Indirect scaling methods applied to the identification and quantification of auditory attributes

Ph.D. thesis by
Florian Wickelmaier

Indirect scaling methods applied to the identification and quantification of auditory attributes

Ph.D. thesis by
Florian Wickelmaier

Sound Quality Research Unit (SQRU)
Department of Acoustics, Aalborg University
Denmark

November 2005

Preface

This thesis is submitted to the Faculty of Engineering and Science at Aalborg University, Denmark, in partial fulfillment of the Ph.D. study programme. The work reported here was done at the Department of Acoustics, Aalborg University between May 2002 and August 2005.

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 comes from the Ministry for Science, Technology, and Development (VTU), and from the Danish Research Council for Technology and Production (FTP).

I would like to thank Professor Wolfgang Ellermeier for his guidance throughout the Ph.D. period, and for being a role model for me in any professional respect, be it as a researcher, as a scientific writer, or as a supervisor. I would also like to thank Sylvain Choisel for having the idea to combine our different backgrounds and to run a joint study. This led to a close and productive cooperation from which I benefited a lot, and which I truly enjoyed.

The efforts of Jody Ghani when proofreading manuscripts for mistakes in English, and of Søren Legarth when helping to translate the summary of this thesis into Danish are kindly acknowledged. For supporting my work in one way or another, I am grateful to Karin Zimmer, Christian Schmid, Søren Bech, Torben Holm Pedersen, Geoff Martin, to my colleagues at SQRU and at the Department of Acoustics, and to all the motivated participants in the listening experiments.

Finally, I am indebted to the members of the assessment committee Dorte Hammershøi (Aalborg Univerity), William L. Martens (McGill University), and Hans Irtel (Mannheim University) for their helpful comments on an earlier version of this thesis.

Florian Wickelmaier
November 2005, Aalborg

Table of contents

1 Summary

Auditory attributes, like for example loudness, pitch, sharpness, or tonal prominence, reflect how human listeners perceive their acoustical environment. The identification of which auditory attributes characterize the perceptual effects of a sound source and their quantification are therefore of major concern for different applications of sound quality research: be it in product-sound design or in the design of sound reproduction systems.

In this Ph.D. thesis, so-called *indirect* scaling methods were applied to the identification and quantification of auditory attributes. Such methods are characterized by requiring only simple qualitative judgments from listeners (as opposed to complex, often numerical judgments in direct scaling procedures), and by providing built-in tests of the validity of the underlying theoretical construct (as opposed to assuming its validity). In indirect scaling procedures, the numerical scale or the representation of the attributes are derived from modeling the listeners' judgments.

Specifically, the potential of probabilistic choice models to derive ratio-scale measures of auditory attributes from binary paired-comparison judgments was investigated. Previous studies have shown that simple choice models, e. g. the Bradley-Terry-Luce (BTL) model are too restrictive to account for judgments of complex, multidimensional stimuli. Therefore, software was developed which allows for parameter estimation and model testing of a more general class of probabilistic choice models.

Furthermore, a method for identifying auditory attributes, a feature-based representation of auditory stimuli was proposed, and tested experimentally. The method avoids confounding listeners' perceptual and verbal abilities, in that it strictly separates the process of identifying auditory features from labeling them. The approach was applied to simple synthetic sounds with well-defined physical properties (narrow-band noises and complex tones). For each stimulus triad, listeners had to judge whether the first two sounds displayed a common feature which was not shared by the third, by responding with a simple "Yes," or "No". Due to the high degree of consistency in the responses, feature structures could be derived for most of the 18 participants.

In a final sequence of three experiments, auditory feature structures and probabilistic choice models were applied to the identification and quantification of attributes of multichannel reproduced sound. A selection procedure was devised in order to select listeners for these experiments. 91 participants filled in a web-based questionnaire. 78 of them took part in an assessment of their hearing thresholds, their spatial hearing, and their verbal production abilities. The listeners displayed large individual differences in their performance. 40 subjects were selected based on the test results.

In the experiments proper, short musical excerpts were presented in mono, stereo and several multichannel formats to the panel of 40 selected listeners. The first experiment aimed at an assessment of the overall preference for the reproduction modes by means of paired comparisons and probabilistic choice models, and an exploratory analysis of the salient perceptual dimensions using multidimensional scaling. In the second experiment, individual auditory attributes were elicited and selected employing auditory feature structures and a direct elicitation method. In the third experiment, the selected attributes were quantified, and their contribution to overall preference was investigated. Scaling of preference and of the attributes was based on consistency tests of the paired-comparison judgments and on predicting the choice frequencies using probabilistic choice models. As a result, the preferences of non-expert listeners could be measured reliably on a high scale type. Principal components derived from the quantified attributes predict overall preference well. The findings allow for a careful generalization regarding the perception of spatial audio reproduction across musical program materials.

2 Resumé (Summary in Danish)

Auditive attributter, som for eksempel loudness, pitch, sharpness eller tonal prominence, afspejler hvordan mennesker oplever deres akustiske omgivelser. Identificeringen af hvilke auditive attributter der karakteriserer de perceptuelle effekter af en lydkilde og deres kvantificering er derfor af stor betydning for forskellige anvendelser indenfor lydkvalitets-forskning: det være i produktlyd design eller i design af reproduktionssystemer.

I denne Ph.D. afhandling blev såkaldt indirekte skaleringsmetoder anvendt til identificering og kvantificering af auditive attributter. Den type metoder er kendetegnet ved kun at kræve simple kvalitative bedømmelser fra forsøgspersoner (i modsætning til komplekse, ofte numeriske bedømmelser i direkte skaleringsprocedurer), og at de leverer "indbyggede" validitetstests af det underligende theoretiske konstrukt (i modsætning til at formode dets validitet). Ved indirekte skaleringsmetoder afledes den numeriske skala eller repræsentation af attributter udfra en modellering af forsøgspersonernes bedømmelser.

Specielt potentialet for at anvende "probabilistic choice models" til at aflede ratioskaler for auditive attributter fra binær parvise sammenligningsbedømmelser undersøges. Tidligere studier har vist, at simple "choice models", for eksempel Bradley-Terry-Luce modellen (BTL-model), er for restriktiv til at beskrive bedømmelser af komplekse, multidimensionale stimuli. Derfor blev noget software udviklet, som tillader parameterestimerig og modelprøvning af en mere generel klasse af "probabilistic choice models".

Endvidere blev en metode til identificering af auditive attributter, en egenskabsbaseret repræsentation af auditive stimuli, foreslået og afprøvet eksperimentelt. Det undgås med metoden, at perceptive og verbale evner hos personen forveksles, da den nøje adskiller processerne med at identificere og navngive auditive egenskaber. Metoden blev afprøvet med brug af simple syntetiske lyde med veldefinerede fysiske egenskaber (smalbåndet støj og komplekse toner). For hver sæt af tre stimuli skulle personen, ved at besvare med et simpelt ja eller nej, bedømme, om de to første lyde havde en egenskab til fælles, som ikke deltes med den tredje. Da bedømmelserne blev afgivet meget konsistent, kunne der – for de fleste personer – findes en struktur i egenskaberne.

I en afsluttende række af tre forsøg blev auditive egenskabsstrukturer og "probabilistic choice models" anvendt til identificering og kvantificering af attributter for mulitkanals-lyd. En udvælgelsesprocedure blev udviklet for at vælge personer til disse forsøg. 91 deltager udfyldte et web-baseret spørgeskema. 78 af disse deltog i en test af deres høretærskel, rumlige lydopfattelse og verbale udtryksevne. Forsøgspersonerne udviste store individuelle forskelligheder i deres præstationer. 40 forsøgspersoner blev udvalgt på baggrund af testresultaterne.

I det egentlige forsøg blev de 40 udvalgte deltagere præsenteret for korte musikeksempler i mono, stereo og forskellige multikanals-formater. Formålet med det første eksperiment var en bedømmelse af præference for reproduktionsformaterne ved hjælp af parvise sammenligninger og “probabilistic choice models”, samt en eksplorativ analyse af relevante opfattede dimensioner ved hjælp af multidimensional skalering. I andet forsøg blev individuelle auditive attributter fremkaldt og udvalgt ved hjælp af auditive egenskabsstrukturer og en direkte metode. I tredje forsøg blev de udvalgte attributter kvantificeret og deres bidrag til overordnet præference undersøgt. Skalering af præference og af attributterne blev baseret på konsistenstests af bedømmelserne fra de parvise sammenligninger og på prædiktion af valghyppigheder ved hjælp af “probabilistic choice models”. Som resultat kunne naive forsøgspersoners præference blive målt pålideligt på højt skalaniveau. De opnåde hovedkomponenter udledt fra de målte attributter kunne tilfredsstillende forudsige præferencen. Resultaterne tillader en forsigtig generalisering med hensyn til opfattelse af rumlige audio-reproduktion på tværs af musik materiale.

3 Overview of the thesis

3.1 Introduction

Both the identification of auditory attributes, and the measurement of sensation strength constitute ongoing challenges in psychoacoustics which have led to the emergence of a diversity of psychophysical methods. A particularly prominent part in sound quality assessments is played by so-called methods of *direct* scaling in which listeners are asked to directly provide an estimate of their sensation magnitude; this might be done by making a mark on a graphical (visual analogue) scale, or by judging the sounds on numerical or verbal rating scales. Direct scaling procedures have in common that they convert the listeners' judgments directly into numerical values, assuming the validity of the resulting scales.

Not only is the scaling of auditory attributes frequently carried out using direct methods, also the identification (or "elicitation") of relevant attributes often rests on procedures, which require complex judgments about sounds and typically an explicit labeling of the dimensions encountered. Such direct methods come with a number of disadvantages, in that they (1) demand a considerable amount of expertise and training and (2) involve the risk of biasing the judgments by a priori presenting verbal categories and therefore (3) prevent the detection of possibly unknown and unlabeled auditory attributes.

By contrast, *indirect* scaling methods require only simple qualitative judgments (e. g., which of two stimuli is greater than the other one with respect to a certain attribute), and the numerical representation depends on certain structural conditions, for example *transitivity* which postulates – in its simplest form – that whenever stimulus A is judged to dominate B, and B to dominate C, then A must also dominate C.

In this Ph.D. thesis indirect scaling methods are applied and analyzed systematically with respect to their usefulness for sound quality evaluation. It is the ultimate goal to derive meaningful representations of auditory attributes. It is, therefore, of particular interest to explore whether the structural conditions required for such representations hold for judgments of complex stimuli, as they are usually encountered in sound quality assessments. Two research questions will recur throughout the study:

(1) Are listener judgments structured enough to allow for a feature representation of auditory stimuli to be derived from simple "Yes"/"No" answers without requiring an explicit labeling of the features?

(2) Are listener judgments structured enough to allow for a numerical representation of the strength of an auditory attribute to be derived from simple binary paired comparisons without requiring a direct estimate of the sensation strength?

3.2 Methodological considerations in sound quality evaluation

3.2.1 Auditory attributes and their verbal descriptors

In psychoacoustic research, there is usually a strict distinction between the acoustic stimulus, which is described in physical terms, and its perceptual representation or *auditory event* (Blauert, 1997, chap. 1) which builds up inside a listener when acoustic waves impinge on his or her auditory system. An ongoing challenge in applied psychoacoustics constitutes the problem of how the auditory event can be characterized in terms of more elementary sensations, which have been referred to as sensory features (Nakayama, Miura, Kosaka, Okamoto, & Shiga, 1971), perceptual dimensions (Gabrielsson & Sjögren, 1979), hearing sensations (Zwicker & Fastl, 1999), or auditory attributes (e. g. Rumsey, 2002). In sound quality evaluation, seen as a discipline of applied psychoacoustics, one or several of the following questions are addressed:

(1) What auditory attributes are relevant in the context of the sounds under study?

(2) How can the auditory attributes be quantified?

(3) What is the relationship between preference for (or overall quality of) the sounds and the auditory attributes elicited by them?

(4) What is the relationship between physical characteristics of the sounds and the auditory attributes?

While there is agreement in the literature that the auditory attributes, i. e. the *sensations*, cannot directly be equated with their verbal labels or *descriptors* (e. g. Martens & Giragama, 2002), the methods for the elicitation of auditory attributes which are traditionally used in sound quality evaluation, like the *repertory grid technique* (Berg & Rumsey, 1999) or *descriptive analysis* (Zacharov & Koivuniemi, 2001) do not make this distinction very clear; in fact these methods rely on a close correspondence between auditory attributes and their verbal descriptors. In this thesis, a method is introduced which allows to identify auditory features independently of their verbal labeling by the subjects. The applicability and limitations of this method in the context of sound quality evaluation are investigated.

A second aim of this study is the investigation and application of methods for scaling or measurement of auditory attributes. The following section presents two different philosophical viewpoints on the problem of measurement which underlie the methodological considerations throughout this thesis.

