Extraction of Auditory Features and Elicitation of Attributes for the Assessment of Multichannel Reproduced Sound*

SYLVAIN CHOISEL,** *AES Member,* **AND FLORIAN WICKELMAIER**
(sc@acoustics.aau.dk) (florian.wickelmaier@med.uni-muenchen.de)

Sound Quality Research Unit, Department of Acoustics, Aalborg University, 9220 Aalborg, Denmark

The identification of relevant auditory attributes is pivotal in sound quality evaluation. Two fundamentally different psychometric methods were employed to uncover perceptually relevant auditory features of multichannel reproduced sound. In the first method, called repertory grid technique (RGT), subjects were asked to assign verbal labels directly to the features when encountering them, and to subsequently rate the sounds on the scales thus obtained. The second method, perceptual structure analysis (PSA), required the subjects to consistently use the perceptually relevant features in triadic comparisons, without having to assign them a verbal label; given sufficient consistency, a lattice representation—as frequently used in formal concept analysis (FCA)—can be derived to depict the structure of auditory features.

0 INTRODUCTION

The assessment of sound quality is a multidimensional problem, in which a crucial part is concerned with the identification of perceptual dimensions, or auditory attributes. The elicitation of relevant attributes is not straightforward, and it has attracted increasing interest in the last few years. A generalized set of attributes would certainly help research on sound quality by allowing standardized assessments and improving comparability between studies. However, considering the diversity of the applications, it is more likely that a list of specific attributes will have to be established for each context.

Recently several studies have addressed the problem of eliciting auditory attributes in the context of reproduced sound [1]–[6]. In [1] Bech gives an introduction to descriptive analysis (DA), a technique frequently used in other sensory research such as food quality [7]. In this method a panel of trained listeners jointly develops a set of verbal descriptors. The repertory grid technique (RGT) [2], [8], on the other hand, is based on individual elicitation and is applicable to expert as well as nonexpert listeners. Various verbalization methods have been used in other studies (such as [3] or [4]), with the same goal of arriving at a common descriptive language for auditory perception, by reducing the redundancy within the subjects' verbal descriptors.

All of these direct elicitation methods rely on the basic assumption of a close correspondence between a sensation on the one hand and its verbal descriptor on the other. This is problematic in at least two ways. First, the elicitation of auditory attributes will be dependent on the availability of an adequate label in the subject's lexicon. This means that the verbal abilities of a participant will always bias the outcome of direct elicitation procedures. Second, it cannot be ensured that when a listener provides a verbal expression, it is related to an actual sensation at all. That the listener had the sensation of, for example, being enveloped by the sound field, when he or she said that it was "enveloping," is assumed, not justified. This becomes even more of a problem if the elicitation procedure encourages the subject to produce many descriptors for a given set of sounds.

Indirect methods have been developed in order to disentangle sensation and verbalization, such as multidimensional scaling (MDS, see for example [9]), which aims at uncovering salient perceptual dimensions without having the subject name them, or even be aware of them. MDS is frequently used as an exploratory analysis tool in sound quality evaluation (for example, [10]) and only requires a judgment of perceived distances between stimuli, typically in the form of dissimilarity ratings. An outcome is a map of the stimuli in a multidimensional space. The interpre-

*Portions of this work were presented at the 118th Convention of the Audio Engineering Society, Barcelona, Spain, 2005 May 28–31. Manuscript received 2005 September 22; revised 2006 June 22.

**Also with Bang & Olufsen A/S, 7600 Struer, Denmark.

tation of the dimensions, however, is not straightforward, but often requires additional knowledge about the stimuli, for instance, as obtained from (subjective) rating scales.

This paper presents perceptual structure analysis (PSA) as a novel method to extract auditory features from a set of sounds. The method is based on Heller [11], who developed the measurement-theoretical framework for an experimental procedure to extract semantic features of verbal concepts. Heller's approach is based on knowledge space theory [12], a formalized theory for the representation and assessment of knowledge, and on formal concept analysis (FCA) [13], a technique derived from applied set theory, which results in a graphical representation of data structures by means of lattice diagrams. Heller's method was adapted to the extraction of auditory features and was tested experimentally with synthetic sounds by Wickelmaier and Ellermeier [14]. In addition to its mathematical foundation, the major advantage of PSA lies in the fact that it strictly separates the identification of auditory sensations from their labeling and allows for testing the identifiability of the extracted features.

In the present study both RGT and PSA were used as methods to elicit auditory attributes in the context of multichannel reproduced sound. Both elicitation methods are introduced and illustrated using results of an experiment on the perception of a common set of sounds consisting of various reproduction modes: mono, stereo, and several multichannel formats.

1 METHOD

1.1 Setup and Stimuli

1.1.1 Experimental Setup

The loudspeaker configuration (represented in Fig. 1) consisted of a five-channel surround setup following the ITU-R BS.775-1 recommendation [15], with two additional loudspeakers at ±45°. This configuration allows for the reproduction of mono, stereo, and 5.0 multichannel formats as well as wide-angle stereo. The loudspeakers were Genelec 1031A monitors, placed in a listening room complying with the ITU-R BS.1116 requirements [16]. It had an area of 60 m^2 and a reverberation time of between 0.25 and 0.45 s. The setup was hidden from the subject by a curtain. The sounds were played back by a PC placed in the control room, equipped with an RME Hammerfall HDSP sound card connected to an eight-channel digital-to-analog (D/A) converter (RME ADI-8 DS).

The response interface consisted of a 15-in flat screen placed in front of the listener, at a height of 0.45 m above the floor (in its center), a keyboard, and an optical mouse. A head rest fixed to the armchair ensured that the subject's head was always centered during the listening test. This could be monitored from the control room via a camera fixed to the ceiling above the listener.

1.1.2 Program Material

Four musical excerpts (two pop, two classical) were selected from commercially available multichannel material (Table 1). Their different musical contents (genre, instrumental versus vocal) as well as the varying spatial information present in the multichannel mix (natural room reverberation in the classical recordings, distributed instruments in the pop music) made this selection suitable for eliciting different spatial sensations. The two classical recordings were made with five omnidirectional micro phones placed in a circular array, and the two pop recordings were mixed with a standard surround panning technique. These excerpts were transfered from their original medium—Super Audio Compact Disc (SACD) or Digital Versatile Disc-Audio (DVD-A)—onto a computer (48 kHz, 24 bit) using a Denon 2200 player connected to an eight-channel A/D converter (RME ADI-8 DS), and carefully cut to include a musical phrase, their duration ranging from 4.7 to 5.4 s.

