system in its current form with respect to repetitive practice without feedback.

6. FURTHER WORK

Recently a new method for examining the ranking data has been suggested [31] and will be implemented as soon as possible. It involves treating each rank as a single data point, increasing the sample size and potentially allowing parametric analysis to be conducted.

Additional ways of examining the data will be investigated, including examining how listening strategies change with practice and training. In order to accomplish this a method has been devised by which the user interface interaction stored by the computer administering the tests can be displayed graphically and examined. In addition, the rating experiments will be conducted with experienced listeners in order to create a performance 'yardstick' against which to measure the

performance of the three listener groups. Answers given during the grading experiment will also be examined to see how the way in which subjects rate the sounds concurs with other group members and with other listeners, including the experienced listeners. This will check whether the trained or practice listeners begin to fall into agreement with one another after their respective regimes.

A further experimental session is planned to coincide with the next academic year, and it is anticipated that 15-30 additional students can be recruited to take part. The purpose of this could be to add to the current data set in order to increase the statistical power and hence uncover previously hidden transfer effects. The experiment could also be adjusted in light of the additional analysis.

7. ACKNOWLEDGEMENTS

The research presented in this paper is supported by a studentship grant from the Engineering and Physical Sciences Research Council (EPSRC).

8. REFERENCES

1. Quesnel, R. and W.R. Woszczyk. *A Computer-Aided System for Timbral Ear Training*. Presented at the *AES 96th Convention*. 1994. Amsterdam: Preprint No: 3856.
2. Quesnel, R. *Timbral Ear Trainer: Adaptive, Interactive Training of Listening Skills for Evaluation of Timbre Differences*. Presented at the *AES 100th Convention*. 1996. Copenhagen: Preprint No: 4241.
3. Miskiewicz, A., *Timbre Solfege: A Course in Technical Listening for Sound Engineers.* Journal of the Audio Engineering Society, 1992. **40**(7/8): p. 621-625.
4. Letowski, T., *Development of Technical Listening Skills: Timbre Solfeggio.* Journal of the Audio Engineering Society, 1985. **33**(4): p. 240-244.
5. Olive, S.E. *A Method for Training Listeners and Selecting Program Material for Listening Tests*. Presented at the *AES 97th Convention*. 1994. San Francisco, CA: Preprint No: 3893.
6. Brixen, E.B. *Spectral Ear Training*. Presented at the *AES 94th Convention*. 1993. Berlin: Preprint No: 3474.
7. Neher, T., *Towards A Spatial Ear Trainer*, Doctoral Thesis. *Department of Music & Sound Recording*, University of Surrey, Guildford, 2004.
8. Detterman, D.K., *The Case for the Prosecution: Transfer as an Epiphenomenon*, in *Transfer on Trial: Intelligence, Cognition, and Instruction*, D.K. Detterman and R.J. Sternberg, Editors. 1993, Ablex: Norwood, NJ. p. 1-24.
9. Clark, R.E. and A. Voogel, *Transfer of Training Principles for Instructional Design.* Educational Communication and Technology Journal, 1985. **33**(2): p. 113-123.
10. ITU-R BS.1116-1, *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*. 1994-1997, International Telecommunications Union: Geneva, Switzerland.
11. Bech, S., *Selection and Training of Subjects for Listening Tests on Sound-Repoducing Equipment.* Journal of the Audio Engineering Society, 1992. **40**(7): p. 590-610.
12. Shaw, M.L.G. and B.R. Gaines, *Comparing Conceptual Structures: Consensus, Conflict, Correspondence and Contrast.* Knowledge Acquisition, 1989. **1**(4): p. 341-363.
13. Kassier, R., T. Brookes, and F. Rumsey. *A Simplified Scene-Based Paradigm for Use in Spatial Audio Listener Training Applications*. Presented at the *AES 117th Convention*. 2004. San Francisco, CA.: Preprint No: 6292.
14. Kassier, R., T. Brookes, and F. Rumsey. *A pilot study into listener training for spatial audio evaluation*. in *Proceedings of the Digital Music Research Network (DMRN) 1st Summer Conference*. 2005. Glasgow.
15. Kassier, R., T. Brookes, and F. Rumsey. *Designing a Spatial Audio Attribute Listener Training System for Optimal Transfer.* Presented at the *AES 120th Convention*. 2006. Paris: Preprint No: 6819.
16. Salomon, G. and D.N. Perkins, *Rocky Roads to Transfer: Rethinking Mechanisms of a Neglected Phenomenon.* Educational Psychologist, 1989. **24**(2): p. 113-142.
17. Rumsey, F., *Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm.* Journal of the Audio Engineering Society, 2002. **50**(9): p. 651-666.

18. Alessi, S.M. and S.R. Trollip, *Multimedia for Learning: Methods and Development*. Third ed. 2001, Needham Heights, MA.: Allyn and Bacon.
19. Ellis, H., *The Transfer of Learning*. The Critical Issues in Psychology, ed. M.H. Marx. 1965, New York: Macmillan.
20. Keller, J.M. and K. Suzuki, *Use of the ARCS motivation model in courseware design*, in *Instructional designs for microcomputer courseware*, D.H. Jonassen, Editor. 1988, Lawrence Erlbaum: Hillsdale, NJ. p. 401-434.
21. Malone, T.W. and M.R. Lepper, *Making Learning Fun: A Taxonomy of Intrinsic Motivations for Learning*, in *Aptitude, Learning, and Instruction*, R.E. Snow and M.J. Farr, Editors. 1987, Lawrence Erlbaum Associates, Inc.: Hillsdale, NJ. p. 223-253.
22. Salisbury, D., *Effective drill and practice strategies*, in *Instructional designs for microcomputer courseware*, D.H. Jonassen, Editor. 1988, Lawrence Erlbaum: Hillsdale, NJ. p. 103-124.
23. Meilgaard, M., G.V. Civille, and B.T. Carr, *Sensory Evaluation Techniques*. Second ed. 1991, Boca Raton, Fl.: CRC Press.
24. ITU-R BS.775-1, *Multichannel stereophonic sound system with and without accompanying picture*. 1992-1994, International Telecommunications Union: Geneva, Switzerland.
25. Sternberg, R.J. and P.A. Frensch, *Mechanisms of Transfer*, in *Transfer on Trial: Intelligence, Cognition, and Instruction*, D.K. Detterman and R.J. Sternberg, Editors. 1993, Ablex: Norwood, NJ. p. 25-38.
26. Kassier, R., et al. *An Informal Comparison between Surround-Sound Microphone Techniques*. Presented at the *AES 118th Convention*. 2005. Barcelona, Spain: Preprint No: 6429.
27. Fukada, A., K. Tsujimoto, and S. Akita. *Microphone Techniques for Ambient Sound on a Music Recording*. Presented at the *AES 103rd Convention*. 1997: Preprint No: 4540.
28. Theile, G. *Multichannel natural recording based on psychoacoustic principles*. in *Audio Engineering Society 19th International Conference*. 2001. Schloss Elmau, Germany.
29. Herrmann, U. and V. Henkels. *Main Microphone Techniques for the 3/2-Stereo-Standard*. in *20th Tonmeister Tagung*. 1998. Karlsruhe, Germany.
30. Field, A., *Discovering Statistics Using SPSS*. 2nd ed. ISM Introducing Statistical Methods, ed. D.B. Wright. 2005, London: Sage Publications.
31. Busenitz, K. and M. Mendoza, *Personal communication with the author*. 2006.