3.2.2 Operational versus representational measurement

The question of how measurement can be defined, and consequently what conclusions about the status of a measured attribute can be drawn, has led to controversies in philosophy of science (cf. e. g. Narens & Luce, 1986; Hand, 1996). While this discussion and the associated development of a formal theory of measurement have passed the physical sciences largely unnoticed (cf. Falmagne, 1976), the possible implications for the behavioral sciences are considerable. This is hardly surprising, since physical measurement—unlike "subjective" measurement—is not as vulnerable to the skepticism towards whether or not measurement of a certain attribute is possible *in principle*.

Operationalism (Bridgman, 1927) defines an attribute by its measuring procedure, and the attribute has no "real" existence beyond that. Thus, the measurement problem is solved when a procedure is devised by which the attribute can be measured. In psychology, operational ideas are reflected in behaviorism (Skinner, 1976) where introspection as a method of gaining knowledge about psychological processes—such as, e. g., perception—was criticized. Rather, the basic datum is the response of a subject to a given stimulus; the underlying psychological processes are considered non-observable and inaccessible to introspection. Influenced by the operational view on measurement, Stevens (1975) proposed several procedures for the direct estimation of sensation magnitudes. According to an operational view, the loudness *sensation*, for example, is equated with the loudness *scale* as it results from a magnitude estimation experiment. In response to the viewpoint propagated by physicists (e. g. Campbell, 1928) that measurement of an attribute requires an interpretation of addition (which is hard to find for most psychological attributes), Stevens (1946) proposed that measurement can be performed on different *levels*, which led to the notion of scale types: nominal, ordinal, interval, and ratio scales.

The pragmatic approach of defining an attribute by specifying a procedure for its measurement has met with criticism. Narens (1996), for example, points out that scale properties, and thus validity, of scales derived from magnitude estimation procedures (Stevens, 1975) cannot be justified by the finding that such measures correlate with some other phenomena. In fact, an empirical evaluation of the assumptions implicit in Stevens' approach (Ellermeier & Faulhammer, 2000) shows them to not stand up to closer scrutiny. Irwin & Whitehead (1991) go as far as stating that direct methods, like category scaling, are of "unknown and unknowable validity" (p. 234). A mathematical framework in which the validity of an attribute scale can be investigated is presented by the *representational measurement theory* (*RMT*; Krantz, Luce, Suppes, & Tversky, 1971; Narens & Luce, 1986). RMT strictly separates the *objects* which need to be measured from the *numbers* which are assigned to them. The objects and the relations among them (like, for example, "is larger than", "is at least as heavy as", "is not louder than", etc.) form the *empirical relational structure*. The (real) numbers and the arithmetic relations and operations (like $>$, $\leq$, or $+$) form the *numerical relational structure*. A mapping from the empirical to the numerical relational structure which preserves the empirical relations such that they are represented by the numerical relations, is called a *homomorphic* mapping, or a *scale* in RMT.

In RMT, three types of problems are addressed; these are the problems of representation, of uniqueness, and of meaningfulness. The problem of *representation* asks the question of what structural properties the relations in the empirical relational structure must satisfy in order for the scale to exist. The representation problem is solved if these properties (or *axioms*) can be explicitly stated. Axioms may be empirically testable, or non-testable (technical). An example of a testable axiom, which underlies ordinal measurement, is *transitivity* which asserts for any three objects a, b, c that whenever $a \succ b$ and $b \succ c$ then also $a \succ c$ (the symbol $\succ$ denotes the empirical relation). In a paired-comparison experiment it can be tested if $\succ$ satisfies transitivity, and if this is the case (and other axioms hold in addition), an ordinal scale can be constructed.

If all axioms required by a certain measurement system are found to hold, usually more than just one homomorphism (mapping from empirical to numerical relational structure) exists. The *uniqueness* problem poses the question of how many such scales can be obtained. This problem is solved by stating the family of permissible transformations for a scale. For an ordinal scale, all strictly monotonic increasing functions are permissible transformations. By contrast, a ratio scale is unique up to multiplication by a positive constant (similarity transformations) and has, therefore, a higher degree of uniqueness. The problem of *meaningfulness*, finally, is concerned with which (numerical) statements have an empirical interpretation when using a scale: Meaningful are only those statements which remain true under all permissible transformations. Consequently, whether or not the expression "Sound A is twice as loud as sound B" is meaningful depends on the scale level of the loudness scale; only a loudness ratio-scale permits such a statement.

3.2.3 Potential methodological implications for sound quality research

Representational measurement theory has some practical consequences which might influence methodological considerations in sound quality research. According to Rumsey (2002), an important property of an auditory attribute is its *meaningfulness*. Here "meaningful", as in everyday language, might just denote that it "makes sense" to a listener. It is not so clear, however, how to arrive at a meaningful measurement of an auditory attribute using direct scaling methods. RMT gives a more rigorous definition of meaningfulness. If the scale type of a measured attribute can be justified by experimentally testing the axioms required for its measurement, it can also be shown which numerical statements remain true under all permissible transformations, and are thus meaningful.

Furthermore, from the attribute elicitation methods traditionally employed in sound quality research it is not immediately obvious, which aspects of the results have an empirical interpretation. For example, can an elicited descriptor be directly interpreted as an auditory attribute? On the other hand, a measurement-theoretically founded elicitation method, which will be introduced in this thesis, allows for an interpretation of the elicited constructs as auditory features, given the structural requirements have been shown to hold in the experimental data.

A second often postulated property of an auditory attribute (Rumsey, 2002) is its *unidimensionality*. Using direct scaling methods it is difficult to justify unidimensionality, worse yet, it remains unclear how a critical test of unidimensionality should be performed. By contrast, the transitivity axiom in RMT, specifies the structural restrictions the data must satisfy in order to guarantee that subjects are able to integrate multiple stimulus dimensions into a unidimensional sensation magnitude.

In summary, representational measurement theory might provide a framework for addressing methodological problems encountered in sound quality evaluation. It is, therefore, the objective of this study to investigate whether the structural requirements implied by measurement theory hold for judgments on complex sounds. In particular, two research questions will be investigated in this study:

(1) Traditional methods for the elicitation of auditory attributes (Berg & Rumsey, 1999; Zacharov & Koivuniemi, 2001) make the assumption of a close correspondence between the auditory attribute, that is the sensation, on the one hand, and its verbal descriptor on the other hand. In this study, an indirect elicitation method is presented which does not depend on the labeling of the encountered auditory features. It is investigated whether listener judgments are consistent to such an extent that auditory features can be derived from them.

(2) Traditional methods for scaling auditory attributes (e. g. Guski, 1997) do not question the possibility of obtaining an attribute scale by asking listeners to provide a direct estimate of sensation magnitude. Based on the principles of measurement theory, a more rigorous approach is pursued in this study: It is argued, that a numerical representation of auditory attributes can only be derived from highly consistent judgments. It is investigated whether such judgments can be observed in a sound quality evaluation setting.

3.3 Organization of the thesis

This Ph.D. thesis is organized in five manuscripts. Although each of them contains independent work, they are related to each other in the following way:

Manuscripts A and B focus on the development of methods for scaling and identifying attributes (or features), respectively, which are then applied to auditory attributes of multichannel reproduced sound in Manuscripts D and E. Manuscript C describes the subject selection procedure for the experiments reported in Manuscripts D and E.

Manuscript A:
Wickelmaier, F. & Schmid, C. (2004). A Matlab function to estimate choice model parameters from paired-comparison data. *Behavior Research Methods, Instruments, & Computers*, **36**, 29–40.

Manuscript B:
Wickelmaier, F. & Ellermeier, W. (2005). Deriving auditory features from triadic comparisons. Accepted for publication in *Perception & Psychophysics.*

Manuscript C:
Wickelmaier, F. & Choisel, S. (2005). Selecting participants for listening tests involving multichannel reproduced sound. Portions of this work were presented at the *118th Convention of the Audio Engineering Society*, May 28-31, Barcelona, Spain. Preprint 6483.

Manuscript D:
Choisel, S. & Wickelmaier, F. (2005). Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound. Portions of this work were presented at the *118th Convention of the Audio Engineering Society*, May 28-31, Barcelona, Spain. Preprint 6369.

Manuscript E:
Choisel, S. & Wickelmaier, F. (2005). Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference.

The manuscripts included in this thesis have been composed in collaboration with Christian Schmid, Wolfgang Ellermeier, and Sylvain Choisel. Manuscripts C, D, and E are also included in the thesis submitted by Sylvain Choisel entitled "Spatial aspects of sound quality – subjective assessment of sound reproduced by stereo and by multichannel systems".

Substantial interrelations between the manuscripts

As regards to contents, the thesis is divided into two major parts, which are each subdivided into a theoretical section on the development of an indirect scaling method, and a section where the method is applied to the evaluation of multichannel sound. Figure 1 displays the the substantive relation of the manuscripts schematically.

The first part is concerned with the problem of identifying auditory attributes. In the theoretical section (Manuscript B), an indirect elicitation method which separates the elicitation of auditory features from their labeling, and which is based on principles of representational measurement theory, is presented and illustrated in an experiment on auditory features of well-defined synthetic sounds. In the application section (Manuscript D), this method is employed in order to extract auditory features which are elicited by several audio reproduction formats (mono, stereo and multichannel formats). The substantive research question in the first part of the thesis is whether listener judgments obey the structural requirements that allow for a representation of the underlying auditory features.

The second part of the thesis discusses methodological issues related to quantifying auditory attributes and overall preference. In the theoretical section (Manuscript A), probabilistic choice models are presented as a method for scaling psychological attributes. It is argued that complex stimuli lead to choice behavior which requires more general models for an adequate representation. In the application section (Manuscript E), probabilistic choice models are applied to quantifying eight selected auditory attributes (based on the results of Part I) and overall preference for the audio reproduction formats. The substantive research question in this part of the thesis is whether listener judgments display structural requirements which allow for a meaningful numerical representation of auditory attributes of and overall preference for complex stimuli as usually encountered in sound quality evaluation.

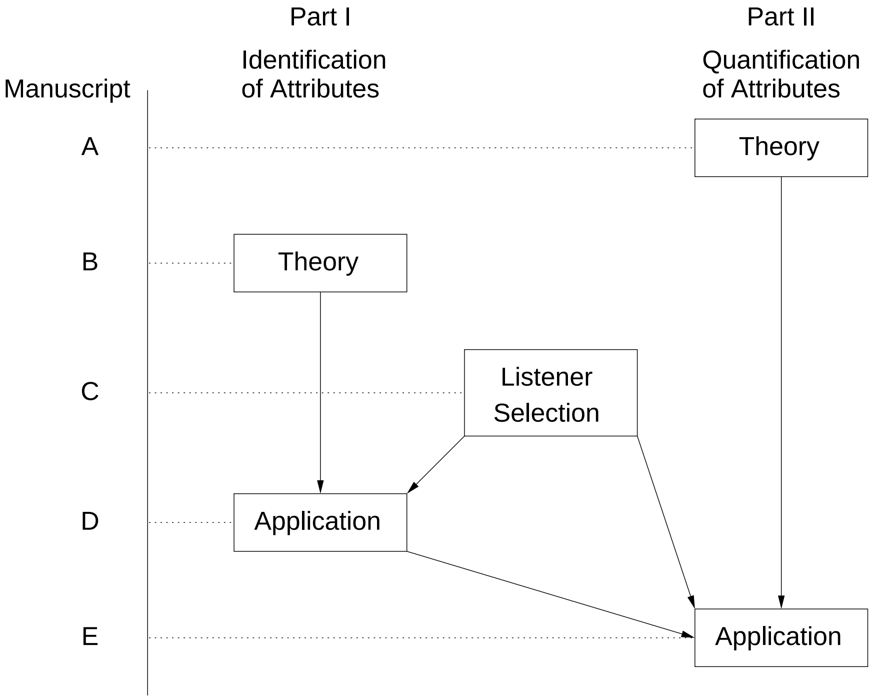


Figure 1: Schematic overview of the interrelations between the five manuscripts included in the thesis

Manuscript A: A Matlab function to estimate choice model parameters from paired-comparison data

A software algorithm is developed and described which allows for estimation and testing of a general class of probabilistic choice models, which are less restrictive than the Bradley-Terry-Luce (BTL) model; this means that those models require less restrictive forms of consistency in the judgments and are thus more realistic when attempting to quantify attributes of stimuli that are more multidimensional in nature. The manuscript further introduces the indirect scaling methods which were extensively employed for quantifying auditory attributes in the experiments described in Manuscript E. The statistical concepts of maximum likelihood estimation, likelihood ratio tests, and model selection are explained which are applied in Manuscript E for testing hypotheses about auditory attributes.