1.1.3 Downmixing and Upmixing

From the original five-channel recordings several formats were derived, as shown in Table 2. When present

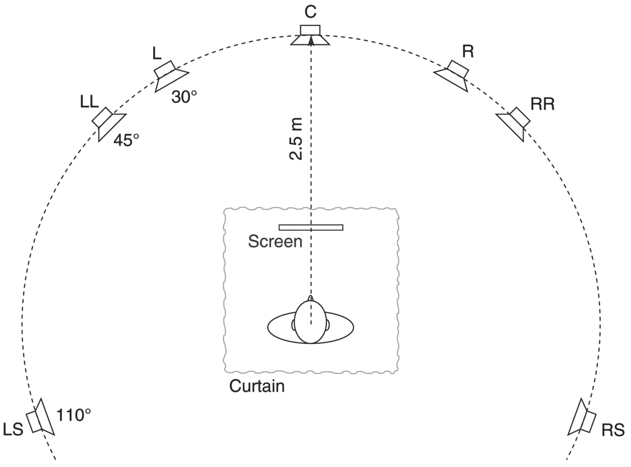


Fig. 1. Playback setup consisting of seven loudspeakers: left (L), right (R), center (C), left-of-left (LL), right-of-right (RR), left surround (LS), and right surround (RS). Setup was placed symmetrically with respect to room width and was hidden from subject by an acoustically transparent curtain. A computer flat screen was used as a response interface.

Table 1. List of musical program material.

Disc	Title	Medium	Track	Time
Beethoven: Piano Sonatas Nos. 21, 23 & 26—Kodama	Sonata 21, op. 53 (Rondo)	SACD	03	1'51
Rachmaninov: Vespers—St. Petersburg Chamber Choir/Korniev	Blazen Muzh	SACD	03	2'04
Steely Dan: Everything Must Go	Everything Must Go	DVD-A	09	0'52
Sting: Sacred Love	Stolen Car	SACD	06	1'55

(only in the pop recordings), the low-frequency effect (LFE) channel was disregarded.

The original five-channel program material was mixed down to stereo using Eq. (1), as recommended in [15],

$$L_{st} = \frac{1}{\sqrt{2}}\left(L + \frac{1}{\sqrt{2}}C + \frac{1}{\sqrt{2}}LS\right)$$
$$R_{st} = \frac{1}{\sqrt{2}}\left(R + \frac{1}{\sqrt{2}}C + \frac{1}{\sqrt{2}}RS\right). \quad (1)$$

From the stereo version, mono and phantom mono were computed as described by Eqs. (2) and (3), respectively,

$$C_{mo} = \frac{1}{\sqrt{2}}(L_{st} + R_{st}) \quad (2)$$

$$L_{ph} = R_{ph} = \frac{1}{2}(L_{st} + R_{st}). \quad (3)$$

All processing was done in MATLAB using floating-point precision, and all intermediate files were stored with 24-bit resolution. The wide stereo format was identical to stereo, but played on loudspeakers LL and RR positioned at ±45°.

Finally three upmixing algorithms were used to reconstruct a multichannel sound from the stereo downmix: two commercially available algorithms, Dolby Pro Logic II and DTS Neo:6 (referred to as upmixing 1 and 2, in no specific order), and a simple matrix decoding algorithm. Dolby Pro Logic II was implemented on a Meridian 861 surround processor, with parameters defined in Table 3. The processor was fed with a digital signal (S/PDIF) coming from the RME sound card, and the five analog output signals were recorded through the RME converter using 24-bit resolution. A Yamaha RX-V 640 receiver was used as a DTS Neo:6 decoder. The only parameter (C. Image) was set to default (0.3). The matrix upmixing was implemented in MATLAB. The left and right surround channels were fed with the difference between the left and right signals (L − R and R − L, respectively) attenuated by 6 dB [Eq. (4)],

Table 2. Reproduction modes (see Fig. 1).

Name	Abbreviation	Loudspeakers Used
Mono	mo	C
Phantom mono	ph	L,R
Stereo	st	L,R
Wide stereo	ws	LL,RR
Matrix upmixing	ma	L,R,LS,RS
Dolby Pro Logic II	*	L,R,C,LS,RS
DTS Neo:6	*	L,R,C,LS,RS
Original 5.0	or	L,R,C,LS,RS

* Referred to u1 and u2 (in no specific order) in the rest of this paper.

Table 3. Parameters for Dolby Pro Logic II upmix on Meridian 861 processor.

Parameter	Value
Treble	+1
Bass	−2
Balance	0
Center	0 dB
Depth	0.5
Width	3
Dimension	+1
Panorama	No
Rear	0 dB
R delay	0.0
Lip sync	0.0

$$\begin{aligned} L_{ma} &= L \\ R_{ma} &= R \\ LS_{ma} &= \frac{1}{2}(L - R) \\ RS_{ma} &= \frac{1}{2}(R - L). \end{aligned} \quad (4)$$

1.1.4 Equalization and Calibration

The frequency responses of the seven loudspeakers were measured in an anechoic chamber by means of a 14th-order maximum-length sequence (MLS) at a sampling frequency of 48 kHz, using a microphone (B&K 4133) placed at 2.5-m distance. After an adjustment of the sensitivities, the loudspeakers showed differences up to 1 dB at some frequencies. In order to match them further, FIR filters were designed to equalize for their anechoic frequency responses. They were calculated based on the first 7000 samples of the impulse responses and were truncated to 1024 samples. The loudspeaker responses and the calculated filters can be seen in Fig. 2(a), the resulting equalized responses are shown in Fig. 2(b).

Each channel of the stimuli was equalized using the corresponding filter. This equalization was based on the

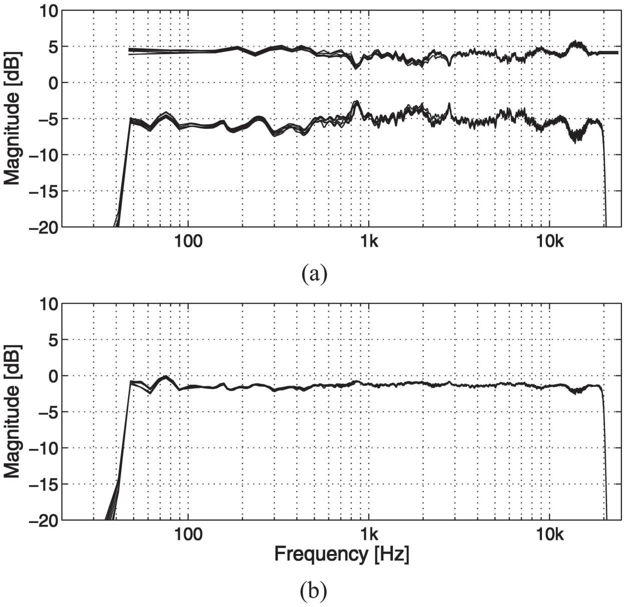


Fig. 2. (a) Measured frequency responses of seven loudspeakers (lower curves) and their corresponding equalization filters (upper curves). (b) Equalized loudspeaker responses.

anechoic measurements, and no attempt was made to correct for the response in the listening room. Floor, wall, and ceiling reflections, as well as standing waves, might affect the sound differently for each channel, resulting in interchannel level differences. Therefore it is recommended to align the playback level of the individual channels [17], [18], but there does not seem to be a general agreement on what stimulus to use for this purpose. In the present study, band-limited pink noise (200 Hz to 2 kHz) was used and recorded for 10 s at the listening position using a B&K 4134 pressure-field microphone pointing upward. This signal was equalized for each channel based on the anechoic loudspeaker responses, in the same way for all the musical excerpts. The A-weighted sound pressure level was then calculated from the recordings. The interchannel level differences were measured to be within 0.3 dB, but the differences between left/right pairs did not exceed 0.1 dB. Therefore no additional correction was applied.

