Verification and correlation of attributes used for describing the spatial quality of reproduced sound

Jan Berg[1] **and Francis Rumsey**[2]

[1]School of Music, Luleå University of Technology, Sweden

[2]Institute of Sound Recording, University of Surrey, Guildford, UK

When the spatial quality of reproduced sound is to be assessed, knowledge of the dimensions forming the quality is essential since the quality is known to be multi-dimensional. The dimensions could be indicated by attributes describing them. Attributes encountered during a previous experiment are considered by a group of subjects and their responses are analysed for finding the attributes' applicability and dimensionality over an extended number of sound stimuli.

INTRODUCTION

One main characteristic of the more recent audio recording and reproducing systems is an increased ability to reproduce the spatial features of sound. Spatial features can be exemplified by location of sounds, sense of the acoustical environment in which the sound source is located, or as expressed in [1], "the three-dimensional nature of the sound sources and their environment". The performance of a sound system in that respect could be denoted as "spatial quality".

A knowledge and an understanding of the factors or dimensions forming and thereby affecting the spatial quality in a sound reproduction system is essential in the different fields of audio work:

- recording – for knowing the chosen recording technique's influence on the spatial quality of the recording;
- post-production – for knowing how to enhance spatial quality;
- reproduction – for knowing possibilities and limitations of different modes of reproduction, e g two-channel mixdowns of multichannel recordings;
- coding – for assessing audio coding algorithms' impact on spatial quality;
- verification of 'objective' measurements – for verifying the correlation between measurements of physical parameters and effects perceived by listeners;
- audio quality assessments – for evaluating the spatial quality part of total audio quality.

One key issue for those involved in audio work – from recording to reproduction – is the perceived result at the listener's end. Since this raises the question of which components in a reproduced sound that are perceivable or not, and how these are interrelated, methods used in the behavioural sciences have to be considered for getting a better understanding of the phenomena. The central problem seems to be how to 'measure' a person's (a subject) conception of an auditory event. This information has to be elicited and communicated from the subject in some way. Ways of collecting and analysing data are reviewed by Rumsey in [2]. The methods available for this mainly rely on verbal or graphical techniques. Examples of verbal techniques are the Repertory Grid Technique (used in Personal Construct Psychology) [3, 4, 5] and Quantitative Descriptive Analysis (used in food research) [6]. Graphical elicitation has been discussed and used by Mason et al [7] and Wenzel [8]. Both verbal and graphical methods have their advantages and disadvantages. The authors' approach is to use verbal communication with the subjects.

In an attempt to find the dimensions of spatial quality, an experiment was conducted in 1998. The experiment is described in [1], and its approach was to try to elicit information from the participating subjects by playing a number of reproduced sounds to them, where after they were asked for verbal descriptions of similarities and differences between the sounds. The subjects then graded the different sounds on scales constructed from their own words. This was an example of a technique where the subjects came up with descriptions using their own vocabulary with known meaning to them, instead of being provided with the experimenter's descriptors for the scales. The data was subsequently analysed by methods used in the Repertory Grid Technique, aimed to find a pattern or a structure not necessarily known to the subjects (or the experimenters) themselves. The ex-

perimental idea was to investigate if a pattern with distinguishable groups of descriptors showed, and if so, it would be regarded as an indicator of the presence of the dimensions searched for. The results from this experiment have been reported in [1,9,10,11], and indicated the existence of a number of dimensions described by attributes generally used by the subjects for describing perceived dimensions of spatial audio. In [11] the correlation between some of the attributes was reported.

The problem with verbal descriptions of a perceived event is addressed by the authors in [1]. Many factors affect a person's conception and his/her description of the event, e g memory, other senses, emotion/sentiment, training, terminology, etc. Another important issue is the interpretation of verbal data and the bias involved with this [10]. With this in mind, in addition to the experiences and the results from the 1998 experiment, a number of questions arise concerning the feasibility of attributes in the form of verbal descriptors as means of assessing spatial quality:

- Are these attributes valid for describing the spatial quality of (a subset of) reproduced sounds?
- Are scales defined by words interpreted similarly within a group of subjects?
- If such scales are found to be valid, are certain attributes interrelated and if so, in which way?

In order to answer these questions, a new experiment was designed and conducted in early 2001. The new experiment made use of the attributes resulting from the 1998 experiment. Scales were constructed from the attributes and were provided to a new group of subjects. The subjects assessed a number of sound stimuli on the provided scales. The hypothesis to be tested in the experiment and its alternative were:

- If the scales are not relevant for describing parts of spatial quality of a subset of reproduced sounds, they will have insufficient common meaning to the subject group, which will not be able to make distinctions between any stimuli at a significant level, i e the data will contain mostly randomly distributed points.
- If, however, the scales are relevant in this respect, the scales will have sufficient common meaning to the group, which will be able to make distinctions between some or all of the stimuli in the experiment at a significant level.

If the alternative hypothesis is true, the interrelations of scales and attributes can be analysed subsequently.

The main purpose of the experiment is to investigate the feasibility of using verbal attributes as components in assessment of spatial quality of reproduced sound. It is not the authors' intention to focus on the source material used as stimuli, their physical differences or different recording/downmixing techniques. Differences in these are mainly used in this experiment as means of exciting different dimensions of the subjects' perception of sound, in order to span the spatial quality space. However, some observations of the interaction between attributes and certain stimuli have been made. The results from this experiment are reported in this paper.

ATTRIBUTES

The analyses of the 1998 experiment yielded a number of attributes used by the subjects. Some of the attributes occurred in more than one analysis. One problem encountered during the work was the fact that an attribute could refer to different parts of the auditory event, such as the sound source itself (a single instrument), a group of sources (a section of instruments), groups of sources (an orchestra) or the whole scene including the reflected sounds (from the hall). This encouraged the authors to assign certain attributes to defined parts of the auditory scene in order to avoid confusion about what the attributes were referring to. Since the task of the new experiment was to possibly verify the findings of the previous one, all these findings had to be compiled in a comprehensible form with ample size for the new group of listeners to consider. This implies a reduction of the number of elicited attributes from the 1998 experiment, with the intention to keep the main part of them for the new experiment. This compilation was made by the authors. The omitted attributes were "externalisation", "phase" and "technical device", since they were a result of the use of phase reversed signals in the 1998 experiment and that no phase reversed signals were considered for the new experiment. Externalisation (to perceive sound as coming from outside one's head in contrast to 'internalisation' where the sound is perceived as coming from within the head) was also considered as being a dichotomous attribute hard to grade on the linear scales that were going to be used in the experiment. The attributes included in the new experiment were divided into four attribute classes:

- General attributes – referring to the whole sound as an entity.
- Source attributes – referring to the sound of the sound source.
- Room attributes – referring to sound relating to the acoustical environment as a result from a sound source's initial action, e g reverberation.
- Other attributes – sounds generated neither by the source nor its interaction with the room.

The attribute classes contained attributes accompanied by a written description of each attribute. The attributes and their descriptions are found in Appendix A. The following 12 attributes (with their abbreviations and attribute classes) were used in the experiment

- *Naturalness* *nat* General
- *Presence* *psc* General
- *Preference* *prf* General
- *Envelopment* *env* General
- *Source width* *swd* Source
- *Localisation* *loc* Source
- *Source distance* *dis* Source
- *Room width* *rwd* Room
- *Room size* *rsz* Room
- *Room spectral bandwidth* *rsp* Room
- *Room sound level* *tlv* Room
- *Background noise level* *bgr* Other

Which attributes could be purely spatial or non-spatial or a mixture of both was not considered, since the origin of the attributes is the previous experiment, where subjects came up with these attributes as descriptors of their experiences. The attributes yielded from the analyses of the previous experiment and their relation to the derived attributes (abbreviated) used in the new experiment are shown in fig. 1.

EXPERIMENTAL DESIGN

The outline of the experiment was to provide a non-naïve group of subjects with a list of attributes with associated descriptions and, for every attribute, listen to a number of different sound stimuli and grade the stimuli on scales defined by the attributes. The subjects performed the experiment one at a time in a listening room equipped with loudspeakers and a user interface in the form of a computer screen, a keyboard and a mouse. All communication with the subjects was made in Swedish.