Manuscript B: Deriving auditory features from triadic comparisons

The manuscript describes an indirect method for the extraction of auditory features, so-called *auditory feature structures*. This method is *indirect*, because the features may be derived from qualitative triple-comparison judgments only (given sufficient consistency), and the labeling is optional. This is in contrast with direct elicitation methods, which require the participants to be able to directly provide verbal labels for each encountered attribute. Such a direct elicitation method is described in Manuscript D. The crucial difference between auditory feature structures and multidimensional scaling (MDS) is explained in the introduction of the paper (Manuscript B): While MDS can also be considered an indirect method (since a dimensional representation of the stimuli is obtained by modeling dissimilarity judgments), the conditions necessary for an MDS representation remain usually untested. By contrast, auditory feature structures can only be derived from *transitive* judgments. Therefore, the method allows the experimenter to test whether a listener can identify the auditory features consistently.

The manuscript describes the first application of this method to the domain of (auditory) perception. It is investigated whether the structural requirements implied by the method hold for judgments on simple acoustic stimuli. The results encourage a further application of the method in the context of sound quality evaluation, for the purpose of identifying auditory features of complex multichannel sounds (Manuscript D).

Manuscript C: Selecting participants for listening tests involving multichannel reproduced sound

The manuscript details the selection procedure employed in order to select the participants for the experiments reported in Manuscripts D and E. The procedure consisted of a questionnaire, pure-tone audiometry, a spatial hearing test, and a verbal fluency test. 40 subjects were selected based on their results. The questionnaire data did not predict the performance in the spatial hearing test or in the fluency test. This led to the conclusion that task-specific performance tests might be the preferable means of selecting a listening panel.

Manuscript D: Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound

Auditory feature structures as introduced in Manuscript B are applied to different audio reproduction formats (mono, stereo, and several multichannel formats) in an attempt to derive salient auditory features. Two additional proposals are made in order to make the method more suitable for the application to such complex stimuli. First, subjects were introduced to the task in a tutorial session using pictures having eye-catching visual features (see Appendix); this might help them to understand the requirements of the procedure before judging the sounds. Second, a technique is proposed to evaluate the transitivity violations encountered by means of simulating response patterns and comparing simulated and observed patterns. It is argued that a violation might have occurred at random (rather than systematically) if the discrepancy between simulated and observed responses is small.

In order to illustrate the distinction between direct and indirect methods for the identification of auditory attributes, the application of a well-known direct elicitation procedure in also reported in Manuscript D, and some of its (untested) assumptions are pointed out in the introduction of the paper.

Manuscript E: Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference

A selected list of eight auditory attributes which were elicited in the experiments described in Manuscript D are being quantified, using probabilistic choice models as introduced in Manuscript A. In addition, the overall preference for the reproduction modes is scaled. The results indicate that the listeners were highly consistent in their judgments. Consequently, overall preference and the more specific auditory attributes can be measured on a high scale type. Based on the scaling results, the contributions of the single attributes to overall quality are summarized in a statistical model. From this model, hypotheses of the influence of specific auditory sensations on the quality of multichannel sound can be derived.

3.4 General conclusions

3.4.1 Applicability and limitations of indirect scaling methods

This Ph.D. thesis is concerned with the application of indirect scaling methods to the identification and quantification of auditory attributes. It has been demonstrated that such methods do not only display theoretical advantages over direct scaling procedures, but that they are also applicable to “real-life” problems, such as the elicitation and scaling of attributes of multichannel reproduced sound. Before summarizing the specific findings about applicability and limitations of the indirect scaling methods investigated in this study, the following general statements can be made:

A potential disadvantage of the methods discussed is that they are time consuming. Since the data collection involves paired or triple comparisons, the number of stimuli which can be investigated in a single study has to be limited. For the assessment of a large variety of sounds there seems to exist no practical alternative to direct scaling procedures yet.

The advantages of the indirect methods, on the other hand, are that they

(1) rely on simple qualitative judgments,

(2) are in principle falsifiable,

(3) provide a test of the validity of the underlying theoretical constructs, and

(4) explicitly formulate a theory of the underlying psychological processes.

This makes them attractive alternatives to direct scaling procedures at least in basic-research studies of sound quality.

Auditory feature structures

The applicability of auditory feature structures to investigate simple sounds which elicit strongly salient auditory features was demonstrated in Manuscript B. The highly consistent judgments given by the subjects indicate that listeners have a well-defined representation of acoustic stimuli in terms of their auditory features. These auditory features are accessible in a listening experiment via simple qualitative judgments, and do not require an explicit labeling by the listener.

Also for complex stimuli, like multichannel reproduced sounds, the assessment of the underlying features by means of auditory structures is possible (Manuscript D). The degree of consistency with which features could be identified, however, was lower than for simple sounds, and was found to be insufficient for deriving a representation in about half of the sample. This indicates that the structural assumptions of the method might be too restrictive when differences between stimuli are subtle. It further suggests that attributes elicited using direct methods might not always have a clear perceptual basis.

In summary, application of auditory feature structures is recommended when

(1) a test of identifiability of the underlying features is needed,

(2) the risk of biasing the subjects by forcing them to introduce verbal descriptors for their sensations is high.

Limitations of auditory feature structures

The strongest limitation for a wide application of auditory feature structures in sound quality evaluation is arguably the required independence from the local context. This form of independence implies that the decision whether a feature is present or absent must not depend on the *local* context of a given stimulus triple, but has to be made based on the *global* context provided by all sounds under study. To relax this independence assumption constitutes a great challenge for further development of the method. In order to introduce subjects to the demands of the task, careful instruction is vital. A tutorial in form of pictures with strongly salient visual features (as used in Manuscript D, see Appendix) might be of help.

As other applications of representational measurement theory, auditory feature structures provide no probabilistic framework for the classification of the encountered transitivity violations as random errors or as systematic. In order to constitute a more realistic model of the underlying psychological processes, identification of auditory features should be formulated in a probabilistic manner. Ideally, the probability of overlooking a feature or of wrongly identifying one in the local context should be estimated from the data. A major challenge for the development of *probabilistic* auditory feature structures is answering the question of what restrictions the data need to satisfy in order for such probabilities to be estimable.

A further limitation originates from the lack of a statistical test of the validity of an auditory feature structure. As yet, the decision about the validity of a given structure has to be based on descriptive indices of reliability, consistency, and discrepancy between the data and that structure. A statistical test would greatly reduce uncertainty in this decision. Recent developments in the theory of knowledge spaces (Doignon & Falmagne, 1999), which is closely related to auditory feature structures, have shown that the introduction of probabilistic models also facilitates the formulation of statistical validity tests.

Paired-comparison scaling using probabilistic choice models

Many of the problems faced by auditory feature structures are already solved in case of probabilistic choice models. The "independence from irrelevant alternatives" (Luce, 1959) as required by the BTL model which in fact implies that the choices be made independently of the context provided by a given pair of stimuli, has been relaxed by less restrictive models, such as EBA (Tversky, 1972), and software for the estimation of these models has been made available (Manuscript A). Especially for complex multidimensional stimuli—as usually encountered in sound quality evaluation—context independence of choice behavior cannot be readily assumed (cf. the results on *envelopment* and *width* in Manuscript E, but also Zimmer, Ellermeier, & Schmid, 2004).

A further strength of these models is their conception of choice as a probabilistic process, which has a long tradition in psychophysics (Fechner, 1860; Thurstone, 1927) and stems from the basic observation that even under nearly identical stimulus conditions subjects show—both intra- and inter-individual—variation in their choice behavior. Momentary fluctuations in sensitivity, attention, or motivation are accounted for in a probabilistic framework.

Finally, estimation and testing in probabilistic choice models is based on standard maximum likelihood procedures. Therefore, statistical decision criteria (likelihood ratio tests, information criteria) can be employed to test the validity of the models, evaluate violations of the restrictions in the data (e. g. transitivity), and justify the scale type of the measured attribute. However, when relying on the statistical properties of the estimators and test statistics which often hold only asymptotically, it is the responsibility of the experimenter to collect data of a sufficient sample size.

In summary, application of probabilistic choice models is recommended when

(1) a test whether the subjects are able to map multiple stimulus dimensions onto a unidimensional sensation scale has to be performed,

(2) response biases resulting in different strategies of scale usage need to be eliminated.

Limitations of probabilistic choice models

A potential limitation of the application of probabilistic choice models in sound quality research comes up when scaling highly discriminable stimuli, the choices among which cannot be considered probabilistic, but rather deterministic. Such a situation might for example arise when scaling loudness of sounds having great level differences. When sounds having the higher SPL are *always* chosen over sounds having the lower SPL (and this happens to be for most stimuli under study), probabilistic choice models will not be adequate. Many, but certainly not all, such problems can be overcome by increasing the sample size: a choice *frequency* of zero for a given sound in a pair does not necessarily imply that the *probability* is zero; and it is conceivable that the sound might have been chosen in a larger sample.

3.4.2 Substantive findings regarding listener preference and auditory attributes

In general, the paired-comparison judgments collected in this study (and reported in Manuscript E) were found to be highly structured and consistent. Thus, it was possible to scale listener preferences for different audio reproduction formats using choice models which imply a very restrictive form of transitivity. This *strong* stochastic transitivity was observed to hold—independently of the program material—for judgments collected at two points in time. This indicates that listeners were able to integrate the multiple stimulus dimensions into a unidimensional preference scale. It further underlines that non-experts have a very stable concept of their listening preferences.

Not only the preference judgments, but also the judgments on the elicited auditory attributes were highly structured and were consequently measurable on high scale types. For one type of program material and two attributes (*envelopment* and *width*), however, systematic violations of the strong stochastic transitivity were encountered. This suggests that the choices were based on different aspects of the sounds depending on a given pair. Since this type of program material contained dry sources both in the front and in the surround channels, it may be described as having a *foreground-foreground* spatial characteristic (Zieliński, Rumsey, & Bech, 2003). By contrast, the other three program materials containing predominantly reverberation in the surround channels may be classified as having a *foreground-background* spatial characteristic. From the violations of the strong stochastic transitivity encountered in the present study, it is hypothesized that foreground-foreground material elicits more perceptually salient aspects which need to be integrated when judging upon an attribute. These results are also consistent with Rumsey's (2002) proposal of splitting up so-called "macro attributes" (like *envelopment* or *width*) into several "micro attributes" (like *individual source envelopment* and *ensemble envelopment*).

Zieliński et al. (2003) and Zieliński, Rumsey, Kassier, & Bech (2005) reported that the perceived quality of foreground-foreground material is more impaired by downmixing than is foreground-background material. The result of the present study that the stereo downmix was less preferred than the original *only* for the foreground-foreground material supports this hypothesis. The present study, however, suggests that the original is not always judged to be of highest quality, if the subjects are not explicitly instructed to assign the maximum rating to the original (as in, e. g., Zieliński et al., 2005).

Finally, the regression models established in Manuscript E were able to predict overall preference well from predictors based on the elicited auditory attributes. Letowski (1989) has proposed the idea that overall sound quality is comprised of timbral and spatial quality. Recently, Rumsey, Zieliński, Kassier, & Bech (2005) provided experimental evidence that global quality judgments can be predicted by judgments on timbral and spatial fidelity scales in the context of multichannel audio reproduction. In the present study, the elicited attributes were reduced to two principal components, which might be described as primarily spatial and timbral, respectively: *width*, *envelopment*, and *distance* loaded on one of the components, while *brightness* and *elevation* loaded on the other one. Only for the classical music, however, *clarity* could be uniquely assigned to the timbral component, whereas *spaciousness* loaded mainly on the spatial component only for the pop music. This suggests that there is some uncertainty when the classification of auditory attributes is based not on their verbal labels but on how listeners use them when judging sound quality. In general, however, the results from the present study provide support to the notion that both timbral and spatial auditory attributes are important predictors of overall listener preference.

References

Berg, J., & Rumsey, F. (1999). Identification of perceived spatial attributes of recordings by repertory grid technique and other methods. *106th Convention of the Audio Engineering Society, Munich, Germany, May 8–11.* Preprint 4924.

Blauert, J. (1997). *Spatial Hearing.* MIT Press, Cambridge, USA.

Bridgman, P. W. (1927). *The logic of modern physics.* New York: Macmillan.

Campbell, N. R. (1928). *An account of the principles of measurement and calculation.* London: Longmans, Green.

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces.* Berlin: Springer.

Ellermeier, W., & Faulhammer, M. (2000). Empirical evaluation of axioms fundamental to Stevens's ratio-scaling approach: I. Loudness production. *Perception & Psychophysics*, **62**, 1505–1511.

Falmagne, J.-C. (1976). Random conjoint measurement and loudness summation. *Psychological Review*, **83**, 65–79.

Fechner, G. T. (1860). *Elemente der Psychophysik.* Leipzig: Breitkopf und Härtel.

Gabrielsson, A., & Sjögren, H. (1979). Perceived sound quality of sound-reproducing systems. *Journal of the Acoustical Society of America*, **65**, 1019–1033.

Guski, R. (1997). Psychological methods for evaluating sound quality and assessing acoustic information. *Acta Acustica united with Acustica*, **83**, 765–774.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society: Series A*, **159**, 445–492.