1.1.5 Loudness Matching

After the interchannel level alignment the next goal was to obtain the gains to be applied to the reproduction modes, in order to eliminate loudness differences as much as possible between them. Thereby no attempt was made to match the four types of program material in loudness. Rather, each of them was adjusted to a comfortable level by the experimenters.

Eight subjects (six male, two female) performed the loudness-matching task. All of them were experienced listeners, either professionally involved in acoustics or having extensive experience in subjective listening tests, but were not taking part in the main experiment. The different reproduction modes in Table 2 were matched in loudness by using an adaptive procedure (2AFC, 1 up/1 down [19], [20]). On each trial the task of the subject was to decide which of the two presented sounds was louder, one being the standard, the other the comparison, in random order. When the listener indicated that the comparison was louder than the standard, the level of the comparison was reduced; it was increased otherwise. After four reversals the step size was halved from 1 to 0.5 dB. After eight reversals the track was completed, and the average level of the last four reversals yielded an estimate of the loudness match, or point of subjective equality. The first two seconds of the musical excerpts in the eight reproduction modes served as stimuli. For all four types of program material the standard was chosen to be the stereo reproduction mode. Its playback level was adjusted beforehand and measured in the listening position to be 65.8, 59.4, 66.5, and 67.7 dB A-weighted SPL, respectively (averaged over the duration of the stimuli). In order for the procedure to be less transparent for the subject, the eight adaptive tracks were interleaved randomly in a single block, with a probability proportional to the number of remaining reversals. Each track had a random starting level of between ±3 dB. On average one block lasted 12.7 min for Beethoven. 14.0 min for Rachmanivov, 12.9 min for Steely Dan, and 15.2 min for Sting. Altogether the eight subjects gave 3808 loudness judgments.

The resulting matches (averaged across subjects) were applied as gains to the final stimuli. After equalization and loudness matching, all sounds were saved as multichannel wave files, dithered and quantized to 16-bit (±1 LSB, triangular probability density function) and with a sampling frequency of 48 kHz.

1.2 Subjects

Thirty-nine listeners (27 male, 12 female) were selected among 78 candidates, according to their listening abilities and verbal fluency (see [21] for details on the selection procedure).[1] In summary the selected subjects outperformed the rejected ones in their ability to discriminate sounds varying in stereo width, and in their ability to promptly produce words belonging to different semantic categories. The selected participants were all native Danish speakers, and their ages ranged from 21 to 39 (median 24). Because of their participation in a previous experiment, all subjects were familiar with the stimuli. The subjects were assigned randomly to one of three groups, two of which took part in the repertory grid technique, the third in the perceptual structure analysis.

1.3 Repertory Grid Technique

The repertory grid technique (RGT) typically consists of two parts—an elicitation part, in which the subject describes in what way the sounds differ or are alike, and a rating part, in which the stimuli are rated along the elicited descriptors.

1.3.1 Elicitation of Verbal Descriptors

A triadic elicitation procedure was implemented following Berg and Rumsey [2]. On each trial the subject was presented with a triple of sounds and instructed to indicate which of the three sounds differed most from the other two. He or she was then asked *in what way* the selected sound differed from the other two, and *in what way* the other two were alike. A pair of words or expressions was thus obtained for each triple, which were used later as poles of a rating scale. The subject was allowed to reuse already mentioned descriptors, available in a pull-down list. He or she also had the possibility to listen to the sounds as many times as needed.

An advantage of this triadic elicitation method is that it avoids asking the subjects explicitly for opposite expressions. Rather, it was assumed that asking the subjects to describe first the similarities between two stimuli and then the differences from the third one would implicitly elicit descriptors opposite in meaning. A disadvantage, however, of using stimulus triples is that salient differences between two sounds might be overlooked if they are always presented together with a more dissimilar sound.

Therefore an alternative elicitation method was employed in addition, using pairs of stimuli. The subjects were asked to describe the difference between sounds *a* and *b* with a pair of opposite words or expressions. Ten

[1] One of the 40 participants originally selected had left the panel before the present experiment.

subjects took part in the triadic elicitation (referred to as RGT-3), whereas another group of ten took part in the pairwise elicitation (RGT-2). Because of the higher number of triples (56) than pairs (28), the subjects in the first group performed the task on only two types of program material (balanced across subjects), whereas the second group completed the task for all four types of music.

A possible problem in the elicitation phase is that, in a given pair (or triple), a dominant attribute might mask another, less dominant but yet important attribute. Berg [22, p. 3] addresses this so-called *construct masking* by iteratively querying the listener for a given set of sounds and, thereby, "exhausting all perceived attributes of the stimulus, or combination of stimuli, under consideration by the subject, before the stimuli are changed." In the present study the responses were not limited to one word only, and it was possible to enter several descriptors per triple. Subjects were not, however, explicitly encouraged to do so. Nonetheless, it was hypothesized that pairwise elicitation is less prone to construct masking than is triadic elicitation, for the reasons mentioned.

1.3.2 Scaling

The scaling procedure remained identical for the two groups (RGT-3 and RGT-2). For each pair of opposite descriptors the eight reproduction modes were to be rated by making a mark on a line using the mouse. The descriptors were shown at the top left- and top right-hand side of the screen; underneath eight lines were displayed, and eight buttons (labeled from A to H) placed next to them allowed for the playback of the sounds. Once all sounds were rated, the subject could proceed to the next pair of descriptors. The order of the reproduction modes was randomized on each trial.

1.3.3 Reduction to Fewer Attributes

When a large number of descriptors is obtained, it might be desirable to reduce them to fewer—ideally independent—attributes. Two main approaches are typically used. The first involves classifying the verbal data into semantic categories (for example, [3]), the second makes use of the ratings of the stimuli on the elicited descriptors [2]. For the latter approach several statistical methods are available to reduce the dimensionality of a set of variables, the most common ones being factor analysis, principal component analysis, and cluster analysis. The latter was used in this study.