Stimuli

The main part of the experiment comprised grading of attributes relating to the source or the room. Therefore it was important to have stimuli in the form of recordings with a relatively low complexity, in this case meaning one single stationary centre-positioned source within a room or a hall. This was to avoid the difficulties with assessing too complex recordings, e g containing multiple sources, moving sources, changes of reverberation, etc. Consequently, the stimuli were all recordings of single sources in acoustical environments, i e no anaechoic conditions were used. To achieve ecological validity ('real-

ATTRIBUTES FROM PREVIOUS ANALYSES	REF	ATTRIBUTES FOR NEW EXPERIMENT											
		nat	*psc*	*prf*	*env*	*swd*	*loc*	*dis*	*rwd*	*rsz*	*rsl*	*rsp*	*bgr*
		G	G	G	G	S	S	S	R	R	R	R	O
authenticity/naturalness	[9]	x											
lateral positioning/source size	[9]					x	x						
envelopment	[9]				x								
depth	[9]							x					
room/reverberation properties: spectral, level and clarity	[9]									x	x	x	
source width	[9]					x							
(externalisation)	[9]												
frontal image	[9]						x						
localisation, left – right and front – back	[10]						x						
depth/distance	[10]							x					
envelopment	[10]				x								
width	[10]					x			x				
room perception	[10]									x	x	x	
(externalisation)	[10]												
(phase)	[10]												
source width	[10]					x							
source depth	[10]							x					
detection of background noise	[10]												x
frequency spectrum	[10]											x	
naturalness	[11]	x											
presence	[11]		x										
(technical device)	[11]												
positive/negative	[11]			x									

Fig 1. Table showing attributes yielded from previous analyses of the 1998 experiment and the derived attributes for the new experiment. Omitted attributes within brackets. The attribute class is indicated by letters under the attribute abbreviation.

world validity'), sound sources with high probability to encounter in a natural listening situation were used: trumpet, saxophone, speech and flute. The flute and the trumpet recordings can be characterised as more reverberant ('hall') than the speech and the saxophone ('room'). These differences are aimed to span the scaling of the room parameters.

In the 1998 experiment it was noted that different modes of reproduction excited a number of spatial sensations. Since one aim of the new experiment was to verify the findings in the previous one, this method was employed once again. The four original recordings were available in a 5-channel format, intended for reproduction on five different speakers. The 5-channel recordings were all downmixed for reproduction in 2-channel stereo and 2-channel phantom mono. In addition to that, the speech recording only was mixed down to mono. The downmix coefficients for 2-channel stereo were

$$L = L_F + 0.71C_C + L_R$$
$$R = R_F + 0.71C_C + R_R$$
$$C = L_S = R_S = 0$$

where

L, R, C, L_S and R_S are the output signals to the 5-speaker system

and where

L_F, R_F, C_C, L_R and R_R are the original signals from the 5-channel recordings.

For the phantom mono downmixes the coefficients were

$$L = R = L_F + R_F + 1.4C_C + L_R + R_R$$
$$C = L_S = R_S = 0$$

and finally, the coefficients for the mono stimulus

$$C = L_F + R_F + 1.4C_C + L_R + R_R$$
$$L = R = L_S = R_S = 0$$

Every sound stimulus was assigned a number for identification throughout the test, "item 1…13". The complete list of sound stimuli with their identifying number and the mixdown mode used is shown in fig. 2.

The flute and the trumpet recordings were collected from the part "Musik in Surround", track id 6 and 12 on the "Multichannel Universe" DVD [12], and the saxophone and the speech recording were made at the School of Music in Piteå, Sweden. Both the saxophone and the speech recordings were used in the 1998 experiment and details of the recording technique were given in [1]. All recordings were downmixed on a ProTools system and stored as *.wav files with a resolution of 16 bits and 48 kHz sampling frequency. The level difference between the different 5-channel recordings were adjusted in the listening room "by ear" by two persons until consensus was reached about the plausible sound level for the different sources. The downmixed versions of the original 5-channel recording were level aligned – also in the listening room – against the original by measuring the equivalent continuous sound level, $L_{eq}(A)$, for the first 10 seconds of the sound files and adjusting the difference to be within ±1 dB.

Item	Programme	Downmix mode	$L_{eq}(A)$
1	Flute	5	65.3
2	Flute	2	65.3
3	Flute	p	65.6
4	Saxophone	5	76.9
5	Saxophone	2	77.5
6	Saxophone	p	77.1
7	Speech	5	71.6
8	Speech	2	72.0
9	Speech	p	71.9
10	Speech	m	71.5
11	Trumpet	5	78.0
12	Trumpet	2	78.6
13	Trumpet	p	78.3

5 = 5-channel stereo
2 = 2-channel stereo
p = phantom mono
m = mono

Fig. 2. Stimuli used in the experiment, their identification and measured sound level in the listening room

Subjects

The number of subjects completing a whole test session was 19. All subjects were students from the sound recording programme at the School of Music. They had previously participated in listening tests aimed to assess the total audio quality of coding algorithms in bit-reduction systems. They had neither received any special training in assessing spatial quality or any instructions in using common language for describing the spatial features of recordings. In conclusion, the subjects should be regarded as experienced listeners of reproduced sound, without any particular bias for a certain way of describing spatial quality.

Listening conditions

The experiment was executed in a reproduction room at the School of Music. The dimensions of the room was 6 × 6.6 × 3.2 m (w × d × h). All reproduction was made through Genelec 1030A loudspeakers, configured according to BS-1116 [13] at a 2 m distance from the listening position, fig. 3. The settings of each loudspeaker were: Sensitivity = +6 dB, Treble tilt = +2 dB, Bass tilt = -2 dB. Only one subject at a time was present in the listening room during the experiment. Equipment with fans was acoustically insulated to avoid noise in the listening room. The room had no windows and the only light in the room were a small spotlight at the listening position for reading. This was to increase the subject's concentration on the user interface and minimise visual distraction from the room.

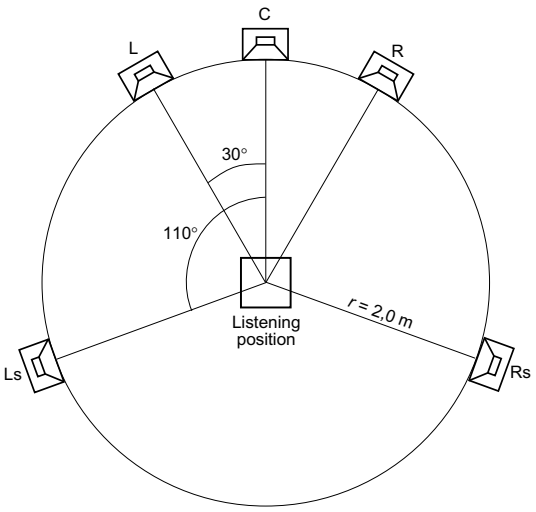


Fig. 3. Loudspeaker set-up

Experiment equipment

The experiment was performed on a computer (PC) by which each test session was controlled. All sound files were stored on the computer's disk and played back via an 8-channel sound card installed in the computer. (Only five channels were used.) The sound card output delivered audio data in the T-DIF format. This was converted by a DA-88 tape recorder into five discrete analogue signals directly feeding the speakers. The equipment had been tested for phase differences between the channels and as well as for general audio quality by Sveriges Radio (the Swedish Broadcasting Corporation).

For controlling the test, special software was designed. Both playback controls as well as collecting subject responses were handled by the software.

Experiment execution

Prior to the test, every subject received a written instruction, where the experiment was described. A list of the attributes (Appendix A), to be used in the test accompanied the written instruction. The subjects were allowed to ask questions about the instruction, but not about the attributes and their descriptions. The instruction and the attribute list were available for the subjects during the whole session.

A session started with a training phase with a content corresponding only to 25% of the actual test to avoid subject fatigue at the end of the test. The purpose of the training phase was to familiarise the subjects with the equipment and the stimuli used in the test.

Each subject was first presented one attribute with its description. All 13 stimuli were available for listening by clicking on buttons on the computer screen. The task was to grade all stimuli one by one on the attribute presented. This was accomplished by providing continuos sliders on the screen, one slider per stimulus. The subjects were instructed to regard the scale on the sliders as linear. The slider had only two markings, one at each endpoint, the lower marked "0" (zero) and the upper marked "MAX". The subject was also instructed to use the MAX grade for at least one stimulus, but did not necessarily have to give any stimulus the grade 0. When the subject was satisfied with his/her grading on the first attribute, the scores were stored by clicking a button, whereupon the next attribute was presented. All stimuli were graded again, but now on the new attribute. This was repeated until all attributes were graded. When this was completed, the test finished.