Irwin, R. J., & Whitehead, P. R. (1991). Towards an objective psychophysics of pain. *Psychological Science*, **2**, 230–235.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement I.* New York: Academic Press.

Letowski, T. (1989). Sound quality assessment: Concepts and criteria. *87th Convention of the Audio Engineering Society, New York, USA, October 18–21.* Preprint 2825.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York: Wiley.

Martens, W. L., & Giragama, C. N. W. (2002). Relating multilingual semantic scales to a common timbre space. *113th Convention of the Audio Engineering Society, Los Angeles, USA, October 5–8.* Preprint 5705.

Nakayama, T., Miura, T., Kosaka, O., Okamoto, M., & Shiga, T. (1971). Subjective assessment of multichannel reproduction. *Journal of the Audio Engineering Society*, **19**, 744–751.

Narens, L. (1996). A theory of ratio magnitude estimation. *Journal of Mathematical Psychology*, **40**, 109–129.

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignment. *Psychological Bulletin*, **99**, 166–180.

Rumsey, F. (2002). Spatial quality evaluation for reproduced sound: terminology, meaning, and a scene-based paradigm. *Journal of the Audio Engineering Society*, **50**, 651–666.

Rumsey, F., Zieliński, S., Kassier, R., & Bech, S. (2005). On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *Journal of the Acoustical Society of America*, **118**, 968–976.

Skinner, B. F. (1976). *About behaviorism.* New York: Vintage books.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, **103**, 677–680.

Stevens, S. S. (1975). *Psychophysics. Introduction to its perceptual, neural, and social prospects.* New York: Wiley.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, **34**, 273–286.

Tversky, A. (1972). Elimination by aspects: a theory of choice. *Psychological Review*, **79**, 281–299.

Zacharov, N., & Koivuniemi, K. (2001). Audio descriptive analysis & mapping of spatial sound displays. In *Proceedings of the 2001 International Conference on Auditory Displays, Espoo, Finland*, (pp. 95–104).

Zieliński, S. K., Rumsey, F., & Bech, S. (2003). Effects of down-mix algorithms on quality of surround sound. *Journal of the Audio Engineering Society*, **51**, 780–798.

Zieliński, S. K., Rumsey, F., Kassier, R., & Bech, S. (2005). Comparison of basic audio quality and timbral and spatial fidelity changes caused by limitations of bandwidth and by down-mix algorithms in 5.1 surround audio systems. *Journal of the Audio Engineering Society*, **53**, 174–192.

Zimmer, K., Ellermeier, W., & Schmid, C. (2004). Using probabilistic choice models to investigate auditory unpleasantness. *Acta Acustica united with Acustica*, **90**, 1019–1028.

Zwicker, E., & Fastl, H. (1999). *Psychoacoustics. Facts and models.* Berlin: Springer.

Behavior Research Methods, Instruments, & Computers
2004, 36 (1), 29-40

A Matlab function to estimate choice model parameters from paired-comparison data

FLORIAN WICKELMAIER and CHRISTIAN SCHMID
Aalborg University, Aalborg East, Denmark

Tversky (1972) has proposed a family of models for paired-comparison data that generalize the Bradley– Terry–Luce (BTL) model and can, therefore, apply to a diversity of situations in which the BTL model is doomed to fail. In this article, we present a Matlab function that makes it easy to specify any of these general models (EBA, Pretree, or BTL) and to estimate their parameters. The program eliminates the time-consuming task of constructing the likelihood function by hand for every single model. The usage of the program is illustrated by several examples. Features of the algorithm are outlined. The purpose of this article is to facilitate the use of probabilistic choice models in the analysis of data resulting from paired comparisons.

In many empirical and experimental psychological studies, researchers rely on paired-comparison data in order to measure perceived magnitudes. Often, subjects are not capable of expressing their perceptions, impressions, or attitudes by means of an exact numerical value. On the other hand, a paired comparison between alternatives is generally feasible even if the stimuli involved are hard to distinguish or the dimension measured is not easy to comprehend. Making the task easier for the subject need not result in a measurement that lacks statistical power.

The first section of this article gives an outline of a family of probabilistic choice models as formulated by Bradley and Terry (1952), Luce (1959), Tversky (1972), and Tversky and Sattath (1979). These models are capable of deriving ratio scale measures of the stimuli from only binary judgments resulting from paired comparisons. Readers who are already familiar with these models are referred to the remainder of the article, where the Matlab function *OptiPt.m* is introduced, which allows for the straightforward fitting of the models.

Probabilistic Choice Models

A widely applied method of analyzing paired-comparison data is the BTL model as formulated by Bradley and Terry (1952) and Luce (1959). These authors showed that measurement on a ratio scale level can be established if the data satisfy certain structural assumptions. The probability of choosing x from a set of alternatives A can then be represented as

$$P(x, A) = \frac{u(x)}{\sum_{y \in A} u(y)}, \tag{1}$$

where u is a ratio scale. The BTL model leads to strong testable consequences. One of them is the independence of irrelevant alternatives. This implies that the probability ratio of choosing x from the set $\{x,y\}$ versus choosing y is not affected by any other alternative—for example, $\{x,y,z\}$. Formally,

$$\frac{P(x;\, y)}{P(y;\, x)} = \frac{P(x;\, y, z)}{P(y;\, x, z)} \tag{2}$$

is called the *constant ratio rule*. The constant ratio rule, however, is not likely to hold if the set of stimuli has some natural structure, as has been suggested by several counterexamples—for example, those from Debreu (1960) or Savage (see Luce & Suppes, 1965).

Consider the situation in which a subject is asked to choose between a trip to Florida (f) and two trips to California (c and c^+). (Rumelhart & Greeno, 1971, used a similar example to illustrate the shortcomings of the BTL model.) The two trips to California are identical except for a \$10 bonus for c^+. Suppose the subject is indifferent as to whether to travel to Florida or to California. Therefore, $P(f;c) = P(c;f) = .5$. Surely, he or she would prefer the bonus trip to California to the regular one; hence, $P(c^+;c)$ is close to one. Nevertheless, choosing between the three alternatives would presumably not result in a probability close to zero or one, which violates the constant ratio rule as formulated in Equation 2:

$$\frac{\overbrace{P\left(c;\, c^+\right)}^{\to 0}}{\underbrace{P\left(c^+;\, c\right)}_{\to 1}} \neq \frac{\overbrace{P\left(c;\, c^+, f\right)}^{>0}}{\underbrace{P\left(c^+;\, c, f\right)}_{<1}}. \tag{3}$$

This article was completed while the authors were employed at the Sound Quality Research Unit at Aalborg University. This unit receives financial support from Delta Acoustics & Vibration, Brüel & Kjær, and Bang & Olufsen, as well as from the Danish National Agency for Industry and Trade (EFS) and the Danish Technical Research Council (STVF). C.S. is presently at the Department of Experimental Psychology, University of Regensburg, Germany. The authors thank Sylvain Choisel for his contribution to the code of the Matlab function *OptiPt.m*. We also thank Karin Zimmer and Wolfgang Ellermeier for numerous helpful comments on earlier drafts of the article. We are grateful for the remarks of two anonymous reviewers and of the editor, which further improved the article. Correspondence concerning this article should be addressed to F. Wickelmaier, Department of Acoustics, Aalborg University, Fredrik Bajers Vej 7 B5, 9220 Aalborg East, Denmark (e-mail: fw@acoustics.dk).

Thus, stimulus similarity seems to be a major reason to abandon the BTL model.

Therefore, Tversky (1972) introduced a family of models that can cope with subgroups consisting of similar stimuli. He called the most general strategy, when choosing between alternatives, *elimination by aspects* (EBA). According to EBA, a subject prefers one stimulus over another because of a certain attribute this stimulus has that the other one does not have. Stimuli without this attribute are eliminated from the set of possible alternatives. If all the stimuli under consideration share the preferred attribute, it will be disregarded for the current decision. Thus, another discriminating attribute has to be found, and the elimination process restarts. In the example given above, EBA would predict that the choice between a trip to California and a trip to California plus a \$10 bonus would, in essence, be a choice between obtaining an extra \$10 or not.

EBA is a rather general approach for modeling paired-comparison data. It can be shown that the BTL model is a special case of the EBA model. If only one unique attribute characterizes each stimulus, EBA reduces to BTL.

Another special case of EBA describes a hierarchical decision strategy leading to so-called *preference tree*, or *Pretree*, models, as proposed by Tversky and Sattath (1979). According to Pretree, the attributes of the stimuli investigated are ordered in a hierarchical manner. Hence, the elimination process arrives at the final outcome much faster than it does with EBA. Consider the decision between several dishes at a restaurant (see Tversky & Sattath, 1979, for a similar example). These dishes may have hierarchically ordered attributes. They might, for example, fall into the two main categories of meat and fish. The meat category might again be divided into subcategories of, say, beef and other meats. Thus, a subject choosing the meat attribute could eliminate all alternatives of the fish category. If he or she chose the beef attribute, he or she could eliminate all nonbeef alternatives, and so on. The final outcome exhibits all of the desired attributes.

More formally, let $T = \{x,y,z, \ldots\}$ be the total finite set of alternatives or stimuli under study, and let A denote any nonempty subset of T. Furthermore, let $x' = \{\alpha,\beta,\gamma, \ldots\}$ be the set of attributes that characterizes the alternative x. Then, according to EBA, the probability of choosing x from A is

$$P(x, A) = \frac{\sum_{\alpha \in x' \setminus A^0} u(\alpha) P(x, A_\alpha)}{\sum_{\beta \in A' \setminus A^0} u(\beta)} \tag{4}$$

(cf. Tversky, 1972, Equation 6), where A^0 is the set of attributes shared by *every* alternative in A (formally, $A^0 = \cap_{x \in A} x'$), A' is the set of attributes that belongs to *at least one* alternative in A ($A' = \cup_{x \in A} x'$), and A_α is the set of all alternatives in A sharing the attribute α ($A_\alpha = \{x \in A : \alpha \in x'\}$). If only binary choice probabilities are analyzed as they result from paired-comparison data, the general Equation 4 simplifies to

$$P(x; y) = \frac{\sum_{\alpha \in x' \setminus y'} u(\alpha)}{\sum_{\alpha \in x' \setminus y'} u(\alpha) + \sum_{\beta \in y' \setminus x'} u(\beta)} \tag{5}$$

(cf. Tversky, 1972, Equation 7), where $x' \setminus y'$ is the set of attributes characterizing alternative x, but not alternative y. Note that the EBA model distinguishes between the scale values of the stimuli and the values of their attributes (which are the model parameters): The scale values are defined as the sum of the respective parameters. In the BTL model, there is only one parameter per stimulus. It is not hard to see that Equation 1 is a special case of Equation 4 and, in the case of binary data, of Equation 5 as well. Also, the probabilities of any Pretree model for paired-comparison data can be expressed by Equation 5, but then the attributes have to be hierarchically structured.

Parameter Estimation

In order to obtain maximum likelihood estimates (MLEs) of the model parameters, the likelihood function of the model has to be specified. Since a paired-comparison matrix of n stimuli can be perfectly described by $\binom{n}{2}$ binomially distributed random variables, the likelihood function takes the shape

$$L = \prod_{i<j} \pi_{ij}^{Nij} \left(1 - \pi_{ij}\right)^{Nji}, \tag{6}$$

where i and j are the row and column indices, respectively, of the data matrix and N_{ij} is the ijth element. In the EBA model, the probabilities π_{ij} are computed by Equation 5. The MLEs, $\hat{u}(\alpha), \hat{u}(\beta), \ldots$, are the values that maximize Equation 6. Most often, analytical solutions for maximizing the likelihood do not exist. Therefore, the MLEs have to be found by numerical optimization, using iterative methods.

The u parameters define a ratio scale on the set of attributes. Thus, one of them can be set to an arbitrary unit. Consequently the number of *free* parameters of an EBA model is always one less than the number of parameters in the likelihood function. This parameter surplus is not crucial when the MLEs are determined by numerical optimization. When statistical tests such as those presented below are employed, however, the number of free parameters has to be considered.