Cluster analysis was performed on the ratings associated with each descriptor, in a similar way as proposed by Berg and Rumsey [2]. First, a matrix of distances between the scales was calculated; the distance between two scales X_i and X_j was chosen as $d_{ij} = 1 - |r_{ij}|$, where r_{ij} is the correlation coefficient between the two scales. Uncorrelated scales will therefore be at a distance of 1, while highly correlated scales, either positively or negatively, would result in a distance close to 0. From the distances the cluster analysis derives a treelike representation, the so-called dendrogram, where the descriptors/scales are the leaves and the nodes are clusters. The closer to the bottom two leaves are connected in the dendrogram (the lower the clustering level), the more similarly the two corresponding scales were used by the subject. Verbal descriptors clustering together can then be merged into a common construct, according to a criterion chosen by the researcher.

1.4 Perceptual Structure Analysis

Perceptual structure analysis (PSA) attempts to extract auditory features arising from a set of sounds. A feature may be defined as some perceptual effect which allows the subject to categorize the stimuli. In this section the basic theoretical background is introduced (for more details, the reader is referred to [11], [14]), and the experimental and analysis procedures used in this study are presented.

1.4.1 From Triadic Comparisons to a Feature Representation

Let X denote the total set of sounds under study, the so-called *domain,* and σ a collection of subsets of X, which will be interpreted as the set of auditory features of the sounds in X. In accordance with [11], $<X, \sigma>$ is called a *perceptual structure.* Fig. 3 displays the lattice graph of a hypothetical perceptual structure $\sigma = \{\emptyset, \{a\}, \{b\}, \{c\}, \{d\}, \{a, b, c\}, X\}$ on the domain $X = \{a, b, c, d\}$. Each node in this graph represents a feature shared by the sounds connected to it. Let $A \subseteq X$ denote a subset of X, and $\sigma(A)$ the intersection of all sets in σ of which A is a subset,

$$\sigma(A) = \bigcap_{A \subseteq S,\, S \in \sigma} S.$$

This means that $\sigma(A)$ is the smallest set in σ which includes the sounds in A. In the example shown in Fig. 3, $\sigma(\{a, b\}) = \{a, b, c\}$, implying that all features shared by a and b are also shared by c; and $\sigma(\{a, d\}) = X$, implying that a and d do not share any other feature than the one shared by all sounds in X.

A relation Q that relates the subsets of X to X can be defined in the following way. The sounds in A are said to be in relation to a sound $x \in X$, formally AQx, if and only if the subject answers "No" to the question:

Do the sounds in A share a feature that x does not have?

If the answer is "Yes," the relation between A and x does not hold, formally $A\overline{Q}x$.

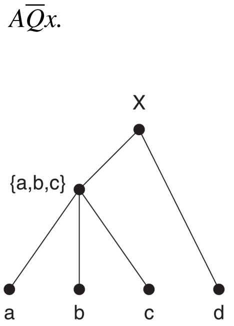


Fig. 3. Lattice graph of hypothetical perceptual structure. Sounds are denoted by *a, b, c,* and *d;* $\{a, b, c\}$ represents a feature shared by *a, b,* and *c,* but not by *d;* $X = \{a, b, c, d\}$ is the domain.

The relation Q is said to be *transitive* if

$$AQb(\forall b \in B) \text{ and } BQc \;\Rightarrow\; AQc \qquad (5)$$

for all nonempty $A, B \subseteq X,$ and $c \in X.$

To illustrate this, let us assume that Q has been established by querying a subject and that the responses $\{a, b\}Qc$ ("No") and $\{a, b\}\overline{Q}d$ ("Yes") have been observed, which are in line with the structure in Fig. 3. If in addition, however, the response $\{b, c\}Qd$ ("No") was given—and assuming $\{a, b\}Qb$ to hold trivially—it follows from Eq. (5) that transitivity is violated, since transitivity would require that

$$\{a, b\}Qb, \{a, b\}Qc \text{ and } \{b, c\}Qd \;\Rightarrow\; \{a, b\}Qd$$

and consequently a perceptual structure cannot be derived, given this pattern of responses.

If and only if transitivity holds, Q can be represented by a perceptual structure σ on X such that

$$AQb \quad \text{if and only if} \quad b \in \sigma(A) \qquad (6)$$

for all nonempty $A \subseteq X$ and $b \in X$ (see [11, theorem 2]).

If the set of all sounds that are in relation to A is defined as

$$AQ = \{x \in X : AQx\}$$

then it follows from Eq. (6) that $AQ = \sigma(A)$. In an experiment $\sigma(A)$ will have to be determined from the responses, and the perceptual structure σ can then be constructed by

$$\sigma = \{\sigma(A) : A \subseteq X\}.$$

In practice the number of questions is usually too large to be accommodated in an experimental session if all possible subsets $A \subseteq X$ are presented. For n stimuli, $(2^n - 1)n$ questions are required, which would amount to 2040 questions for $n = 8$. For that reason, in the present experiment the subsets are restricted to containing pairs of sounds only. The consequence of such an incomplete design is a potentially nonunique representation [11], [14], meaning that the subject's responses might result in more than one representing perceptual structure. This uniqueness problem is addressed in a later section.

Single-element subsets are not included in the querying procedure, assuming that the sounds can be discriminated, and therefore each has at least one characteristic feature. Finally Q is assumed to hold for both sounds in each pair (Q is assumed to be reflexive), that is, $\{a, b\}Qa$ and $\{a, b\}Qb,$ so that each pair will only be presented together with the remaining $|X| - 2$ sounds.

1.4.2 Experimental Procedure

Let n be the number of sounds under study in an experiment. For each of the $n(n - 1)/2$ unordered pairs of sounds $\{a, b\}$ and each of the $n - 2$ remaining sounds $c,$ the following question was asked:

> Do sounds a and b share a feature that c does not have?

The number of such triples $\{a, b, c\}$ is $n(n - 1)(n - 2)/2.$ This procedure relies on—and thereby verifies—the ability of the subject to consistently identify the salient features in a certain context given by a set of stimuli. In order to increase consistency of the judgments, it will help if the subject develops a clear idea of the features already prior to proceeding with the triadic comparisons. For that purpose a familiarization session preceded the main experiment, in which the subject was instructed to listen to the sounds (arranged in a playlist) as many times as needed, and identify the features characterizing the sounds. Previous to the familiarization with the sounds, in order to introduce the participants to the task, they went through a short tutorial using drawings (simple geometric shapes having strongly salient visual features) together with the experimenter before the task was applied to sound stimuli. Such a tutorial has the advantage that 1) the task is intuitive when visual stimuli are used and 2) it avoids introducing auditory concepts that might bias the judgments.