To avoid systematic errors, the presentation order and assignation of playback buttons were randomised: When a session starts the attribute class is chosen randomly. The order of which the attributes within the chosen class are presented is also picked randomly. When all attributes within the class are assessed by the subject, a new attribute class out of the remaining ones is randomly chosen. This is repeated until all attribute classes with their attributes are assessed. For every new attribute, the assignation of the stimuli to the 13 playback buttons is re-randomised.

Data acquisition

The slider position representing a subject's assessment of a given stimulus on a given attribute was converted into integer numbers from 0 to 100, where 0 corresponds to the marking "0" and 100 to "MAX". The converted grades with proper identification of subject, associated stimulus, attribute and date/time were stored on the computer in one text file per subject. The text files were later converted into MS Excel files for the upcoming analysis.

RESULTS – ANALYSIS OF ATTRIBUTE SIGNIFICANCE

The analysis seeks to answer the introductory question in the paper: Is the subject group able to make significant distinctions between the stimuli in the test using the provided attributes?

The analysis started with normalisation of the data and check for normal distribution. Analysis of variance (Anova) was used for the significance test.

Data structure

The data acquired consisted of 19 subjects assessing 13 stimuli on 12 attributes. This yields 2964 data points. Every subject delivered 156 grades.

Normalising

In order to facilitate the comparison of grades between items, the subjects' different use of the scales provided must be equalised. This was accomplished by, for each subject, normalising the grades given to an attribute. This way, each attribute had the same mean value and standard deviation as the other attributes. There are 13 stimuli per attribute and the mean value

$$\bar{x}_{ik} = \frac{1}{13}\sum_{j=1}^{13} x_{ijk}$$

and the standard deviation

$$s_{ik} = \sqrt{\frac{1}{12}\sum_{j=1}^{13}(x_{ijk} - \bar{x}_{ik})^2}$$

where

x_{ijk} = grade given on attribute i for item j by subject k

are used for calculating the z-score

$$z_{ijk} = \frac{x_{ijk} - \bar{x}_{ik}}{s_{ik}}$$

which now is the normalised value of the original grade. The mean value of z-scores per subject and per attribute is 0 and the standard deviation is 1. Consequently, the data now consists of normalised values in the form of z-scores suitable for the coming steps in the analysis.

Normal distribution test

To examine if the subjects' scores given for each item on each attribute were normally distributed, the Shapiro-Wilks' test was performed. Since the subjects graded 13 stimuli on 12 attributes, the number of cases to be tested was 156. The outcome of this test, expressed as probabilities for normal distribution for the different cases in the experiment, is found in Appendix B. When the level of confidence is set to 95%, the test shows that a normal distribution can not be excluded in 129 of the 156 cases. Since more than 80% of the cases seem to have a normal distribution, this is a first indication of some agreement between the subjects in their grading of the stimuli. The existence of normal distribution also does not exclude commonly used statistical methods.

Attribute applicability test

Analysis of variance (Anova) was performed on each attribute to determine whether it was sufficient or not for discriminating between stimuli. The dependent variable was the normalised grade (*z-score*) and the factor was stimulus (*item*). Since the data was normalised, the *F*-ratio of the factor subject (*subno*) became zero, which confirmed that the subject effect had been removed from the analysis, as intended. The null hypothesis

H_0 : The attribute provided is not sufficient for enable the subjects to find a significant difference between any stimuli

and its alternative hypothesis

H_A : The attribute provided is sufficient for enable the subjects to find a significant difference between at least one stimulus and the other stimuli

The analysis showed that all the 12 attributes had *F*-ratios with significance levels $p < 0.05$. The null hypothesis was therefore rejected in favour of its alternative for every attribute. This means that all attributes in the experiment showed to be sufficient for making distinctions between at least one stimulus and the other stimuli. The attributes must therefore have some common meaning to the subjects; otherwise, the individual subject differences would have been randomly distributed across the stimuli, resulting in insignificant differences between the stimuli. The Anova tables are found in Appendix B.

Attribute consistency test

Since all attributes were found to be sufficient for discriminating between stimuli, it is of interest to examine the attributes for how consistently they were graded. A relatively high consistency is likely to indicate a more similar perception of the attribute than a relatively low one. To test this, the residual (or error) variances for the attributes were taken from the previous Anovas and compared between them. Since the between-subject variability was removed earlier from the Anova model by the normalisation procedure, the residual variance only consists of the differences in magnitude and direction of the trends in subject performance. Consequently, a low residual variance indicates a high consistency in trends [14].

When the attributes' residual variances were ordered in ascending order and these variances were inspected, they formed two groups, one with relatively high variances and another with relatively low ones. One attribute stood out as being in between the two groups. The group with low residual variances comprises the attributes (in ascending order of variances*): envelopment, room size, background noise level, source distance, preference* and *room width*. The group with high residual variances includes the attributes in (descending order of variances): *room spectral bandwidth, localisation, source width, naturalness,* and *room sound level*. Between these groups, the attribute *presence* is found. The residual variances and *F*-ratios are shown in fig 4.

Attribute	F-ratio	Residual
env	39.54	0.33019
rsz	35.10	0.35783
bgr	33.69	0.36761
dis	32.84	0.37371
prf	28.96	0.40464
rwd	28.77	0.40628
psc	19.90	0.50138
rlv	13.83	0.59701
nat	13.79	0.59776
swd	13.58	0.60161
loc	12.53	0.62226
rsp	10.17	0.67441

Fig. 4. From the Anova: F-ratios and residual variances of the attributes sorted in ascending order of their residual variances.

Mean scores per item and attribute

Analysis of the stimuli themselves was not the main purpose of the experiment, but since the attributes were found to enable subjects to discriminate between some stimuli, these were examined to find which of the stimuli was separable from the others by the different attributes. This is shown in fig. 5a-b, where, for every attribute, every stimulus' mean score with its associated confidence interval (95%) is plotted. Some observations are commented on here.

The 5-channel and 2-channel stimuli showed higher scores (not necessarily equal to 'better') on most of the attributes. For some attributes there were no significant difference between the reproduction modes of a source, e g *room level* and *background noise level*. If the mono speech stimulus was disregarded this applied for *room size* too.

The 5-channel and 2-channel recordings were equally scored per source on most of the attributes. The exceptions were on *presence* and *envelopment*, where the 5-channel version of both the saxophone source and the speech source was higher scored than the 2-channel version. For *preference* there was also a difference between 5 and 2 channels for speech.