For the BTL model, Bradley (1955) has described how to estimate confidence intervals for the MLEs, but his approach can easily be generalized to apply to the whole EBA family. The Hessian matrix of the log-likelihood function is defined as the square matrix of second partial derivatives with respect to the model parameters:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 \log L}{\partial u_1^2} & \cdots & \frac{\partial^2 \log L}{\partial u_1 \partial u_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \log L}{\partial u_k \partial u_1} & \cdots & \frac{\partial^2 \log L}{\partial u_k^2} \end{bmatrix}, \tag{7}$$

where k is the number of parameters. Plugging the vector of MLEs $\hat{\mathbf{u}}$ into Equation 7 allows one to construct the matrix **C**, which is the inverse of the negative Hessian augmented by, respectively, a column and a row vector of ones and a zero in the bottom right corner:

$$\mathbf{C} = \begin{bmatrix} -\mathbf{H} & \mathbf{1} \\ \mathbf{1}' & 0 \end{bmatrix}^{-1}. \tag{8}$$

The first k rows and columns of **C** form the estimated covariance matrix of $\hat{\mathbf{u}}$. The variances, and thus the standard errors, of the MLEs can be estimated from the main diagonal of the covariance matrix. The 95% confidence intervals are obtained by

$$\hat{\mathbf{u}} \pm 1.96\sqrt{\mathbf{Diag}\left[\widehat{\mathbf{cov}}\left(\hat{\mathbf{u}}\right)\right]}. \tag{9}$$

Goodness of Fit

To check the goodness of fit of EBA models (including Pretree and BTL), it is convenient to compare the likelihood of the model with the saturated model that fits the data perfectly (cf. Wickens, 1982, chap. 6). In the saturated model, the probabilities π_{ij} are estimated by the relative frequencies, $\hat{\pi}_{ij} = N_{ij}/(N_{ij} + N_{ji})$. Therefore, the saturated model has $\binom{n}{2}$ free parameters. In the simplest case of the EBA model (i.e., BTL), the number of free parameters reduces to $n - 1$. Every additional parameter—for example, for a branch in a Pretree—has to be added, so in general the EBA model has $n - 1 + c$ free parameters. Note that the maximum number of Pretree parameters is $2n - 2$, whereas the maximum number of EBA parameters, $2^n - 2$, has to be reduced to a maximum of $\binom{n}{2}$ parameters if only binary data resulting from paired comparisons are available. For the statistical decision on whether or not the EBA model can account for the data, the likelihood ratio is computed. The expression

$$\chi^2 = -2\log\left[\frac{L_{\text{EBA}}}{L_{\text{SAT}}}\right] \tag{10}$$

is approximately χ^2-distributed, with $\binom{n}{2} - (n - 1 + c)$ degrees of freedom. The EBA model is rejected if the p value is less than 10% (rather than the conventional 5%, since one would like to increase the chance of detecting violations of the model).

A test statistic alternative to the one in Equation 10 is the common χ^2 test for goodness of fit:

$$\tilde{\chi}^2 = \sum_{i,j}\frac{\left(N_{ij} - \hat{\pi}_{ij}N\right)^2}{\hat{\pi}_{ij}N}, \tag{11}$$

where $N = N_{ij} + N_{ji}$. It accounts for the lack of fit between observed and predicted values. Equation 10 and Equation 11 are asymptotically equivalent—that is, $\chi^2 = \tilde{\chi}^2$ as $N \rightarrow \infty$.

Comparison of Models

In many applications, it is desirable to compare two EBA models directly. If the parameter space Ω' of one model is a proper subset of the other model's parameter space, Ω, the two models are nested. The restricted model EBA′ and the unrestricted model EBA can be tested against each other by using the likelihood ratio test of Equation 10:

$$\chi^2 = -2\log\left[\frac{L_{\text{EBA}'}}{L_{\text{EBA}}}\right] = 2\left[\log L_{\text{EBA}} - \log L_{\text{EBA}'}\right]. \tag{12}$$

The test statistic is approximately χ^2-distributed, having as many degrees of freedom as the difference between the numbers of parameters in EBA and in EBA′. The restricted model can be rejected if the likelihood ratio test is significant.

It is not in every case that two different EBA models are nested. The likelihood ratio test, however, is not appropriate if the two models are not nested. If this is the case, it is common practice to employ so-called information criteria as a tool for model selection. Akaike (1977) has provided a penalty function that takes into account both the likelihood of the model and the number of free parameters. For the EBA model, Akaike's information criterion (AIC) is defined as

$$\text{AIC} = -2\log L_{\text{EBA}} + 2(n - 1 + c). \tag{13}$$

When two models are compared using the AIC, the model with the smaller AIC should be selected.

The Function OptiPt.m

Pretree and EBA models can explain a great diversity of data sets for which the BTL model has to be rejected. Nevertheless, the BTL model is very popular among researchers, whereas the more general EBA and Pretree models are only rarely used. One of the main reasons appears to be that several computer programs are readily available for the estimation of BTL parameters, but not for EBA in general. Moreover, specifying the likelihood function of an EBA or Pretree model to a computer program is usually very cumbersome, and it takes considerable effort to write an estimation routine in a traditional programming language such as C, to handle more than one Pretree or EBA model.

The Matlab function OptiPt.m was written to overcome this unsatisfactory situation, since it draws heavily on the high-level computational commands lacking in many traditional programming languages. As we will show in the remainder of this article, it is capable of handling the whole EBA family (including the Pretree and BTL models) and is very easy to use and extremely flexible. Matlab is a popular mathematics and visualization software widely used by researchers in different fields to analyze data. OptiPt is designed for Matlab Version 6.0. It has been tested using Solaris and Windows versions of the Matlab software. No additional Matlab toolboxes are required to run the function. The source code of the function is given in the Appendix A. It may also be downloaded from www.acoustics.dk/~fw. Note: OptiPt is also available as an R function from www.acoustics.dk/~fw. R is a free statistical software package (see www.r-project.org).

Using OptiPt.m

To use the function OptiPt.m, copy it to your Matlab directory or to a directory in your Matlab path. For information on the usage of the function, type

```
>> help OptiPt,
```

and a short help message is displayed. (Note that >> is the Matlab prompt.) The complete syntax of the function is

```
>> [p,chistat,u,lL_eba,lL_sat,fit,cova] =
                                        OptiPt(M,A,s).
```

You call the function by its name, *OptiPt*. It expects two mandatory arguments *M* and *A*, and the optional argument *s*. *M* is a paired-comparison matrix, which is square, consisting of absolute frequencies and a main diagonal of zeros. The off-diagonal values should be greater than zero. *A* is a so-called cell array, an array that may consist of rows of different lengths. In *A*, the model will have to be specified, as we will show in the Examples section. Each row in *A* is a vector of all the parameters that belong to one alternative or stimulus. Thus, *A* has as many rows as there are alternatives or stimuli. The vector *s* carries the starting values for the estimation routine. If not specified, the search algorithm starts at $1/k$ for each parameter value, where k is the number of parameters.

The return values (*p*, *chistat*, *u*, *lL_eba*, *lL_sat*, *fit*, and *cova*) are optional. The parameter estimates are stored in the *p* vector. The *chistat* vector reports the χ^2 statistic according to Equation 10 as a measure of the goodness of fit of the model specified in *A*. In *chistat*, you also find the number of degrees of freedom needed for a statistical test. The *u* vector stores the scale values. Each scale value is the sum of the parameters that belong to one alternative or stimulus. Note that *p* and *u* are identical only for the BTL model, since there is only one unique parameter per stimulus. The log-likelihoods of the model specified in *A* and of the saturated model are stored in the *lL_eba* and *lL_sat* variables, respectively. The optimization algorithm searches for parameter values that maximize *lL_eba*. The saturated model, on the other hand, fits the data perfectly. Therefore, *lL_sat* is always greater than *lL_eba* for any model having fewer parameters. The *fit* matrix contains the predicted values of the fitted model. It can be used to calculate the goodness of fit according to Equation 11 and the residuals, in order to check for local misfits. The covariance matrix of the estimated parameters *p* is stored in the *cova* matrix. The diagonal of *cova* displays the variances of *p*. Take the square root of the variances in order to obtain the standard errors.

OptiPt automatically constructs the log-likelihood function of the model specified in *A* by first computing the set difference $x'\backslash y'$ for all pairs of stimuli. Then each factor of the likelihood function is calculated according to Equation 5. As a third step, rather than multiplying these factors as described in Equation 6, OptiPt takes the logarithm of each of them and sums them up, in order to compute the log-likelihood of the specified model. The optimization algorithm performs a direct search for the maximum likelihood estimates by calling the log-likelihood function with different parameter values. Thus, the maximum of the log-likelihood function is approached by numerical optimization without taking derivatives. OptiPt returns the parameter values that maximize the log-likelihood. In order to prevent the optimization algorithm from finding negative estimates, a constraint was built into the log-likelihood function: Whenever this function is called with at least one parameter less than or equal to zero, it returns minus infinity. The optimized parameters are then plugged into Equation 5 to yield the fitted paired-comparison matrix. Finally, a numerical Hessian matrix is computed and, according to Equation 8, augmented and inverted in order to estimate the covariance matrix of the parameters.

Examples

The following gives an extensive example of the usage of the Matlab function OptiPt.m. It is based on the real-world data set reported by Rumelhart and Greeno (1971). Their stimuli consisted of nine celebrities, including three politicians (Lyndon Baines Johnson, Harold Wilson, and Charles de Gaulle), three athletes (Johnny Unitas, Carl Yastrzemski, and A. J. Foyt), and three movie stars (Brigitte Bardot, Elizabeth Taylor, and Sophia Loren). They presented all 36 pairs of stimuli to 234 subjects, asking them with whom they would rather spend an hour of conversation. The results are summarized in Table 1. Suppose you have saved these results in a tab-delimited text file named *matrix.txt*. You can easily read this file into the Matlab matrix *M* by typing

```
>> M = textread('matrix.txt','','delimiter','\t').
```

We will first try to fit a BTL model to the data. Therefore, we are going to specify the model by means of the cell array *A* as follows:

```
>> A = {[1];[2];[3];[4];[5];[6];[7];[8];[9]}.
```

Note that every row in *A* corresponds to one stimulus. Note further that there is only one entry in each row, since every stimulus has only one unique parameter. To estimate the BTL parameters and test the goodness of fit, enter

```
>> [p,chistat] = OptiPt(M,A).
```

This will start the estimation routine. During the estimation process, you will receive feedback about the number of iterations and function calls needed to maximize the log-likelihood function. The message "optimization terminated successfully" indicates that at least a local extremum has been reached before the algorithm stopped searching. The parameter estimates are now stored in *p*, but beware of interpreting them without looking at the fit of the model.

You are going to find a large value of the test statistic [$\chi^2(28) = 78.22$, $p < .001$], indicating that the BTL model cannot account for the data. This is not surprising, since the three politicians, for example, are more similar

Table 1
Aggregate Choice Frequencies Reported by Rumelhart and Greeno (1971)

	L.B.J.	H.W.	C.d.G.	J.U.	C.Y.	A.J.F.	B.B.	E.T.	S.L.
L.B.J.	0	159	163	175	183	179	173	160	142
H.W.	75	0	138	164	172	160	156	122	122
C.d.G.	71	96	0	145	157	138	140	122	120
J.U.	59	70	89	0	176	115	124	86	61
C.Y.	51	62	77	58	0	77	95	72	61
A.J.F.	55	74	96	119	157	0	134	92	71
B.B.	61	78	94	110	139	100	0	67	48
E.T.	74	112	112	148	162	142	167	0	87
S.L.	92	112	114	173	173	163	186	147	0

Note—Row stimuli are chosen over column stimuli.

to each other than to any other celebrity. Typically, the BTL model does not hold for data sets in which subgroups of stimuli are formed on the basis of similarity. Note, however, that sufficient statistical power (provided in the present example by the large sample size of N = 234) is required to be able to reject any given model. A small number of subjects of, say, 20 considerably reduces the chance for any specified model to be rejected. In such a case, however, OptiPt gives no warning messages; it assumes that the sample size is large enough.

As a second example, we are going to fit a Pretree to the data. The list of stimuli naturally suggests a tree structure with three branches corresponding to the three different occupations of the nine celebrities: (L.B.J., H.W., C.d.G.) (J.U., C.Y., A.J.F.) (B.B., E.T., S.L.). Tversky and Sattath (1979) investigated this kind of Pretree, depicted schematically in Figure 1.