Each triple of sounds was presented twice, in two different sessions. All triples for which the two responses did not agree were presented a third time in a third session. Because of the many questions required by this method, only $n = 7$ reproduction modes were included in the experimental design; the matrix-upmixed format [Eq. (4)] was removed from the stimuli. With 105 triples each session was completed in one hour including breaks. Nineteen of the 39 subjects participated in this experimental procedure, with only one type of program material each.

1.4.3 Fitting Perceptual Structures

In principle a single violation of transitivity [Eq. (5)] in the responses of a subject prevents their representation by a perceptual structure. If only a few violations have occurred, it is possible, however, to inspect the pattern of violations visually and attempt to resolve them by reversing as few responses as possible from "Yes" to "No," or vice versa. Often a single response alteration can account for several violations, which makes it likely that these violations are the result of a careless error rather than of systematic inconsistencies. Such a manual procedure, however, is quite cumbersome, especially if one is interested in minimizing the number of response alterations, or evaluate several solutions to resolve the violations.

Therefore a computer program was developed,[2] which searches for the best solution allowing a feature representation, that is, the structure that best fits the subject's responses. A brute-force procedure, such as testing all possible structures, is not practically viable—with $n = 7$ stimuli, there are $2^n = 128$ possible sets. Excluding trivial sets such as the singletons, the empty set $\emptyset$, and the domain X—which would not affect the goodness of fit—there are $2^{(2n-n-2)} = 2^{119} = 6.6 \times 10^{35}$ possible structures. Testing all possible response alterations might also be a possible approach, but the complexity of such an algorithm quickly rises with the number of response alterations. By contrast the proposed method attempts to infer the features that potentially underlie the subject's responses (see [14] for an

[2] Software is available from the authors upon request.

example). In order to do that, it assumes that violations are caused by the subject either overlooking a feature or, on the contrary, erroneously identifying a feature in the local context of a given triple.

From transitivity [Eq. (5)] it follows that $B \subseteq AQ \Rightarrow BQ \subseteq AQ$ for all nonempty subsets A, B of X. Consequently a transitivity violation occurs whenever $B \subseteq AQ$ is true, but $BQ \nsubseteq AQ$ is observed. The violations can be resolved in several ways:

1) Replace AQ by $AQ \cup BQ$.

2) Remove from BQ the elements that are not in AQ.

3) Remove from AQ one or both elements of B, with the restriction that AQ must still contain A.

This can be interpreted as follows. Generally, adding elements to either AQ or BQ corresponds to changing responses from "Yes" to "No," suggesting that a feature has been erroneously identified, whereas removing elements corresponds to changing responses from "No" to "Yes," suggesting that a feature has been overlooked by the subject.

Once a list of possible sets is created from all AQ plus all modified versions according to these three rules, a simulation is performed to estimate the best fitting structure. For all combinations of these candidate sets, the answers in the triadic comparisons are predicted and compared to the observed answers. The number of answers differing between these two sets of responses is used as a measure of fit. The outcome is a list of structures ordered by the number of reversed answers. A simulation of response patterns was not attempted when the observed judgments were classified as unreliable or inconsistent. Unfortunately no simple criteria for such a classification are at hand. Rather, the indices of reliability and consistency (see Table 5), and their development over time (sessions), have to be considered together. Once a potential structure is found, the number of response alterations needed to resolve the transitivity violations might serve as a further index of validity (see Section 2). In the present application a limit for alterations was set at 10% of the observed responses.

1.4.4 Uniqueness Problem

Because the relation Q was established with only pairs of sounds rather than all possible subsets, there is potentially more than one structure representing the responses. In order not to omit any feature, the largest of these representing structures was selected which, according to the uniqueness theorem [11, theorem 3], contains all the other solutions. Whether or not all features could actually be identified by the subject might be inferred from the outcome of a structured debriefing session described in the next subsection.

Furthermore, in the case of transitivity violations the fitting procedure can potentially return several solutions at the same distance to the subject's responses, resulting in another source of uncertainty about the representation. The task is left to the experimenter to choose among the possible structures. The strategy applied in the present study was to choose the structure with most features at minimal distance. Here again, the descriptions collected during the debriefing session might help to clarify whether or not all proposed features had been identified.

1.4.5 Labeling of the Features

When a perceptual structure was obtained, the subject was asked to label each feature during a debriefing session after the actual data collection. All seven sounds were arranged in a play list, some of which—those sharing a common feature according to the structure—were marked with a bullet. The question to the subject was, "What feature do the marked sounds share, which the other sounds do not have?" On each trial the participant entered a short description of the common feature using the keyboard. Subsequently the next feature of his or her perceptual structure was presented. The subject had the possibility of giving no description when he or she did not recognize a feature.

2 RESULTS

2.1 Repertory Grid Technique

The number of verbal descriptors elicited using RGT are shown in Table 4. Overall a slightly higher number of descriptors was obtained per subject with the pairwise (RGT-2) than with the triadic (RGT-3) elicitation, in spite of the fact that the latter involved more comparisons (there are more triples than pairs). This might hint at a slight advantage of RGT-2 over RGT-3 with respect to construct masking (Section 1.3.1).

In order to reduce the number of descriptors, cluster analysis was performed on the ratings associated with each descriptor. This analysis was done individually for each subject. The result for one subject (92) is depicted in Fig. 4 as an example. The choice of a cutoff level—below which sounds clustering together are combined into a single construct—is not straightforward. So far subjects have been encouraged to use their own descriptors, and a grouping of these constructs according to similarities in their ratings requires some interpretation from the researcher. In the present study the same cutoff level was chosen for all subjects after visual inspection of the clusters. A value of 0.3 was found to represent a compromise between the ability to group related constructs together and to discriminate between (at least semantically) unrelated constructs. This resulted in an average number of clusters of 3.6 (Beethoven), 4.2 (Rachmaninov), 5.1 (Steely Dan), and 5.1 (Sting). Berg and Rumsey [2] discussed the different biases possibly introduced by the re-

Table 4. Number of descriptors elicited by two RGT groups.

Program Material	RGT-2			RGT-3		
	Min	Max	Mean	Min	Max	Mean
Beethoven	3	12	7.9	3	9	5.4
Rachmaninov	3	13	8.0	5	13	9.2
Steely Dan	2	19	8.8	2	13	8.6
Sting	4	15	9.9	3	11	8.2

searcher. At this level, in particular, constructs clustering together might erroneously be interpreted as related to the same sensation. In order to reduce such biases associated with cluster labeling, the subsequent attribute selection was based both on the clusters and on the individual constructs. As will be shown in Section 2.3, these two approaches led to similar results.