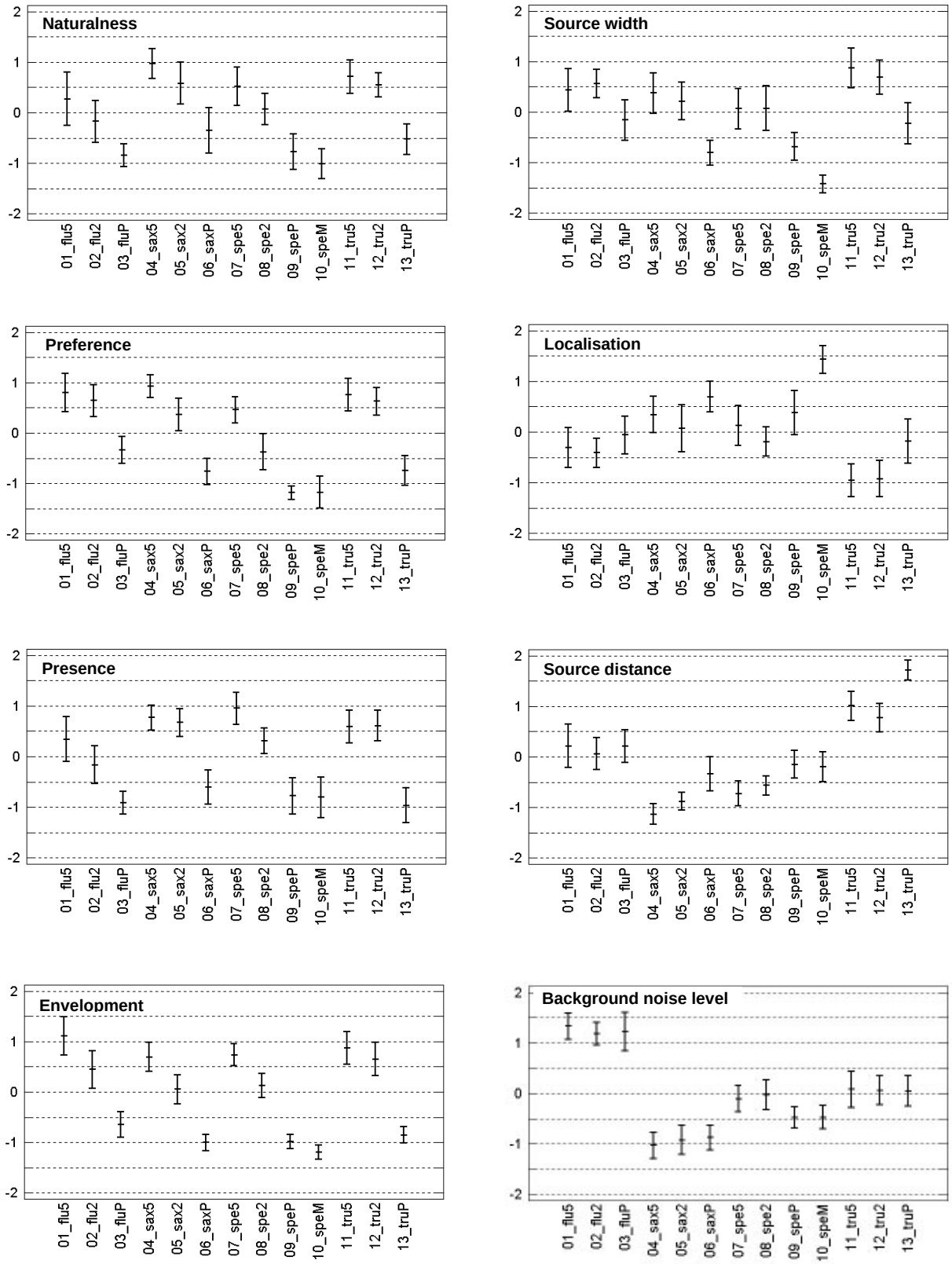


Fig. 5a. Mean scores on attributes for all items. (8 diagrams)

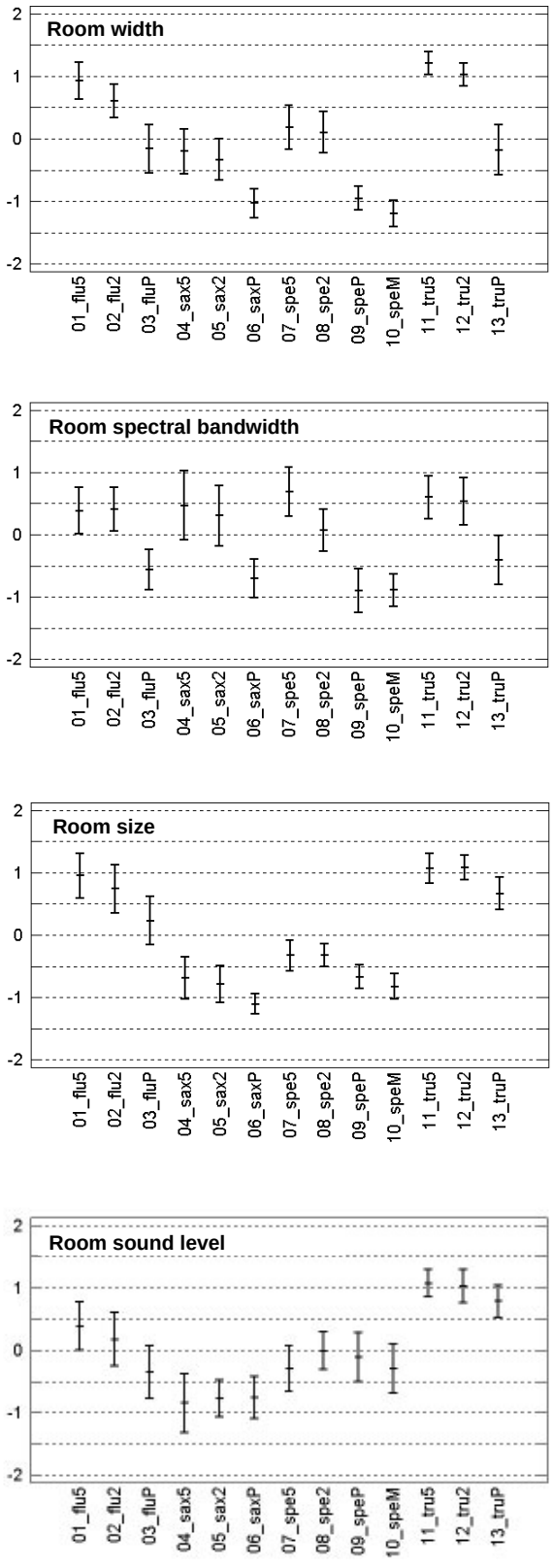


Fig. 5b. Mean scores on attributes for all items (4 diagrams)

THE DIMENSIONALITY OF ATTRIBUTES

In order to investigate if some of the attributes used in the experiment were perceived and thereby likely to be scored similarly by the subjects, multivariate analysis of the data were conducted. The methods used were:

- Principal component analysis (PCA) for reducing the data into fewer components still accounting for a main part of the variance of the scores. In PCA all the variance (both common and unique) of the scores is analysed.
- Principal Factor analysis (FA) with Varimax rotation of the factors for explaining the reduced dimensionality in terms of the attributes. In FA the common variance for the variables is used in the analysis.
- Cluster analysis for visualising relationships between the attributes.
- Correlation analysis for calculating the interrelation between the attributes.

Principal component analysis

A principal component analysis [15] was performed on the set of attributes, which corresponds to the columns in the matrix

$$\mathbf{A} = \begin{bmatrix} z_{1,1,1} & \cdots & z_{12,1,1} \\ \vdots & & \vdots \\ z_{1jk} & \cdots & z_{12,jk} \\ \vdots & & \vdots \\ z_{1,13,19} & \cdots & z_{12,13,19} \end{bmatrix}$$

where

z_{ijk} = z-score on attribute i for item j by subject k

and the attributes were normalised before the PCA. The number of components to keep in the analysis can be determined by the following methods [16]:

- Kasier's criterion – all components with an eigenvalue $\lambda > 1$ should be kept in the analysis. The eigenvalues are shown in fig. 6.
- Cattell's scree test – a method where a plot of eigenvalues versus the number of component is inspected. The scree plot is shown in fig. 7.
- Variance dependent – components are brought in to the analysis until they reach a certain level of cumulative variance [15], usually 70 or 80 percent. The variances are shown in fig. 6.

Component Number	Eigenvalue	Percent of Variance	Cumulative Percentage
1	4,83017	40,251	40,251
2	2,59742	21,645	61,897
3	0,84147	7,012	68,909
4	0,78448	6,537	75,446
5	0,68518	5,710	81,156
6	0,50635	4,220	85,376
7	0,36699	3,058	88,434
8	0,33265	2,772	91,206
9	0,28791	2,399	93,605
10	0,27609	2,301	95,906
11	0,26218	2,185	98,091
12	0,22910	1,909	100,000

Fig 6. Extraction statistics from the PCA

The scree plot (fig. 7) shows that three components should be considered in the subsequent analysis. Their cumulative variance was 68,9%.

An inspection of the component weights (fig. 8) or loadings showed that the component accounting for the highest variance, component 1, was positively loaded by all attributes except for *localisation*. *Room width*, *envelopment* and *source width* together with *preference* loaded component 1 mostly. Component 2 showed the most positive loading for *distance* and the most negative for *presence*. At the positive end of component 3, *source width* were found, and at its negative, *background noise* and *localisation*. Notable was that *source width* and *localisation* seemed to be opposites in any combination of the dimensions. Plots of components weights are found in Appendix C.

The extraction of components in a PCA considers all variance, so the components are likely to consist of more complex functions of the variables (than a FA), which could make the components harder to interpret [17]. Nevertheless, a pattern can be discovered; Component 1 could be interpreted as a width in a general meaning, since the attributes loading the component describes a perceived width and a feeling of being surrounded by sound. A wider(!) interpretation is an experience of a "bigger event" or perhaps a larger listening area. Component 1 also represents a positive attitude towards the sound. Component 2 seems to account for distance to the sound since the opposites are *source distance – presence* and the intermediate attributes follows a 'logical' order from distance through room attributes via width to *envelopment*, stopping at *presence*. The loadings on component 3 indicated the opposite between *source width* and *localisation.* The subjects seemed to interpret the attributes forming this dimension as if the source gets wider it becomes harder to localise, which most likely is an image focus perception.