To specify the Pretree model, we simply expand the cell array A by the three branch parameters 10, 11, and 12:

```
>> B = {[1 10];[2 10];[3 10];[4 11];[5 11];[6 11];
        [7 12];[8 12];[9 12]}.
```

Now, every stimulus is characterized by two parameters, its individual parameter and its branch parameter. Hence, every row of B consists of two elements. In order to estimate the model parameters and check the goodness of fit, type

```
>> [p,chistat,u,lL_eba,lL_sat,fit,cova] =
                                   OptiPt(M,B)
```

to start the estimation procedure. The complete output resulting from this command is listed in Appendix B. Again, the message "optimization terminated successfully" should be displayed. Looking at the model fit, you will find that the tree model can account for the data quite well [$\chi^2(25) = 30.17$, $p = .22$]; the alternative goodness-of-fit statistic according to Equation 11 amounts to $\tilde{\chi}^2(25) = 30.05$. The model parameters are stored in the vector p. They differ from the scale values in u, since each scale value is the sum of the individual and branch parameters characterizing a given stimulus. You can arbitrarily choose any stimulus to define the unit length, because u is a ratio scale. The standardized u-scale with the first stimulus defining the unit is obtained by

```
>> u/u(1),
```

but of course, you can choose any other stimulus by changing the index of the denominator. The standard errors of the parameter estimates p can be extracted from the covariance matrix by typing

```
>> se = sqrt(diag(cova)).
```

To evaluate the program, it is desirable to compare the output of OptiPt with the results obtained by Tversky and Sattath (1979). Unfortunately, these authors reported neither the parameter values nor the scale values estimated by their fitting a Pretree model. Rather, they gave a graphical representation of these values by depicting the estimated Pretree. In their Figure 7 (p. 555), the lengths of the branches are proportional to the parameter values. (Note that this does not hold for the merely schematical Pretree in our present Figure 1, in which the lengths of the branches reveal no information about the parameter values.) In order to extract the parameter values from Tversky and Sattath's Figure 7, we measured the lengths of the branches of the tree with a ruler. Table 2 shows the results of this "measurement" in its third column. The parameter values are standardized, using the first value as the unit length. Obviously, measuring parameters with a ruler might introduce some error. Nevertheless, the standardized parameter estimates of OptiPt are quite close to the ruler-measured values, as may be seen in the forth column of Table 2. Thus, it seems appropriate to conclude that the results of Tversky and Sattath were replicated by

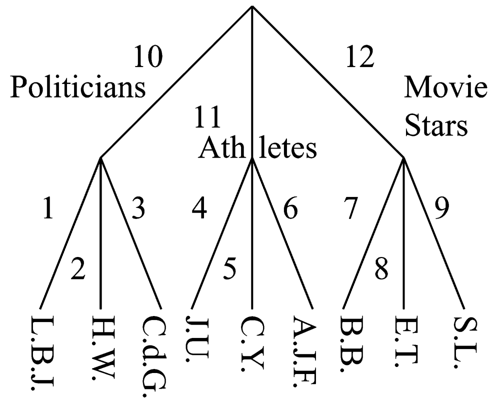


Figure 1. Schematical preference tree for choice among celebrities (Tversky & Sattath, 1979).

reanalyzing the data of Rumelhart and Greeno (1971) with the Matlab function OptiPt.m.

Finally, you may want to compare the BTL model and the Pretree model directly. Since the BTL model is nested into the Pretree, a likelihood ratio test can be employed. According to Equation 12, the test statistic is computed by

$$\chi^2 = -2\log\left[\frac{L_{\mathrm{BTL}}}{L_{\mathrm{PT}}}\right] = 2\left[\log L_{\mathrm{PT}} - \log L_{\mathrm{BTL}}\right], \qquad (14)$$

where L_{BTL} and L_{PT} are the likelihoods of the BTL model and the Pretree, respectively. This is easily achieved by Matlab and OptiPt. Just enter

```
>> [p,chistat,u,btl] = OptiPt(M,A);
>> [p,chistat,u,pt] = OptiPt(M,B);
>> 2*(pt-btl)
```

to receive the test statistic. Subtract the number of free BTL parameters (9 − 1) from the Pretree parameters (9 − 1 + 3) to obtain the degrees of freedom for the test. When the likelihood ratio test is applied to the present data, the BTL model can be rejected in favor of the Pretree [$\chi^2(3) = 48.05$, $p < .001$]. The AIC, as formulated in Equation 13, may alternatively be used for the purpose of model comparison:

```
>> aic_btl = -2*btl + 2*(9-1)
>> aic_pt = -2*pt + 2*(9-1 + 3).
```

The two AICs amount to $\mathrm{AIC}_{\mathrm{BTL}} = 10{,}716.4$ and $\mathrm{AIC}_{\mathrm{PT}} = 10{,}673.5$, respectively, and therefore clearly argue for the Pretree.

Avoiding Local Extrema

Generally, optimization algorithms guarantee only finding local extrema (i.e., locally optimal minima or maxima). The optimization algorithm used by OptiPt is the Matlab function *fminsearch*, which is part of the Matlab standard distribution. It employs the Nelder–Mead simplex method (Nelder & Mead, 1965), which is known to have rather good convergence properties if the parameter space is of low dimensionality (Lagarias, Reeds, Wright, & Wright, 1998, investigated the convergence in one and two dimensions). With increasing number of parameters, however, fminsearch is likely to report just a local solution. The OptiPt.m function provides one with three kinds of diagnostic information that could help you to avoid getting stuck in a local extremum.

Table 2
Comparison of the Results of
Tversky and Sattath (1979) and OptiPt

Parameter		Estimate		
Label	Number	Tversky and Sattath	OptiPt	*SE*
L.B.J.	1	1.00	1.0000	0.1116
H.W.	2	0.57	0.5416	0.0879
C.d.G.	3	0.40	0.3927	0.0735
J.U.	4	0.19	0.1803	0.0431
C.Y.	5	0.09	0.0729	0.0209
A.J.F.	6	0.19	0.1795	0.0454
B.B.	7	0.17	0.1641	0.0292
E.T.	8	0.43	0.4165	0.0538
S.L.	9	0.66	0.6401	0.0685
Politicians	10	0.31	0.3205	0.1300
Athletes	11	0.26	0.2450	0.0431
Movie stars	12	0.26	0.2549	0.0526

Note—The values in the third column were measured from Tversky and Sattath's Figure 7 (p. 555), using a ruler. The parameter estimates are standardized. The *SE* column shows the standard errors.

First, always make sure that the optimization algorithm stops with the message "optimization terminated successfully." Any other message, such as "maximum number of iterations has been exceeded," indicates that the algorithm stopped before reaching the supremum. In this case, the parameter values should not be interpreted. Second, EBA, Pretree, and BTL models are nested models if the parameter space of one model is a proper subset of the other model's parameter space. Thus, there is a natural order for the likelihood values of the models, determined by the number of dimensions of their parameter space. Note that the BTL model must have the lowest likelihood, since it has the lowest number of parameters. Every additional parameter must increase the likelihood. Furthermore, none of the models can have a likelihood that exceeds the likelihood of the saturated model. Any violation of this order indicates a local minimum. To check whether the likelihoods of your models are in the right order, make extensive use of the optional return values *lL_eba* and *lL_sat* of the OptiPt.m function. Since the log-function simply applies a monotonic transformation to the likelihoods, the log-likelihood of a Pretree model, for example, must always lie between the log-likelihoods of the BTL and the saturated model of the same data set. The third type of diagnostic information provided by OptiPt is the covariance matrix of the model parameters, contained in the optional return value *cova*. The main diagonal of the covariance matrix displays the variances of the MLEs. Enter

```
>> diag(cova)
```

to extract the variances from the *cova* matrix in Matlab. Any negative value in the main diagonal hints at an only locally optimal solution.

The general answer to the question of how to avoid local extrema is to optimize the starting values for the search algorithm. To achieve this with OptiPt, the optional function parameter *s* can be passed as an argument. Whenever one encounters a suspicious termination message or one finds the log-likelihoods not to be in the right order, the following method has proven successful: Run the program once, and take the results as starting values for the next run. This is easily done in Matlab by entering

```
>> [p,chistat] = OptiPt(M,A)
>> s = p;
>> [p,chistat] = OptiPt(M,A,s).
```

Note that the starting values have to be greater than zero. Hence, one may have to change elements of s if necessary. When a large number of stimuli and parameters are dealt with, it could also happen that this process of reestimating with optimized starting values has to be iterated several times until the log-likelihood of the specified model does not change any longer.

A second strategy to avoid local minima is to call OptiPt with randomly generated starting values. If this procedure is repeated, say, 10 times and the parameter estimates are the same, there is less doubt about a possible local extremum. The best approach to the problem is presumably to combine both methods: Run the function with different starting values, and plug in the estimates again.

Testing the Performance of OptiPt

The following section will report on a simulation study conducted to test the precision of the estimation algorithm. The general procedure thereby was as follows. First, the model structure needed to be specified. Second, the model parameters were chosen to have some arbitrary but fixed value. Third, a paired-comparison matrix was simulated by plugging the fixed parameters into the model's equations. Finally, the parameters were reestimated from the simulated paired-comparison matrix. The difference between the true parameter values and the estimates is an indicator of the precision of the estimation routine.

OptiPt calls the built-in Matlab fminsearch function to execute the search for the best parameters. Fminsearch evaluates the function to be minimized (in this case, the negative log-likelihood function) and tries different parameter configurations, starting from the initial values in the starting vector, in order to achieve a minimum function value. The search is successfully terminated if the return value of the minimized function and the optimized parameter values do not change by more than 0.0001 in two successive function calls. The stopping criteria are the default values in Matlab; they cannot be passed to OptiPt as optional arguments. Advanced users, however, can change the stopping rules, if they find it necessary, by editing the source code of OptiPt and by passing additional options to fminsearch by means of the *optimset* command. (Consult the Matlab help for detailed information on fminsearch and optimset.)

In the remainder of this section, we will report the method and the results of a simulation study, in which we tried to reestimate the parameters of three nested models in order to check the quality of the search algorithm. To illustrate the findings we encountered in our simulation studies, we will show a typical example of the results.

Consider a set of five stimuli $T = \{a,b,c,d,e\}$. Since we are dealing with paired-comparison data, we have $\binom{5}{2} =$ 10 independent data points. Hence, the total amount of $2^5 - 2 = 30$ parameters of a saturated EBA model (the number of proper nonempty subsets of T) has to be reduced to 10 parameters, in order for the model to be identifiable. We specified the structure of a 10-parameter EBA model, as is depicted in Figure 2; its parameters were randomly chosen from a uniform distribution ranging from 0 to 10. As a result, the parameters were set to

$$\begin{aligned} p_1 &= 1.1228 & p_6 &= 3.1357 \\ p_2 &= 2.8673 & p_7 &= 3.5723 \\ p_3 &= 9.6698 & p_8 &= 3.1550 \\ p_4 &= 2.3594 & p_9 &= 6.2415 \\ p_5 &= 3.3741 & p_{10} &= 6.0702. \end{aligned} \tag{15}$$

We inserted these values into Equation 5 to compute the predicted paired-comparison matrix, multiplying the relative frequencies by $N = 1{,}000$. In Matlab, the 10-parametric EBA model is specified by

```
>> EBA = {[1 6 7 9];[2 6 7 10];[3 7 9 10];
          [4 8];[5 8]}.
```

When calling OptiPt with the predicted matrix and the cell array EBA as arguments, we obtained the parameter estimates. Since there is no error in the data, however, we iterated the estimation procedure as described in the previous section until the χ^2 value was close to zero be-

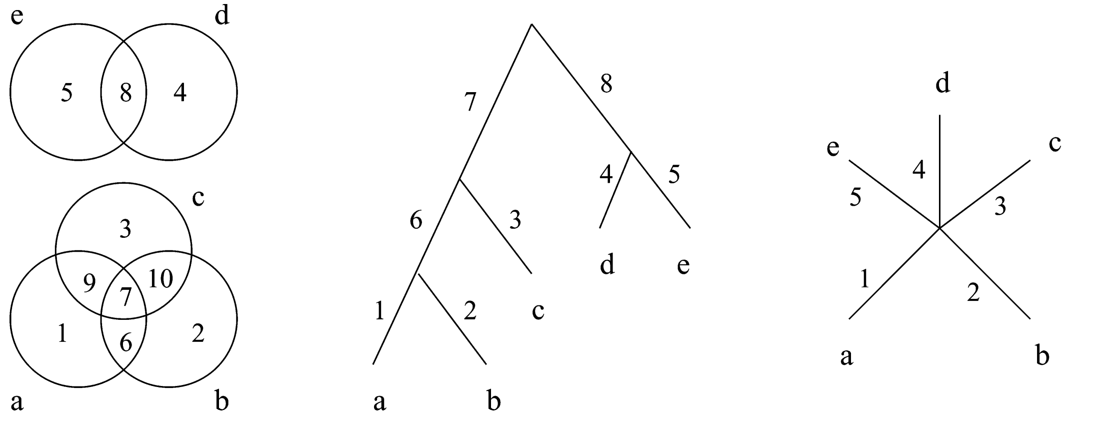


Figure 2. Three nested models: EBA, Pretree, and BTL models (from left to right).

fore we took the parameter estimates for granted. After a few iterations the estimates were

$$p = 0.0508\ 0.1492\ 0.5330\ 0.1331\ 0.1904\ 0.1895\ 0.1890\ 0.1780\ 0.3647\ 0.3551.$$

These estimates are unique up to multiplication by a positive constant, since the parameters are ratio scaled. To find the best constant c, we stored the true values of Equation 15 in the Matlab vector *true* and entered

```
>> c = mean(true./p).
```

Multiplying p by $c = 18.2298$ leads to

$$p = 0.9257\ 2.7201\ 9.7172\ 2.4269\ 3.4706\ 3.4545\ 3.4453\ 3.2451\ 6.6490\ 6.4727$$

as a good approximation of the true values in Equation 15, the maximal estimation error being 17.55%.