In addition to the similarity of the descriptors, the similarity of the reproduction modes can be represented as a dendrogram (Fig. 5). The distance was calculated as the correlation between their respective ratings on all scales. Consequently two reproduction modes rated in a similar way on all scales will cluster at a low level. This was generally the case for mono and phantom mono, which reflects their strong perceptual similarity.

2.2 Perceptual Structure Analysis

From the subject's binary (yes/no) responses to each triple of sounds, the simulation attempted to find the best fitting perceptual structure. A measure of fit $\delta_Q(\sigma)$ is reported in Table 5 for all cases where the fitting of a structure had been attempted. It is calculated as the relative number of response alterations necessary to resolve the transitivity violations in Q in order to be consistent with the structure σ: $\delta_Q(\sigma) = c/T$, where c is the number of response alterations and T the total number of triples (105 in the present case). In Table 5 the number of response changes between the first and second sessions (I–II) and between the second and third sessions (II–III) are also reported. This number is an indicator of the subject's reliability; a value close to 50% would suggest that the subject was guessing. Finally the number of transitivity violations is an indicator of how consistently the features were identified in different contexts (different triples of sounds).

Figs. 6 and 7 show the structures obtained for two subjects, with the corresponding labels (translated from Danish

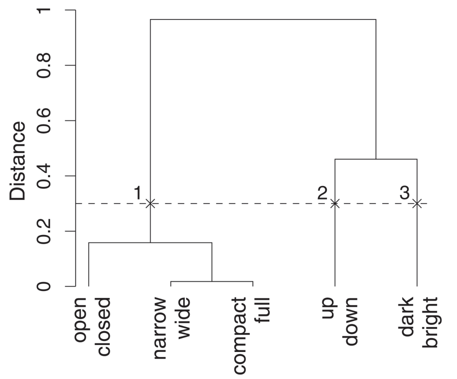


Fig. 4. Cluster analysis of RGT constructs for Sting, rated by subject 92. Distances are derived from absolute correlations. Dashed line indicates cutoff level of 0.3, resulting in three clusters. Verbal descriptors are translated from Danish.

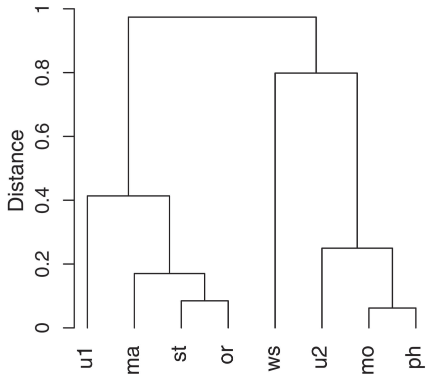
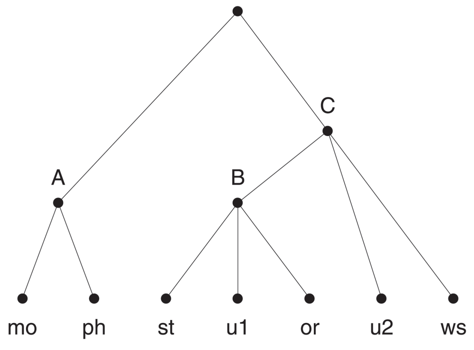


Fig. 5. Cluster analysis of reproduction modes for Sting, rated by subject 92. Distances are derived from correlations.

Table 5. Reliability and consistency of judgments collected in PSA.*

	Response Changes		Transitivity			
Subject	I–II	II–III	I	II	III	$\delta_Q(\sigma)$
Beethoven						
07	32	14	96	80	98	—
29	13	5	33	70	22	0.038
33	20	7	65	43	25	0.048
81	33	14	63	75	50	—
Rachmaninov						
10	17	7	55	59	32	0.095
12	22	11	67	71	55	0.086
35	39	20	75	81	87	—
74	25	6	132	29	30	0.057
88	17	8	49	54	41	0.086
Steely Dan						
04	29	17	67	98	42	0.076
08	18	8	56	74	23	0.057
24	33	15	78	94	74	—
32	39	18	174	113	82	—
39	34	15	91	53	48	—
Sting						
19	33	14	108	138	80	—
27	23	10	46	34	24	0.086
49	47	27	151	123	147	—
73	25	4	80	42	37	0.086
89	26	14	51	70	9	0.038

* Displayed are, for each subject, response changes from session to session, number of transitivity violations in three sessions, and a measure of fit between structures and data. Roman numerals indicate session numbers.



A "Closed. Sound comes from a small area."
B "Deep, bass."
C "Full, volume, depth, reverb."

Fig. 6. Lattice representation of individual perceptual structure for Beethoven (subject 29). Verbal labels are translated from Danish.

ish) obtained during the debriefing session. Each node of the lattice represents a perceptual feature common to the connected reproduction modes. A white node represents a combination of two features not giving rise to a new sensation. This information is obtained from the debriefing session. In the case presented in Fig. 7, *B* referred to the sensation of a nonelevated sound, *C* referred to the width, and the white node deriving from *B* and *C* was labeled as "a wide sound image as well as a sound matching the listening position," which was not considered to be a new feature. From the 19 subjects assigned to this elicitation method, 11 structures were obtained, fitting from 90 to 96% of their answers, that is, from 4 to 10 alterations among 105 answers were necessary. The number of distinct features per structure ranged from 3 to 8.

2.3 Selection of Attributes

So far the elicitation of auditory attributes was performed individually. For each of the 20 listeners assigned to RGT, and for the 11 listeners for whom a feature-structure representation was possible, a number of descriptors was obtained. One goal of the present study, however, was to arrive at a common list of attributes characterizing the sensations evoked by the selected stimuli. For this purpose the individual constructs were sorted into 20 semantic categories emerging from the subjects' descriptors: twelve categories describing the spatial aspects of the sounds—width (wi), envelopment (en), spaciousness (sp), elevation (el), vertical spread (vs), distance (di), depth (de), homogeneity (ho), focus/blur (fo), skew (sk), stability (st), and presence (pr); four categories reflecting the timbral aspects—brightness (br), spectral balance (sb), sharpness (sh), and bass (ba); and four categories not be longing specifically to spatial or timbral aspects—naturalness (na), clarity (cl), loudness (lo), and miscellaneous (mi).

Each labeled feature obtained from PSA was assigned to one of these categories, and the number of occurences in each category was counted separately for each of the four types of program material. For the verbal descriptors obtained from RGT, two strategies were applied. First, the individual descriptor pairs were assigned to one of the 20 categories. The results from the pairwise (RGT-2) and triadic elicitation (RGT-3) were combined for this purpose. The second strategy involved classifying the clusters obtained from cluster analysis. In doing so, the redundancy in the individual descriptors was reduced before the categorization.