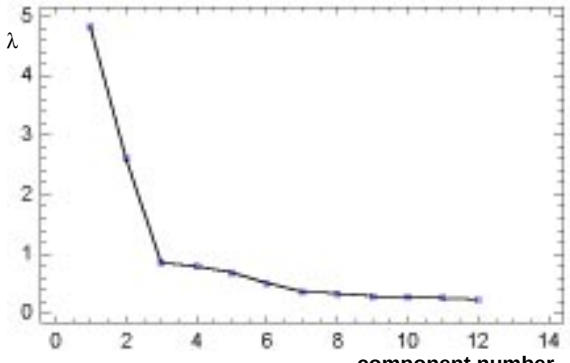


Fig. 7. Scree plot for the PCA

Attribute	Component 1	Component 2	Component 3
nat	0.271	-0.345	-0.236
psc	0.261	-0.402	-0.151
prf	0.338	-0.233	-0.108
env	0.370	-0.160	-0.185
bgr	0.176	0.313	-0.504
rwd	0.390	0.105	0.013
rsz	0.316	0.304	-0.267
rsp	0.269	-0.179	0.234
rlv	0.205	0.384	0.042
swd	0.326	0.039	0.556
loc	-0.305	-0.172	-0.425
dis	0.103	0.478	-0.059

Fig 8. PCA: Component weights on attributes

Factor Analysis

A principal factor analysis was performed on the score matrix **A**, where the scores were normalised prior to the FA in the same way as in the PCA. Factor analysis is used when an accurate description of the domain covered by the variables is desired [17]. The number of factors was determined by the same criterion as in the PCA and was accordingly set to three. To increase the interpretability, the factors were rotated, using Varimax, to maximise the loadings of some of the attributes. These attributes can then be used to identify the meaning of the factors [16]. The factor loadings are presented in fig. 9.

After the Varimax rotation factor 1 was loaded positively by all of the attributes in the General attributes class (referring to the sound as an entity): *naturalness*, *presence*, *preference* and *envelopment*. Factor 2 showed a high loading by *room size*, *back-*

ground noise level, *source distance*, *room sound level* and intermediate loading by *room width*. For factor 3, *source width* loaded strongly positive and *localisation* almost as much but with a negative sign. Plots of factor loadings are found in Appendix C.

Attribute	Factor 1	Factor 2	Factor 3
nat	0.841	-0.036	0.057
psc	0.859	-0.152	0.094
prf	0.787	0.113	0.266
env	0.785	0.266	0.267
bgr	0.099	0.776	-0.079
rwd	0.480	0.520	0.511
rsz	0.247	0.806	0.266
rsp	0.522	-0.041	0.452
rlv	-0.098	0.655	0.385
swd	0.284	0.141	0.822
loc	-0.156	-0.336	-0.735
dis	-0.319	0.707	0.212

Fig 9. Factor analysis: Factor loadings on attributes after Varimax rotation

From the factor analysis it was noted that factor 1 corresponded well to all of the attributes in the General attribute class, whereas factor 2, with one exception (*source distance*), seemed to describe the room attributes. Factor 3 accounted for the source attributes; especially those describing something that can be interpreted as image focus. An alternative interpretation is that factor 1 is a more 'attitudinal' factor where a positive loading indicates both an appreciation of the sound and that enveloping sounds are an important part of a natural and preferable experience. As an alternative approach for factor 2, it could be interpreted as a distance factor, since an increased distance from the source in a room would change the balance between the direct and the reflected sound and thus increase the audibility of the room.

Cluster analysis

Cluster analysis [18] compares and group similar variables according to the metrics and agglomeration technique used. A useful output from a cluster analysis is the dendrogram, which displays the similarity in the form of a tree where the similarity between variables are indicated by 'branches' joining the variables at the point of similarity. The metrics used was squared Euclidean and the agglomeration technique was Complete linkage. The resulting dendrogram is shown in fig 10, The agglomeration distance plot (not shown here), also used in [10] for

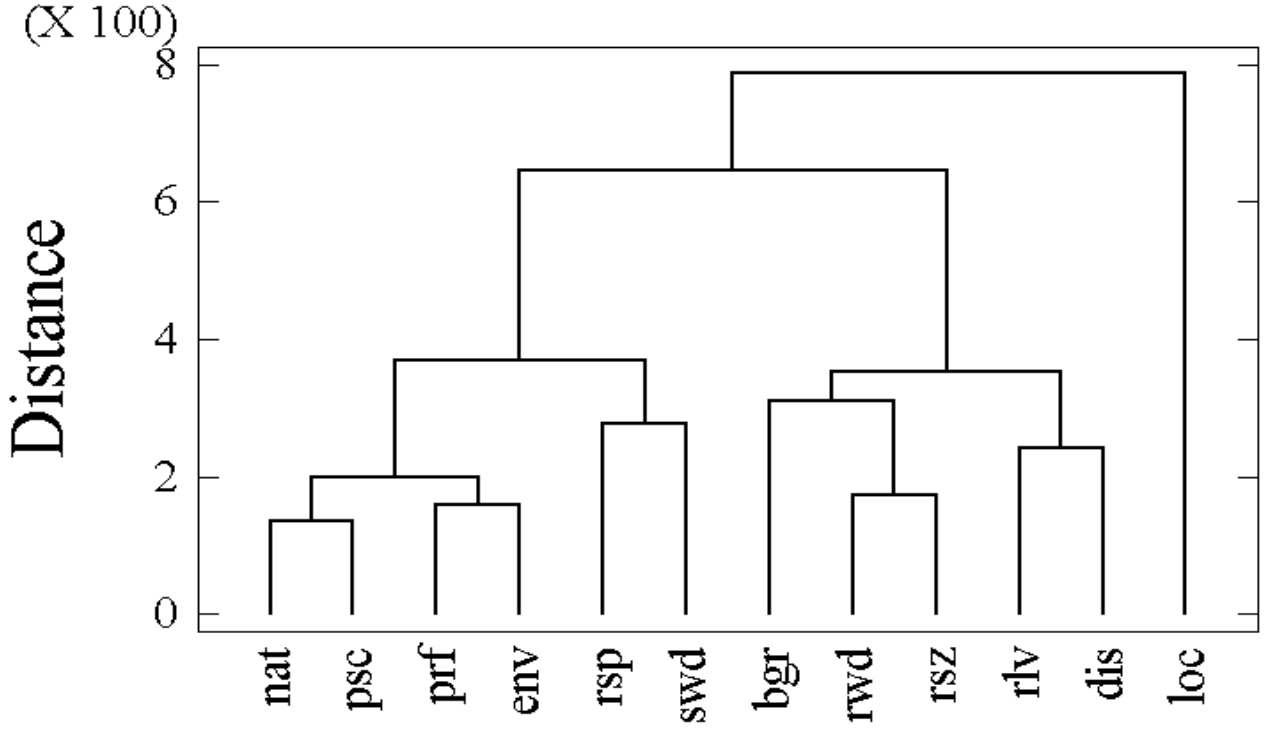


Fig. 10. Cluster analysis presented in dendrogram form. The closer to the baseline the attributes are joined, the more similar they are.

determining the appropriate number of clusters (groups) indicated three groups of attributes. These groups with their attributes were:

Group 1
Naturalness
Presence
Preference
Envelopment
Room spectral bandwidth
Source width

Group 2
Background noise level
Room width
Room size
Room level
Source distance

Group 3
Localisation

An inspection of the dendrogram reveals subgroups within the three groups resulting from the agglomeration distance plot; *naturalness, presence, preference* and *envelopment* show a strong relationship as well as *room width* and *room size*.

Correlation analysis

Finally, to get a complete picture of the correlation between the attributes, the Pearson product moment correlation coefficient, r [19] was calculated. The results are given as a coefficient for every combination of the attributes. The correlation coefficients are found in Appendix B. The strongest correlations ($|r| > 0.6$) are found between

naturalness – presence	$r = 0.727$
preference – envelopment	$r = 0.674$
room width – room size	$r = 0.646$
room width – envelopment	$r = 0.613$
presence – envelopment	$r = 0.611$
source width – localisation (negative)	$r = -0.602$

The lowest correlations are between

naturalness – room sound level	$r = -0.022$
envelopment – source distance	$r = 0.03$
background noise – room spec. bandw.	$r = 0.059$
preference – room sound level	$r = 0.061$
presence – background noise	$r = -0.068$

This analysis also verifies the relatively strong interrelation between *envelopment* and the attributes expressing naturalness and a feeling of presence.