Setting p_9 and p_{10} to zero reduces the EBA model to the nested Pretree depicted in Figure 2. Again, we computed the predicted data matrix and passed it on to OptiPt, specifying the model by

```
>> PT = {[1 6 7];[2 6 7];[3 7];[4 8];[5 8]}.
```

Only one iteration was needed to reach a χ^2 value of 0.00. Multiplying the parameter estimates by $c = 15.7150$ resulted in

$$p = 1.1229\ 2.8674\ 9.6691\ 2.3598\ 3.3748\ 3.1353\ 3.5722\ 3.1542,$$

which is very close to the true values in Equation 15 (maximal estimation error: 0.25%). Finally, we turned the Pretree into a BTL model by setting the branch parameters p_6, p_7, and p_8 to zero. We computed the paired-comparison data and specified the model by

```
>> BTL = {[1];[2];[3];[4];[5]}.
```

Again, OptiPt provided the parameters after the first call. After multiplication by $c = 15.4007$, the parameter estimates were almost perfect copies of the true values (maximal estimation error: 0.07%):

$$p = 1.1228\ 2.8675\ 9.6692\ 2.3594\ 3.3743.$$

Two conclusions can be drawn from this simulation study. First, the precision of the estimation algorithm proved to be very satisfactory. Second, models with fewer than $\binom{n}{2}$ free parameters may be estimated from paired-comparison data, but the more parameters there are, the harder it becomes for the search algorithm to find the global extremum in a single run. As the number of the parameters increases, the precision of the search algorithm decreases slightly because of the enlarged search space. The results are summarized in Table 3.

Conclusions

In many empirical studies, paired-comparison data are collected in order to measure subjective magnitudes on scales resulting from the BTL model. Stimulus similarity, however, often causes the BTL model to be rejected. Therefore, Tversky (1972) proposed a family of models, the EBA models, that can handle stimulus subgrouping due to similarity. Use of these models in applied research, however, has been restricted by a lack of adequate software permitting flexible model specification and stringent testing.

This article introduced the new Matlab function OptiPt.m for the estimation of EBA, Pretree, and BTL model parameters. Its usage was illustrated by a classical example from the literature. Detailed instructions were given on how to apply the function effectively. The precision of the estimation algorithm has been shown to be very satisfactory. It is the authors' hope that this article will encourage other researchers to rediscover Tversky's EBA models and to use them as widely as the BTL model.

Table 3
Exemplary Results of the Simulation Study

Parameter	True Value	EBA	Pretree	BTL
p_1	1.1228	0.9257	1.1229	1.1228
p_2	2.8673	2.7201	2.8674	2.8675
p_3	9.6698	9.7172	9.6691	9.6692
p_4	2.3594	2.4269	2.3598	2.3594
p_5	3.3741	3.4706	3.3748	3.3743
p_6	3.1357	3.4545	3.1353	–
p_7	3.5723	3.4453	3.5722	–
p_8	3.1550	3.2451	3.1542	–
p_9	6.2415	6.6490	–	–
p_{10}	6.0702	6.4727	–	–

Note—The estimates are close to the true values. The accuracy increases as the number of parameters decreases.

REFERENCES

Akaike, H. (1977). On entropy maximization principle. In P. R. Krishnaiah (Ed.), *Applications of statistics* (pp. 27-41). Amsterdam: North-Holland.

Bradley, R. A. (1955). Rank analysis of incomplete block designs: III. Some large-sample results on estimation and power for a method of paired comparisons. *Biometrika*, **42**, 450-470.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, **39**, 324-345.

Debreu, G. (1960). Review of R. D. Luce's Individual choice behavior: A theoretical analysis. *American Economic Review*, **50**, 186-188.

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on Optimization*, **9**, 112-147.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Luce, R. D., & Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 3, pp. 249-410). New York: Wiley.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, **7**, 308-313.

Rumelhart, D. L., & Greeno, J. G. (1971). Similarity between stimuli: An experimental test of the Luce and Restle choice models. *Journal of Mathematical Psychology*, **8**, 370-381.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, **79**, 281-299.

Tversky, A., & Sattath, S. (1979). Preference trees. *Psychological Review*, **86**, 542-573.

Wickens, T. D. (1982). *Models for behavior: Stochastic processes in psychology*. San Francisco: Freeman.

APPENDIX A
Source Code of OptiPt.m

```
function [p,chistat,u,lL_eba,lL_sat,fit,cova] = OptiPt(M,A,s)
% OptiPt parameter estimation for BTL/Pretree/EBA models
%   p = OptiPt(M,A) estimates the parameters of a model specified
%   in A for the paired-comparison matrix M. M is a matrix with
%   absolute frequencies. A is a cell array.
%
%   [p,chistat,u] = OptiPt(M,A) estimates parameters and reports
%   the chi2 statistic as a measure of goodness of fit. The vector
%   of scale values is stored in u.
%
%   [p,chistat,u,lL_eba,lL_sat,fit,cova] = OptiPt(M,A,s) estimates
%   parameters, checks the goodness of fit, computes the scale values,
%   reports the log-likelihoods of the model specified in A and of the
%   saturated model, returns the fitted values and the covariance
%   matrix of the parameter estimates. If defined, s is the starting
%   vector for the estimation procedure. Otherwise each starting value
%   is set to 1/length(p).
%   The minimization algorithm used is FMINSEARCH.
%
%   Examples
%     Given the matrix M =
%                           0    36    35    44    25
%                          19     0    31    37    20
%                          20    24     0    46    24
%                          11    18     9     0    13
%                          30    35    31    42     0
%
%     A BTL model is specified by A = {[1];[2];[3];[4];[5]}
%     Parameter estimates and the chi2 statistic are obtained by
%       [p,chistat] = OptiPt(M,A)
%
%     A Pretree model is specified by A = {[1 6];[2 6];[3 7];[4 7];[5]}
%     A starting vector is defined by s = [2 2 3 4 4 .5 .5]
%     Parameter estimates, the chi2 statistic, the scale values, the
%     log-likelihoods of the Pretree model and of the saturated model,
%     the fitted values, and the covariance matrix are obtained by
%       [p,chistat,u,lL_eba,lL_sat,fit,cova] = OptiPt(M,A,s)
%
% Authors: Florian Wickelmaier (wickelmaier@web.de) and Sylvain Choisel
% Last mod: 03/JUL/2003

I = length(M);  % number of stimuli
mmm = 0;
for i = 1:I
  mmm = [mmm max(A{i})];
end
J = max(mmm);  % number of pt parameters
if(nargin == 2)
  p = ones(1,J)*(1/J);  % starting values
elseif(nargin == 3)
  p = s;
end

for i = 1:I
  for j = 1:I
    diff{i,j} = setdiff(A{i},A{j});  % set difference
  end
end

p = fminsearch(@ebalik,p,optimset(‘Display’,’iter’,’MaxFunEvals’,10000,...
    ‘MaxIter’,10000),M,diff,I);  % optimized parameters
lL_eba = -ebalik(p,M,diff,I);  % likelihood of the specified model

lL_sat = 0;  % likelihood of the saturated model
for i = 1:I-1
  for j = i+1:I
    lL_sat = lL_sat + M(i,j)*log(M(i,j)/(M(i,j)+M(j,i)))...
                   + M(j,i)*log(M(j,i)/(M(i,j)+M(j,i)));
  end
end
```

APPENDIX A (Continued)

```
fit = zeros(I);  % fitted PCM
for i = 1:I-1
  for j = i+1:I
    fit(i,j) = (M(i,j)+M(j,i))/(1+sum(p(diff{j,i}))/sum(p(diff{i,j})));
    fit(j,i) = (M(i,j)+M(j,i))/(1+sum(p(diff{i,j}))/sum(p(diff{j,i})));
  end
end

chi = 2*(lL_sat-lL_eba);
df =  I*(I-1)/2 - (J-1);
chistat = [chi df];  % 1-chi2cdf(chi,df)];  % goodness-of-fit statistic

u = sum(p(A{1}));  % scale values
for i = 2:I
  u = [u sum(p(A{i}))];
end

H = hessian('ebalik',p',M,diff,I);
C = inv([H ones(J,1); ones(1,J) 0]);
cova = C(1:J,1:J);

function lL_eba = ebalik(p,M,diff,I)  % computes the likelihood

if min(p)<=0  % bound search space
  lL_eba = inf;
  return
end

thesum = 0;
for i = 1:I-1
  for j = i+1:I
    thesum = thesum + M(i,j)*log(1+sum(p(diff{j,i}))/sum(p(diff{i,j})))...
                    + M(j,i)*log(1+sum(p(diff{i,j}))/sum(p(diff{j,i})));
  end
end
lL_eba = thesum;

function H = hessian(f,x,varargin)  % computes numerical Hessian

k = size(x,1);
fx = feval(f,x,varargin{:});
h = eps.^(1/3)*max(abs(x),1e-2);
xh = x+h;
h = xh-x;
ee = sparse(1:k,1:k,h,k,k);

g = zeros(k,1);
for i = 1:k
  g(i) = feval(f,x+ee(:,i),varargin{:});
end

H = h*h';
for i = 1:k
  for j = i:k
    H(i,j) = (feval(f,x+ee(:,i)+ee(:,j),varargin{:})-g(i)-g(j)+fx)...
                 / H(i,j);
    H(j,i) = H(i,j);
  end
end
```

APPENDIX B
Sample Output of OptiPt.m (See the Examples Section for Details)

```
Optimization terminated successfully:
 the current x satisfies the termination criteria using OPTIONS.TolX of 1.000000e-04
 and F(X) satisfies the convergence criteria using OPTIONS.TolFun of 1.000000e-04


p =

  Columns 1 through 7

    0.2860    0.1549    0.1123    0.0516    0.0209    0.0513    0.0469

  Columns 8 through 12

    0.1191    0.1831    0.0917    0.0701    0.0729


chistat =

   30.1663   25.0000


u =

  Columns 1 through 7

    0.3777    0.2466    0.2040    0.1217    0.0909    0.1214    0.1198

  Columns 8 through 9

    0.1920    0.2560


lL_eba =

  -5.3257e+03


lL_sat =

  -5.3106e+03


fit =

  Columns 1 through 7

         0  151.7911  168.0190  176.9900  188.5925  177.0748  177.6394
   82.2089         0  135.6456  156.6913  170.9534  156.7932  157.4720
   65.9810   98.3544         0  146.5816  161.8499  146.6893  147.4079
   57.0100   77.3087   87.4184         0  166.6236  117.2719  117.8844
   45.4075   63.0466   72.1501   67.3764         0   67.5996  100.9599
   56.9252   77.2068   87.3107  116.7281  166.4004         0  117.7693
   56.3606   76.5280   86.5921  116.1156  133.0401  116.2307         0
   78.8693  102.4461  113.4627  143.2449  158.7971  143.3542  167.8648
   94.5307  119.1920  130.2268  158.6188  172.6645  158.7194  186.2584

  Columns 8 through 9

  155.1307  139.4693
  131.5539  114.8080
  120.5373  103.7732
   90.7551   75.3812
   75.2029   61.3355
   90.6458   75.2806
   66.1352   47.7416
         0   92.2326
  141.7674         0
```

APPENDIX B (Continued)

```
cova =

  Columns 1 through 7

    0.0010    0.0007    0.0006   -0.0001   -0.0001   -0.0001   -0.0001
    0.0007    0.0006    0.0005   -0.0001   -0.0000   -0.0001   -0.0001
    0.0006    0.0005    0.0004   -0.0001   -0.0000   -0.0001   -0.0001
   -0.0001   -0.0001   -0.0001    0.0002    0.0001    0.0001   -0.0000
   -0.0001   -0.0000   -0.0000    0.0001    0.0000    0.0001   -0.0000
   -0.0001   -0.0001   -0.0001    0.0001    0.0001    0.0002   -0.0000
   -0.0001   -0.0001   -0.0001   -0.0000   -0.0000   -0.0000    0.0001
   -0.0003   -0.0002   -0.0002   -0.0000   -0.0000   -0.0000    0.0001
   -0.0004   -0.0003   -0.0002   -0.0000   -0.0000   -0.0000    0.0001
   -0.0010   -0.0008   -0.0007    0.0001    0.0000    0.0001    0.0001
   -0.0001   -0.0001   -0.0001   -0.0001   -0.0001   -0.0001    0.0000
   -0.0000   -0.0000   -0.0000    0.0000    0.0000    0.0000   -0.0001

  Columns 8 through 12

   -0.0003   -0.0004   -0.0010   -0.0001   -0.0000
   -0.0002   -0.0003   -0.0008   -0.0001   -0.0000
   -0.0002   -0.0002   -0.0007   -0.0001   -0.0000
   -0.0000   -0.0000    0.0001   -0.0001    0.0000
   -0.0000   -0.0000    0.0000   -0.0001    0.0000
   -0.0000   -0.0000    0.0001   -0.0001    0.0000
    0.0001    0.0001    0.0001    0.0000   -0.0001
    0.0002    0.0002    0.0002    0.0000   -0.0001
    0.0002    0.0004    0.0003    0.0000   -0.0001
    0.0002    0.0003    0.0014    0.0002    0.0002
    0.0000    0.0000    0.0002    0.0002    0.0001
   -0.0001   -0.0001    0.0002    0.0001    0.0002
```

(Manuscript received September 19, 2002;
revision accepted for publication September 9, 2003.)