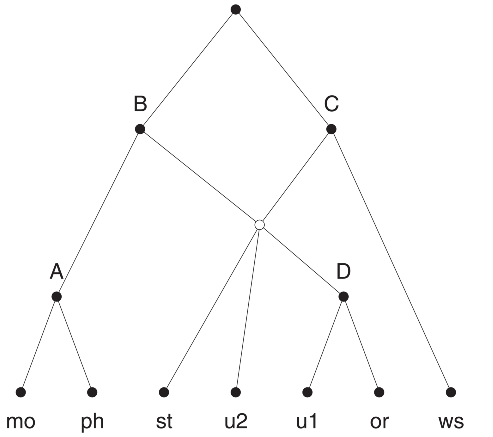
The occurences were then summed across program materials, resulting in one ranking of the categories per technique (PSA, RGT, and cluster analysis of the RGT constructs; see Table 6). Because the descriptors could not be classified without ambiguity in the categories "spectral balance" and "brightness," the scores of spectral balance (sb) were added to those of brightness (br). Furthermore the miscellaneous category was removed from the ranking.

The orders obtained from the RGT constructs and from the cluster labels are very similar, suggesting that the choice of approach would have little impact on the outcome of the attribute selection. Larger differences were observed between the ranking of the RGT constructs and that of the PSA features. It is possible, however, to find a set of attributes common to both methods. Eight categories (out of the 18 displayed in Table 6) appear in the top eleven positions in all three rankings. These are width, envelopment, elevation, spaciousness, brightness, distance, clarity, and naturalness.

3 DISCUSSION

3.1 Repertory Grid Technique

Two groups of ten subjects took part in the RGT, performing either triadic (RGT-3) or pairwise (RGT-2) elicitation. While these two methods yielded a comparable



A "Narrow sound image. The sound comes from one loudspeaker only."
B "The sound comes from a height matching the listening position."
C "Wider sound image (more surround effect)."
D "Surround effect more pronounced."

Fig. 7. Lattice representation of individual perceptual structure for Steely Dan (subject 08). Verbal labels are translated from Danish.

number of descriptors (Table 4), they differed in the difficulty of obtaining bipolar scales from the verbal descriptors. Because the triadic elicitation did not require the subjects to provide opposite words or expressions, some rearrangement of the words by the experimenter was necessary, followed by a verification by the subjects that the two words of each pair had an opposite meaning in the context of the sounds under study. The verbal descriptors elicited by RGT-2, on the other hand, were easier to interpret as end points of a scale. Because of the large number of descriptors obtained (up to 19 per subject), and a certain redundancy among them, a reduction to fewer attributes was performed using cluster analysis. Although synonymous, or at least semantically related, words often grouped together in a same cluster (that is, the stimuli were rated in a similar fashion on the corresponding scales), this was not always the case, and the difficult choice was left to the experimenter to label the heterogeneous clusters. A reason for heterogeneous clusters might be unreliable ratings given by the subjects. Future work should focus on whether including several repetitions of the ratings will result in more homogeneous clusters.

3.2 Perceptual Structure Analysis

An indirect elicitation method was presented, which requires the subjects to consistently identify features in the sounds, without having to name them in the first place. The strength of this method is that consistency of judgments is required in order to obtain a representation of the features. In practice, with complex stimuli such as multichannel reproduced sound, perfect consistency is rarely obtained. Of the 19 subjects who underwent this procedure, none gave responses without any transitivity violations. The difficulty of the task lies in the ability to identify the features independently of the local context. This means that the decision whether or not a feature is present in a sound must not depend on the triple in which this sound is presented. Some features, however, might not be perceived as either present or not present, but rather as present to a variable degree. This does not depend solely on the continuous nature of the sensation, but also on the distribution of the stimuli under study on that given sensation. In the present study the mono and phantom mono sounds were easily identified as narrow, whereas the other sounds could be identified as wide. In the absence of the mono sounds, the feature "wide" would be much harder to attribute, even though the remaining sounds might still differ in width. It must be pointed out that answers based on local similarity in each triple are unlikely to result in a structure representation.

In order to handle transitivity violations, a procedure was proposed that searches for the structure(s) that best fit the subject's responses. Rather than attempting a brute-force simulation of all possible response changes, the information gained from the violations themselves was used to derive a list of features that might have underlain the subject's answers. For eleven of the 19 subjects a structure representation was obtained, with a fit $\delta_Q(\sigma)$ from 90 to 96%. For those eleven subjects the correlation between the number of violations and the number of response alterations was significant [$r = 0.7$; $p = 0.016$]. From this correlation it is reasonable to assume that the fit would have been poorer for the other eight subjects for which no simulation was attempted because of the high number of violations. In the absence of a statistical test to accept or reject the structure representation of the subject's responses in case of violations, $\delta_Q(\sigma)$ might serve as a criterion to assess the validity of the structure. Future methodological developments might consider the adaptation of a probabilistic framework, where response changes and

Table 6. Number of occurrences (Frequency) of PSA features, individual RGT constructs, and RGT clusters in semantic categories.*

PSA		RGT		Clusters	
Category	Frequency	Category	Frequency	Category	Frequency
wi	13	**wi**	60	**wi**	44
en	12	**br**	57	**sp**	37
el	6	**sp**	55	**br**	36
ba	5	**el**	34	**el**	27
br	4	**di**	25	**di**	19
di	4	sh	23	sh	15
sp	4	**en**	20	**cl**	11
cl	2	**na**	16	**na**	11
de	2	**cl**	15	pr	9
vs	2	sk	14	sk	9
na	1	pr	11	**en**	8
ho	1	de	10	de	7
fo	—	lo	7	lo	4
lo	—	ba	6	ba	2
pr	—	fo	4	fo	2
sh	—	vs	4	ho	2
sk	—	ho	3	vs	2
st	—	st	1	st	—

* Categories are explained in text. The eight selected attributes—those having the highest positions in all three rankings—are indicated in boldface.

inconsistencies could be used to estimate the probability of making an error when identifying a feature.

Finally it was generally observed during the debriefing session that the subjects for which no structure was obtained had difficulties producing a list of the features on which they based their answers. This suggests that they did not establish a clear set of features before—or even during—the triadic-comparison task. This, however, seems to be a requirement in order to respond consistently, and for that purpose, the experiment was preceded by a tutorial with drawings and a familiarization session in which the participants were explicitly asked to identify features of the sounds arranged in a play list. One way of increasing the consistency of the judgments would presumably be to conduct a more controlled familiarization or, alternatively, to extend the data collection by including additional sessions until response changes no longer occur. Future experiments should clarify how such an increased exposure to the sounds affects the judgments. Other possible applications of PSA are training and assessment of listening panels. Ideally expert listeners—such as those employed in descriptive analysis—have a clear idea of what to listen for beforehand, and PSA can prove very useful in assessing similarities and differences between individual perceptual structures.