DISCUSSION AND CONCLUSIONS

The experiment's aim was to, if possible, validate the findings in a previous experiment as well as to understand more of how the attributes encountered were working as descriptors for assessing the spatial quality of reproduced audio. In order to draw conclusions from any experiment, its limitations must as far as possible be known to those involved. The experiment at hand made use of recordings of single sound sources in an acoustical environment. There is no imagination needed to realise that the selection is narrow, compared to all recordings available. This of course reduces the generalisability of the results to other types of sound. The decision to use such recordings was a result of some difficulties encountered in the previous experiment concerning complex sound sources. There is a high factor of 'reality' in using complex sounds, but from the experimenter's point of view there is a problem in having too many uncontrolled conditions and a risk to confuse the participating subjects. On the other hand, it is possible to use very simple stimuli but with little connection to what most people normally listens to via loudspeakers, which is sometimes encountered in classical psychophysics. This experiment tries to satisfy both a relatively low complexity as well as sound stimuli that could be found outside the laboratory.

The null hypothesis in the Anova was that if the attributes provided did not make sense or did not have any common meaning to the subjects, the result on the different attributes would have consisted of randomly distributed scores yielding noise only in the data. Such a result would have invalidated the alternative hypothesis, partly (for some attributes) or in total (all attributes). As stated in the analysis section, every attribute was able to produce significant differences on the stimuli selected. This means that every attribute has some common meaning to the group of subjects under the experiment conditions.

Although all attributes showed significant F-ratios, it became clear that some attributes are more consistently scored than others. If this is due to differences in perception or variability in the interpretation of the written descriptions of the attributes is impossible to tell. Most consistently graded were *envelopment, room size, background noise, source distance, preference,* and *room width.* Least consis-

tently graded were: *room spectral bandwidth*, which both had a short description that may have been insufficient for the subjects, in combination with the possible difficult task to quantify the bandwidth of the room sound. The scores on *room spectral bandwidth* separate the stimuli in basically two halves; mono and 2/5-channel reproduction, which either is a perceived difference in bandwidth or an interpretation of some other width attribute. Other attributes with low consistency were *localisation, source width, naturalness* and *room level*. Martin et al [20] points at the problem with defining source width, where there is a risk of confusion between narrower image width, increased distance to the source and less spread of low frequency content. The results from this experiment seem to confirm their findings.

Some attributes seem to be more independent from the number of channels used. The different reproduction modes of a source were not able to cause any significant differences in mean score of the source within the attributes *room level* and *background noise.* When the mono stimulus is disregarded, this also applies to *room size*. This suggests that the properties described with these attributes are not easily corrupted by alterations by means of down-mixing.

Looking at the large trends in the form of dimensionality, the three dimensions extracted from the PCA and the FA mainly leave us with factors representing three attribute classes. The first factor seems to contain attributes concerning general width aspects, naturalness and preference. The second contains an element of decreasing distance from *source distance* and *room level* to *presence*. The third factor comprises source image focus. This would indicate that the subjects actually perceive the attributes classes mainly as orthogonal on the dimensions General, Source and Room. The remaining class, Other, loads the Room factor. The division of attributes into groups were known to the subjects prior to the experiment so this may have biased them. A looser interpretation is that factor 1 represents an attitudinal dimension containing a wide and enveloping sound. This may not be the case, neither when listeners get more experienced and able to be more precise in their discrimination, nor when the differences in terms of spatial quality between stimuli decreases.

The cluster analysis and the correlation analysis both confirm the trends above in greater detail. The general attributes *naturalness* and *presence* show the highest correlation, followed by *preference* and *envelopment*.

In conclusion, the experiment shows that verbal attributes still are a valid approach for describing spatial quality. The findings validate the experiences from the 1998 experiment, showing that these attributes have a common meaning under the conditions of the experiment. Some attributes are more consistently used by the subjects, other attributes are less sensitive to changes in reproduction modes. The use of attribute classes in order to focus a subject onto different parts of the auditory event seems to be a successful approach.

Future work

The attributes used emerged from a technique without provided constructs to encourage the subjects in the previous experiment to find and use words familiar to them. Via PCA and cluster analysis a common meaning of these constructs was sought, and attributes appropriate to describe them were formulated. One idea for future work is to re-iterate that process, since we now know more about stimuli and perceivable dimensions, with the aim of finding more and new attributes. Another approach is that, based on the experiences we have, try to refine the attribute descriptions and repeat an experiment to see if the scoring consistency improves

To carefully examine the use of attributes, the difference between stimuli can be decreased and more precisely controlled. This will make it possible to observe whether the scales depending on certain attributes are still valid under new conditions. These differences could be created in the recording domain, e g by means of different microphone techniques, without changing the modes of reproduction.

In an experiment, the subjects could also be provided with reference stimuli to furthermore narrow down the meaning of the attributes

ACKNOWLEDGEMENTS

The authors wish to thank the students at the School of Music, Piteå, Sweden for their participation in this experiment. The test equipment was constructed as a joint project between the School of Music and Sveriges Radio (Swedish Broadcasting Corporation) in which Ola Kejving and Lars Mossberg are thanked for their support. A special thanks to Jonas Ekeroot, Sveriges Radio, for the construction of the test equipment as well as for the software programming which made this experiment possible.

REFERENCES

1 Berg, J. and Rumsey, F. (1999) Spatial attribute identification and scaling by Repertory Grid Technique and other methods. In *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction, 10–12 Apr*. pp 51-66. Audio Engineering Society.

2 Rumsey, F. (1998) Subjective assessment of the spatial attributes of reproduced sound. In *Proceedings of the AES 15th International Conference on Audio, Acoustics and Small Spaces*, 31 Oct–2 Nov, pp. 122–135. Audio Engineering Society

3 Kelly, G. (1955) *The Psychology of Personal Constructs*. Norton, New York.

4 Danielsson, M. (1991) *Repertory Grid Technique*. Research report. Luleå University of Technology. 1991:23

5 Fransella, F. and Bannister, D. (1977) *A manual for Repertory Grid Technique*. Academic Press, London.

6 Stone, H. *et al* (1974) Sensory evaluation by quantitative descriptive analysis, *Food Technology*, November, pp 24-34.

7 Mason, R., Ford, N., Rumsey, F. and de Bruyn, B. (2000) Verbal and non-verbal elicitation techniques in the subjective assessment of spatial Sound Reproduction. Presented at *AES 109th Convention, Los Angeles*. Preprint 5225.

8 Wenzel, E. M. (1999) Effect of increasing system latency on localization of virtual sounds. In *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction, 10–12 Apr*. Audio Engineering Society. pp 42-50.

9 Berg, J. and Rumsey, F. (1999) Identification of perceived spatial attributes of recordings by repertory grid technique and other methods. Presented at *AES 106th Convention, Munich*. Preprint 4924.

10 Berg, J. and Rumsey, F. (2000) In search of the spatial dimensions of reproduced sound: Verbal Protocol Analysis and Cluster Analysis of scaled verbal descriptors. Presented at *AES 108th Convention, Paris*. Preprint 5139.

11 Berg, J. and Rumsey, F. (2000) Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction. Presented at *AES 109th Convention, Los Angeles*. Preprint 5206.

12 Surround Sound Forum, Balance and Media City (1998) *Multichannel Universe*. DVD BAL-95000-3. Balance, Munich.

13 ITU-R (1996) *Recommendation BS.-1116, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*. International Telecommunication Union.

14 Roberts, M. J. and Russo, R. (1999) A student's guide to analysis of variance. Routledge, London.

15 Johnson, R. A. and Wichern, D. W., (1998) Applied multivariate statistical analysis. Prentice-Hall, New Jersey.

16 Bryman, A. and Cramer, D. (1994) *Quantitative data analysis for social scientists*. Routledge, London.

17 Cureton, E. E. and D'Agostino, R. B. (1983) Factor Analysis – an applied approach. Lawrence Erlbaum, New Jersey.

18 Everitt, B. S. and Dunn, G. (1991) *Applied Multivariate Data Analysis*. Edward Arnold, London

19 Devore, J. L. and Peck, R. (1986*) Statistics, the exploration and analysis of data*. West Publishing Company, S:t Paul.

20 Martin, G., Woszczyk, W., Corey, J. and Quesnel, R. (1999) Controlling phantom image focus in a multichannel reproduction system. Presented at *AES 107th Convention, New York*. Preprint 4996.

APPENDIX A

ATTRIBUTES TO ASSESS IN LISTENING TEST MARCH 2001

GENERAL

Naturalness GA1	How similar to a natural (i.e. not reproduced through e g loudspeakers) listening experience the sound as a whole sounds. Unnatural = low value. Natural = high value.
Presence GA2	The experience of being in the same acoustical environment as the sound source, e g to be in the same room. Strong experience of presence = high value.
Preference GA3	If the sound as a whole pleases you. If you think the sound as a whole sounds good. Try to disregard the *content* of the programme, i e do not assess genre of music or content of speech. Prefer the sound = high value.
Envelopment GA4	The extent of how the sound as a whole envelops/surrounds/exists around you. The feeling of being in the centre of the sound. Feel enveloped = high value.

SOUND SOURCE

Source width SA1	The perceived width of the source. The angle occupied by the source. Does not necessarily indicate the known size of the source, e g one knows the size of a piano in reality, but the task to assess is how wide the sound from the piano is perceived. Disregard sounds coming from the sound source's environment, e g reverberation – only assess the width of the sound source. Narrow sound source = low value. Wide sound source = high value.
Localisation SA2	How easy it is to perceive a distinct location of the source – how easy it is to pinpoint the direction of the sound source. Its opposite (a low value) is when the source's position is hard to determine – a blurred position. Easy to determine the direction = high value.
Source distance SA3	The perceived distance from the listener to the sound source. Short distance/close = low value. Long distance = high value.

ROOM

Room width RA1	The width/angle occupied by the sounds coming from the sound source's reflections in the room – not the sound source itself. Narrow room = low value. Wide room = high value.
Room size RA2	In cases where you perceive a room/hall, this denotes the relative size of that room. Large room = high value. If no room/hall is perceived, this should be assessed as zero.
Room spectral bandwidth RA3	The perceived bandwidth of the room. Room with large bandwidth = high value.
Room sound level RA4	The level of sounds generated in the room as a result of the sound source, e g reverberation – i e not extraneous disturbing sounds. Weak room sounds = low value. Loud room sounds = high value.

OTHER

Background sound level OA1	The level of sounds not generated by the sound source itself. Weak background noises = low value. Loud background noises = high value.

APPENDIX B

Tables

- Test for normal distribution
- Attribute correlation
- Analysis of variance

item	ntl	psc	prf	env	bgr	rwd	rsz	rsp	rlv	swd	loc	dis
1	0,3184	0,2366	0,2066	0,1137	0,1257	0,2251	0,0019	0,4722	0,2308	0,6900	0,5824	0,0120
2	0,2511	0,7942	0,1344	0,2476	0,1065	0,1159	0,0012	0,7649	0,2965	0,9456	0,3314	0,0068
3	0,7363	0,3397	0,4260	0,0317	0,0025	0,1266	0,0806	0,3680	0,3661	0,4145	0,4519	0,8567
4	0,2448	0,4480	0,6034	0,6680	0,5505	0,2091	0,9945	0,3486	0,3945	0,7444	0,0821	0,5020
5	0,9873	0,7216	0,9604	0,5530	0,0856	0,7024	0,1265	0,4944	0,1836	0,7645	0,9079	0,0278
6	0,0874	0,0021	0,2199	0,4455	0,9296	0,0747	0,8143	0,1827	0,2166	0,0106	0,0041	0,5780
7	0,0183	0,8332	0,3395	0,0876	0,0891	0,7273	0,7654	0,8216	0,8291	0,1899	0,0643	0,4131
8	0,1849	0,0074	0,8695	0,8210	0,6496	0,0519	0,9885	0,7790	0,9635	0,1539	0,6262	0,6831
9	0,0316	0,0308	0,5134	0,3739	0,1258	0,8123	0,0834	0,2877	0,3490	0,8342	0,0006	0,0181
10	0,1596	0,0066	0,0081	0,2906	0,3148	0,0013	0,0193	0,2907	0,5073	0,2283	0,6420	0,2560
11	0,0107	0,0341	0,3358	0,0115	0,2598	0,2001	0,6974	0,4186	0,1521	0,6506	0,0035	0,0801
12	0,3118	0,4028	0,9668	0,0672	0,4059	0,6790	0,1506	0,7505	0,4587	0,5708	0,0259	0,0019
13	0,2540	0,0943	0,3226	0,2275	0,2583	0,0970	0,4732	0,3269	0,4855	0,0392	0,0830	0,5613

Shapiro-Wilks' test for normal distribution; p-values are shown.

	nat	psc	prf	env	bgr	rwd	rsz	rsp	rlv	swd	loc	dis
nat		0,727	0,593	0,595	-0,091	0,386	0,204	0,387	-0,022	0,252	-0,203	-0,141
psc	0,727		0,595	0,611	-0,068	0,368	0,095	0,386	-0,075	0,301	-0,200	-0,314
prf	0,593	0,595		0,674	0,137	0,538	0,336	0,483	0,061	0,454	-0,353	-0,077
env	0,595	0,611	0,674		0,256	0,613	0,428	0,455	0,208	0,483	-0,434	-0,030
bgr	-0,091	-0,068	0,137	0,256		0,370	0,523	0,059	0,316	0,218	-0,272	0,314
rwd	0,386	0,368	0,538	0,613	0,370		0,646	0,458	0,486	0,587	-0,558	0,285
rsz	0,204	0,095	0,336	0,428	0,523	0,646		0,276	0,553	0,397	-0,465	0,527
rsp	0,387	0,386	0,483	0,455	0,059	0,458	0,276		0,090	0,434	-0,222	-0,108
rlv	-0,022	-0,075	0,061	0,208	0,316	0,486	0,553	0,090		0,303	-0,383	0,509
swd	0,252	0,301	0,454	0,483	0,218	0,587	0,397	0,434	0,303		-0,602	0,129
loc	-0,203	-0,200	-0,353	-0,434	-0,272	-0,558	-0,465	-0,222	-0,383	-0,602		-0,336
dis	-0,141	-0,314	-0,077	-0,030	0,314	0,285	0,527	-0,108	0,509	0,129	-0,336	

Pearson's product moment correlation coefficients for the correlation between attributes.