3.3 Concluding Remarks

In search of a method to elicit auditory attributes in the context of multichannel reproduced sound, two fundamentally different approaches were investigated. The repertory grid technique proved a useful technique to elicit verbal descriptors. However, that these descriptors correspond to salient attributes that the subjects are able to consistently identify, can only be assumed. Perceptual structure analysis, on the other hand, verifies the consistency with which a subject can identify attributes in the sounds. In addition it makes it possible to uncover auditory features independently of the availability of a verbal label in the subject's lexicon. While RGT generally resulted in a higher number of constructs than PSA, the number of clusters obtained from cluster analysis was comparable to the number of PSA features. Finally, because of the low number of structures derived in PSA per program material, a conclusive comparison of the attributes obtained by the two elicitation methods cannot be drawn. It was possible, however, to derive a common set of eight attributes based on both methods.

4 ACKNOWLEDGMENT

This research was carried out as part of the "Centerkontrakt on Sound Quality," which establishes participation in and funding of the Sound Quality Research Unit (SQRU) at Aalborg University. The participating companies are Bang & Olufsen, Brüel & Kjær, and Delta Acoustics & Vibration. Further financial support came from the Ministry for Science, Technology, and Development (VTU) and from the Danish Research Council for Technology and Production (FTP). The authors would like to thank Wolfgang Ellermeier for helpful comments on an earlier version of the manuscript, and Jody Ghani for helping improve the English style. Furthermore, we are grateful to the two anonymous reviewers for their comments.

5 REFERENCES

[1] S. Bech, "Methods for Subjective Evaluation of Spatial Characteristics of Sound," in *Proc. 16th AES Int. Conf. on Spatial Sound Reproduction* (1999), pp. 487–504.

[2] J. Berg and F. Rumsey, "Identification of Quality Attributes of Spatial Audio by Repertory Grid Technique," *J. Audio Eng. Soc.,* vol. 54, pp. 365–379 (2006 May).

[3] C. Guastavino and B. F. G. Katz, "Perceptual Evaluation of Multi-Dimensional Spatial Audio Reproduction," *J. Acoust. Soc. Am.,* vol. 116, pp. 1105–1115 (2004).

[4] K. Koivuniemi and N. Zacharov, "Unravelling the Perception of Spatial Sound Reproduction: Language Development, Verbal Protocol Analysis, and Listener Training," presented at the 111th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 49, p. 1217 (2001 Dec.), convention paper 5424.

[5] R. Mason, N. Ford, F. Rumsey, and B. De Bruyn, "Verbal and Nonverbal Elicitation Techniques in the Subjective Assessment of Spatial Sound Reproduction," *J. Audio Eng. Soc. (Engineering Reports),* vol. 49, pp. 366–384 (2001 May).

[6] F. Rumsey, "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *J. Audio Eng. Soc.,* vol. 50, pp. 651–666 (2002 Sept.).

[7] H. Stone and J. L. Sidel, *Sensory Evaluation Practices* (Academic Press, London, 1993).

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

[9] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications* (Springer, Berlin, 1997).

[10] W. L. Martens and C. N. W. Giragama, "Relating Multilingual Semantic Scales to a Common Timbre Space," presented at the 113th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 50, p. 974 (2002 Nov.), convention paper 5705.

[11] J. Heller, "Representation and Assessment of Individual Semantic Knowledge," *Meth. Psychol. Res.,* vol. 5, pp. 1–37 (2000).

[12] J. P. Doignon and J. C. Falmagne, *Knowledge Spaces* (Springer, Berlin, 1999).

[13] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations* (Springer, Berlin, 1999).

[14] F. Wickelmaier and W. Ellermeier, "Deriving Auditory Features from Triadic Comparisons," *Percept. Psychophys.,* to be published.

[15] ITU-R Rec. BS.775-1, "Multichannel Stereophonic Sound System with and without Accompanying Picture," International Telecommunications Union, Geneva, Switzerland (1994).

[16] ITU-R Rec. BS.1116, "Methods for the Subjective Assessment of Small Impairments in Audio Systems In-

cluding Multichannel Sound Systems," International Telecommunications Union, Geneva, Switzerland (1997).

[17] AES TD1001.1.01-10, "Multichannel Surround Sound Systems and Operations," AES Technical Council, New York (2001).

[18] P. Suokuisma, N. Zacharov, and S. Bech, "Multichannel Level Alignment, Part I: Signals and Methods," presented at the 105th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 46, p. 1042 (1998 Nov.), preprint 4815.

[19] H. Levitt, "Transformed Up–Down Methods in Psychoacoustics," *J. Acoust. Soc. Am.,* vol. 49, pp. 467–477 (1971).

[20] W. Jesteadt, "An Adaptive Procedure for Subjective Judgments," *Percept. Psychophys.,* vol. 28, pp. 85–88 (1980).

[21] F. Wickelmaier and S. Choisel, "Selecting Participants for Listening Tests of Multichannel Reproduced Sound," presented at the 118th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 53, p. 703 (2005 July/Aug.), convention paper 6483.

[22] J. Berg, "How Do We Determine the Attribute Scales and Questions that We Should Ask of Subjects when Evaluating Spatial Audio Quality?" presented at the Workshop on Spatial Audio and Sensory Evaluation Techniques, Guilford, UK, 2006 April 6–7.

THE AUTHORS

S. Choisel

F. Wickelmaier

Sylvain Choisel received an M.Sc. degree in acoustics from Aalborg University, Denmark, in 2001, and a degree in electrical engineering with a major in signal processing and telecommunications from ESIEE, Paris.

He has worked on several projects in the field of signal processing, psychoacoustics, and computer simulation (BEM). Since 2002 February he has been working with Bang & Olufsen in the Sound Quality Research Unit at Aalborg University, from which he received a Ph.D. degree in 2005. His main research interest is perception of reproduced sound.

Dr. Choisel is a member of the Audio Engineering Society.

Florian Wickelmaier studied experimental psychology and statistics at the University of Regensburg, Germany, from which he received a master's degree in psychology in 2002. Later he joined the Sound Quality Research Unit (SQRU), headed by Wolfgang Ellermeier, at Aalborg University, Denmark, and became a Ph.D. student under his supervision. He was awarded a Ph.D. degree for his thesis on "Indirect Scaling Methods Applied to the Identification and Quantification of Auditory Attributes" in 2005.

Subsequently, Dr. Wickelmaier worked as an assistant professor in the Department of Acoustics, Aalborg University, Denmark. Since 2006 April he has been employed as a statistician in the Department of Psychiatry, University of Munich LMU, Germany.