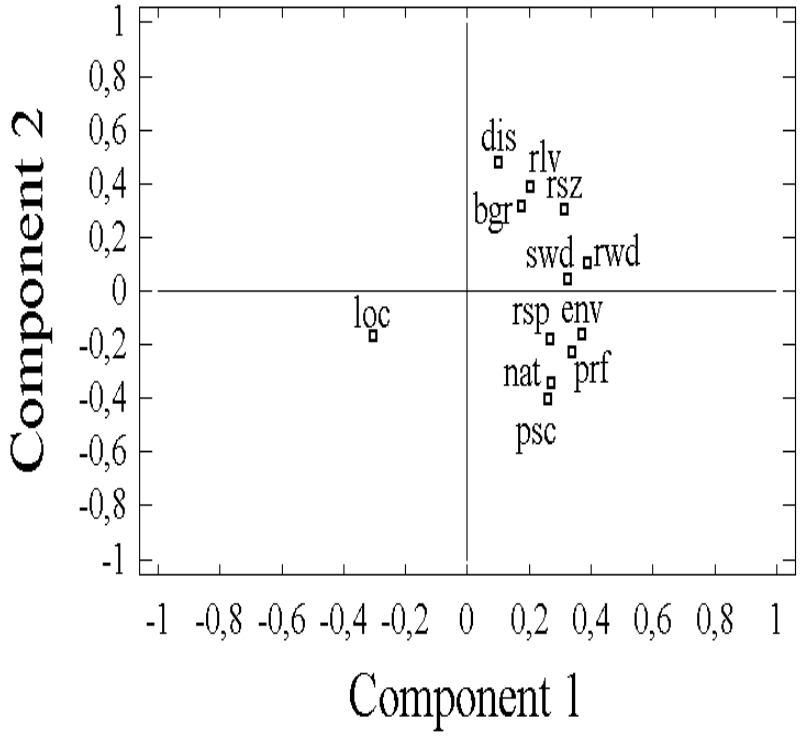
APPENDIX B - CONTINUED

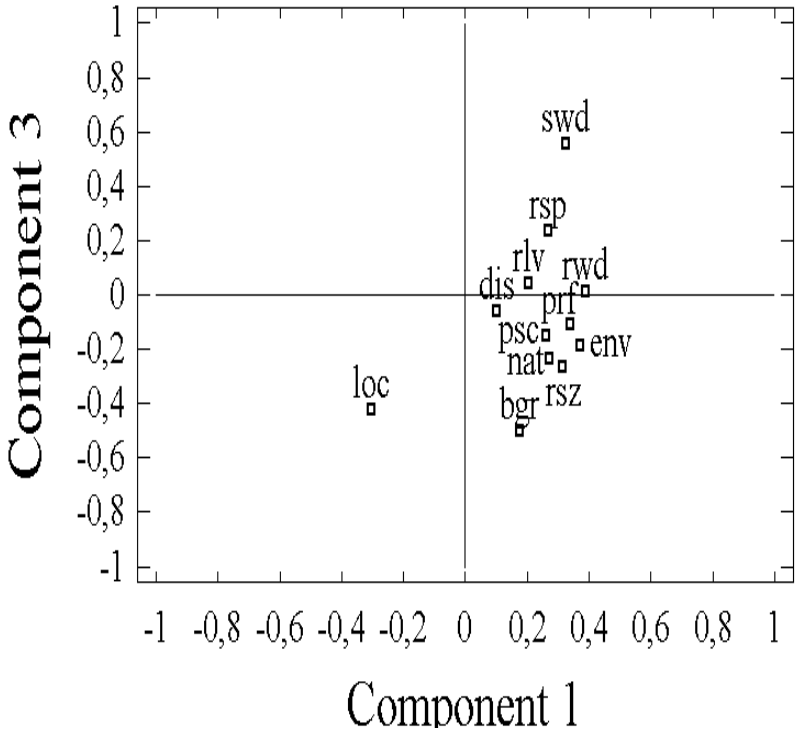
Attribute		Factor	Sums of squares	Degrees of freedom	Mean square	F-ratio	p
Naturalness	nat	A:item	98,8848	12	8,2404	13,79	0,0000
		B:subno	0	18	0	0,0000	1,0000
		RESIDUAL	129,115	216	0,597756		
Presence	psc	A:item	119,702	12	9,97513	19,9	0,0000
		B:subno	0	18	0	0,0000	1,0000
		RESIDUAL	108,298	216	0,501382		
Preference	prf	A:item	140,598	12	11,7165	28,96	0,0000
		B:subno	0	18	0	0,0000	1,0000
		RESIDUAL	87,4021	216	0,404639		
Envelopment	env	A:item	156,678	12	13,0565	39,54	0,0000
		B:subno	0	18	0	0,0000	1,0000
		RESIDUAL	71,322	216	0,330194		
Background noise	bgr	A:item	148,597	12	12,3831	33,69	0,0000
		B:subno	0	18	0	0,0000	1,0000
		RESIDUAL	79,4033	216	0,367608		
Room width	rwd	A:item	140,244	12	11,687	28,77	0,0000
		B:subno	0	18	0	0,0000	1,0000
		RESIDUAL	87,7561	216	0,406278		
Room size	rsz	A:item	150,708	12	12,559	35,1	0,0000
		B:subno	0	18	0	0,0000	1,0000
		RESIDUAL	77,2917	216	0,357832		
Room spectral bandwidth	rsp	A:item	82,3273	12	6,86061	10,17	0,0000
		B:subno	0	18	0	0,0000	1,0000
		RESIDUAL	145,673	216	0,67441		
Room sound level	rlv	A:item	99,0458	12	8,25381	13,83	0,0000
		B:subno	0	18	0	0,0000	1,0000
		RESIDUAL	128,954	216	0,59701		
Source width	swd	A:item	98,0533	12	8,17111	13,58	0,0000
		B:subno	0	18	0	0,0000	1,0000
		RESIDUAL	129,947	216	0,601605		
Localisation	loc	A:item	93,5909	12	7,79924	12,53	0,0000
		B:subno	0	18	0	0,0000	1,0000
		RESIDUAL	134,409	216	0,622264		
Source distance	dis	A:item	147,279	12	12,2733	32,84	0,0000
		B:subno	0	18	0	0,0000	1,0000
		RESIDUAL	80,7207	216	0,373707		

Analysis of variance for attributes.

APPENDIX C

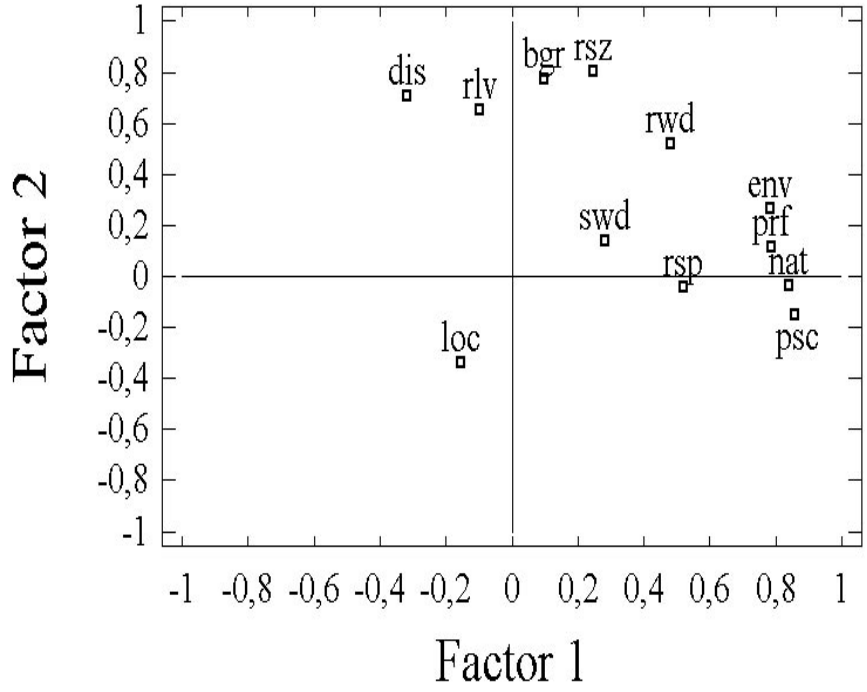
Principal components plots

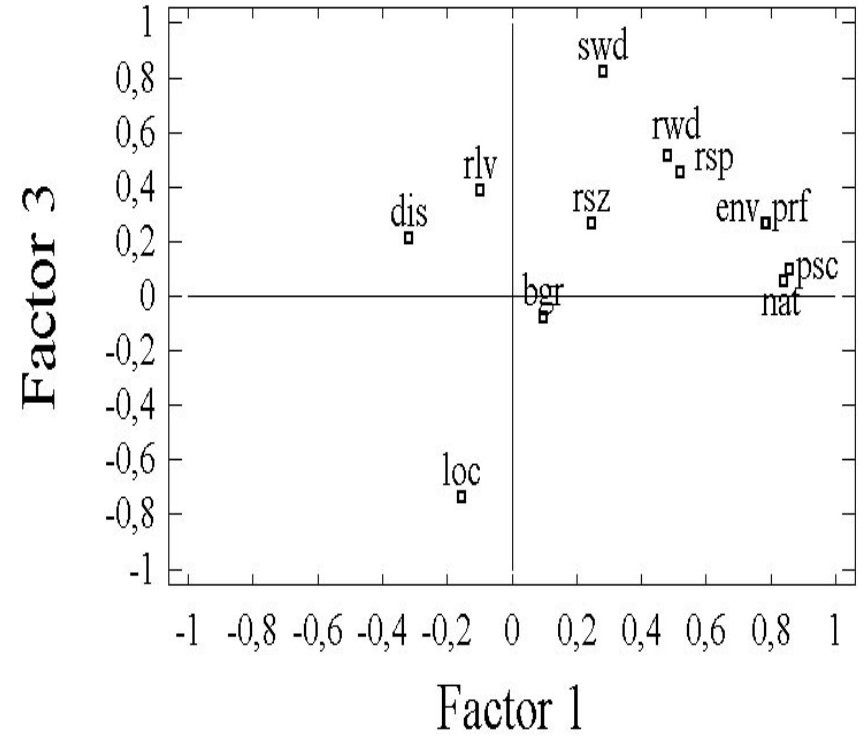




Plots of the three extracted components from the PCA.

APPENDIX C continued
Factor analysis plots





Plots of the three rotated factors in the factor analysis