Accepted for publication in
Perception & Psychophysics

Deriving auditory features from triadic comparisons

Florian Wickelmaier Wolfgang Ellermeier

Sound Quality Research Unit (SQRU)
Department of Acoustics, Aalborg University, Denmark

A feature-based representation of auditory stimuli is proposed, and tested experimentally. Within a measurement-theoretical framework it can be decided whether a representation of subjective judgments by a set of auditory features is possible, and how unique such a representation is. Further, the method avoids confounding listeners' perceptual and verbal abilities, in that it strictly separates the process of identifying auditory features from labeling them. The approach was applied to simple synthetic sounds with well-defined physical properties (narrow-band noises and complex tones). For each stimulus triad, listeners had to judge whether the first two sounds displayed a common feature which was not shared by the third, by responding with a simple "Yes," or "No". Due to the high degree of consistency in the responses, feature structures could be obtained for most of the participants. In summary, the proposed procedure constitutes a supplement to the arsenal of psychometric methods where the main focus is on identifying the type of sensation itself, rather than measuring its threshold or magnitude.

One of the perennial unresolved problems in psychoacoustics is to find out which auditory sensations are elicited by complex acoustic stimuli. Psychometric methods which aim at revealing such underlying sensations have mostly focused on metric and dimensional representations of some measure of the psychological proximity of the stimuli. Most notably, various versions of *multidimensional scaling* (MDS) have been developed (e. g. Carrol & Chang, 1970; Kruskal, 1964; Shepard, 1962; Torgerson, 1952; see Borg & Groenen, 1997, for an introduction and overview) and applied to uncover dimensions of auditory perception (e. g. Grey, 1977; Iverson & Krumhansl, 1993; Lakatos, McAdams, & Caussé, 1997). MDS seeks to represent the stimuli under study in some multidimensional space, such that the metric distances in that space correspond to the psychological proximities.

Beals, Krantz, & Tversky (1968) have studied MDS from the viewpoint of representational measurement theory and formulated qualitative properties that the proximities must satisfy in order to be representable as metric distances. Tversky (1977) has criticized metric and dimensional representations in general, and demonstrated that observed proximities often systematically violate the metric conditions inherent in geometrical models. Instead, he has proposed a feature-based representation, the so-called *contrast model*, which is able to explain many of the empirical findings. Formally, the contrast model predicts the similarity S between two stimuli a and b by

$$S(a,b) = \vartheta f(A \cap B) - \alpha f(A \setminus B) - \beta f(B \setminus A) \quad (1)$$

(cf. Tversky, 1977, p. 332), where S and f are interval scales, $A \cap B$ denotes the features that are common to both a and b, $A \setminus B$ the features that belong to a only, and $B \setminus A$ the features that belong to b only; the parameters ϑ, α, and β are nonnegative weighting factors. The contrast model expresses similarity between stimuli as a weighted function of their common and distinctive features. The main limitation of the contrast model as a method for revealing salient features results from the fact that the features have to be explicitly specified in order for the model to be testable. Thus, from similarity data alone, the characterizing features cannot uniquely be identified. Sattath & Tversky (1987) provided further evidence for this lack of uniqueness inherent in the contrast model.

Heller (2000) concluded from the unsolved uniqueness problem of the contrast model that similarity data generally do not provide enough information to derive the characterizing feature structure. To overcome this problem he introduced a theory of semantic features and an experimental paradigm for their assessment which is closely related to both knowledge space theory (Doignon & Falmagne, 1999; Falmagne, Koppen, Villano, Doignon, & Johannesen, 1990) and formal concept analysis (Ganter & Wille, 1999; Wille, 1982). For these so-called *semantic structures* he formulated both representation and uniqueness theorems in the sense of representational measurement theory (Krantz, Luce, Suppes, & Tversky, 1971). Thus, the feature representation rests on qualitative, experimentally-testable conditions, and its uniqueness can be stated explicitly and is to be determined empirically.

In this paper, Heller's (2000) semantic structures are applied to derive auditory features. In the following section, the theoretical notions needed to characterize *auditory feature structures* are briefly introduced which are in close correspondence with semantic structures. Subsequently, an experiment is reported which has been designed to test the proposed approach for revealing auditory features elicited by simple synthetic sounds.

STRUCTURES OF AUDITORY FEATURES

Let X denote the total finite 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 Heller (2000), $\langle X, \sigma \rangle$ is called an *auditory (feature) structure*. Let further $A \subseteq X$ denote a subset of X, and $\sigma(A)$ the intersection of all sets in σ of which A is a subset

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

which means that $\sigma(A)$ is the smallest set in σ which includes the sounds in A. Then a relation $\mathcal{Q}$ which 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 $A\mathcal{Q}x$, if and only if the subject answers "No" to the question:

> Do the sounds in A have something in common which makes them different from x?

If the answer to that question is "Yes," the relation between A and x does not hold which is denoted by $A\overline{\mathcal{Q}}x$. Further, $\mathcal{Q}$ is called *transitive* if

$$A\mathcal{Q}b \, (\forall b \in B) \text{ and } B\mathcal{Q}c \Rightarrow A\mathcal{Q}c \qquad (2)$$

for all $A, B \subseteq X$ and $c \in X$. In the present application, it is always assumed that $\mathcal{Q}$ is *reflexive*, meaning that $a \in A$ implies $A\mathcal{Q}a$. Thus, only questions where $x \notin A$ are presented to the subject.

The main difference, from a theoretical point of view, between semantic and auditory structures is that the hyponymy relation that exists between two words if they are sub- and superordinate concepts (for example "dog" is hyponymous to "animal") is not expected to exist between sounds. Therefore, $\mathcal{Q}$ is assumed not to hold between any two single sounds a and b, and thus $a\overline{\mathcal{Q}}b \; \forall a, b \in X, \; a \neq b$. Consequently, singleton subsets of X are not presented to the subject. This corresponds to the assumption that the sounds in X are perceptually distinct (i. e. have at least one characteristic feature). It can then be shown that transitivity as defined in Equation 2 is necessary and sufficient for an auditory structure to exist.

As an example, consider a set of four sounds $X = \{a, b, c, d\}$, and a hypothetical auditory structure $\sigma = \{\emptyset, \{a\}, \{b\}, \{c\}, \{d\}, \{a, b, c\}, X\}$ defined on it. Figure 1 displays the lattice graph of σ (without the empty set $\emptyset$). In such a graph, the features are represented as nodes connected by lines, such that lower nodes are subsets of connected higher nodes. For example, the sounds a, b, and c share the feature $\{a, b, c\}$ which is a subset of the domain X. To illustrate, let us assume $\{a, b, c\}$ is a clarinet-like timbre, which the sound d does not have. Suppose now that the relation $\mathcal{Q}$ has been established by querying a listener: In line with the assumed structure, the listener had answered with a "No" to the question whether a and b had something in common that distinguished them from c, and thus $\{a, b\}\mathcal{Q}c$. Further, let $\{a, b\}\overline{\mathcal{Q}}d$ ("Yes"), but $\{b, c\}\mathcal{Q}d$ ("No"). Then it follows from reflexivity that $\{a, b\}\mathcal{Q}b$. Given this pattern of responses, the relation $\mathcal{Q}$ is intransitive, since transitivity would require that

$$\{a, b\}\mathcal{Q}c \text{ and } \{b, c\}\mathcal{Q}d \Rightarrow \{a, b\}\mathcal{Q}d,$$

and consequently, the structure σ cannot be derived from $\mathcal{Q}$. Obviously, the listener was not able to consistently identify the timbral feature $\{a, b, c\}$ because it played a role only in some of the triadic comparisons while it was irrelevant in others, e. g.,

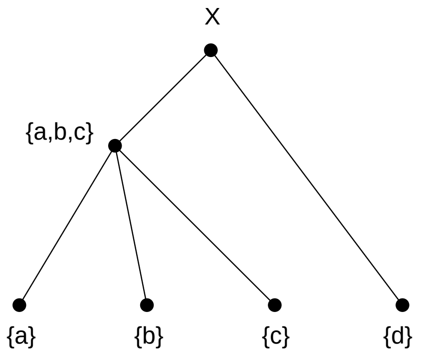


Figure 1: Lattice graph of the auditory feature structure $\sigma = \{\emptyset, \{a\}, \{b\}, \{c\}, \{d\}, \{a,b,c\}, X\}$ on the domain $X = \{a,b,c,d\}$. The empty set is omitted.

$\{b,c\}\mathcal{Q}d$. In the following representation theorem, transitivity is the critical empirical condition for a representation.

Representation

Let $\mathcal{Q}$ be a relation on $2^X \times X$ which is transitive and reflexive. Then there exists an auditory structure $\langle X, \sigma \rangle$, such that

$$A\mathcal{Q}b \text{ if and only if } b \in \sigma(A) \qquad (3)$$

for all nonempty $A \subseteq X$ and $b \in X$.

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

$$A\mathcal{Q} = \{x \in X : A\mathcal{Q}x\},$$

then it follows from Equation 3 that $A\mathcal{Q} = \sigma(A)$. It is the goal of the experiment to determine $\sigma(A)$ from the collected responses, and to construct the auditory structure σ which represents $\mathcal{Q}$.

Uniqueness

Heller (2000) has demonstrated that it is not necessary to establish the relation $\mathcal{Q}$ on *all* possible subsets of X. Instead, the subsets can be restricted to pairs only, resulting in triadic comparisons among the sounds. In doing so, not only is the number of questions drastically reduced, but also the load of the subject's working memory for each question is kept at a reasonable level. Such a reduction of questions comes at the cost of a potential loss of information, yielding a non-unique representation. In general, the equivalence in Equation 3 holds for more than one structure σ, given a set of triadic comparisons.

Formally, let $\sigma^{(2)}$ denote the largest structure (with respect to its number of elements) representing a transitive relation $\mathcal{Q}$ which is based on triadic comparisons. The structure $\sigma^{(2)}$ is constructed by

$$S \in \sigma^{(2)} \text{ if and only if } (S'\mathcal{Q}s \Rightarrow s \in S), \qquad (4)$$

for all $S \subseteq X$, $s \in X$, and all pairs S' in S. Let $\varphi^{(2)}$ denote the collection of all sets $A\mathcal{Q}$ where $A \subset X$ is a pair of sounds; thus, $\varphi^{(2)}$ is the smallest representing structure of $\mathcal{Q}$ (having fewest features). Then any structure σ for which

$$\varphi^{(2)} \subseteq \sigma \subseteq \sigma^{(2)} \qquad (5)$$

holds, is also a representing structure of $\mathcal{Q}$. In particular, the representation is unique if $\varphi^{(2)} = \sigma^{(2)}$. The restriction to triadic comparisons, therefore, does not necessarily result in a loss of information, but it will depend on the complexity of $\mathcal{Q}$ whether a unique representation can be obtained.

In summary, the experimental procedure for deriving auditory structures can be outlined as follows: First, establish $\mathcal{Q}$ based on triadic comparison judgments. Then, test the transitivity of $\mathcal{Q}$. If transitivity holds, construct both $\varphi^{(2)}$ and $\sigma^{(2)}$ (Equation 4). If $\varphi^{(2)}$ and $\sigma^{(2)}$ are equal, they form the uniquely representing auditory structure. Otherwise, all structures which satisfy Equation 5 are representing structures of $\mathcal{Q}$.

The presented approach is able to uncover the underlying auditory features, if and only if at least some stimuli have features in common and are thereby distinct from some other stimuli. If all sounds under study are indiscriminable or each entirely different from all the others (i. e. they possess only unique features), the method will not provide further insight into the auditory organization of the sounds. More precisely, the presented approach can be considered a method to derive *common* auditory features. Situations in which stimuli are perceived as entirely unique entities are, however, potentially rare. Often the context provided by a set of sounds would initiate processes of categorization and organization. It is hypothesized that such processes are feature-based; and the proposed method aims at deriving the features signifying these auditory categories.

This is – to our knowledge – the first experimental attempt to apply semantic structures (Heller, 2000) to *perceptual stimuli* rather than to verbal concepts. Therefore, to increase the chance of finding interpretable auditory structures, highly discriminable sounds will be presented to naïve listeners in two stimulus sets. The first set consists of four sounds varying in center frequency and amplitude envelope. The second set is somewhat larger, affording the representation by more complex structures: Here, the physical variables manipulated are fundamental frequency and the number of overtones. It was hypothesized that different auditory sensations of pitch, brightness, timbre, or loudness changes could be evoked by these sounds, and captured using auditory structures.

METHOD

Subjects

The sample consisted of 18 listeners (nine male, nine female), which were between 21 and 30 years of age (median 23.5 years). None of the subjects reported any hearing problems. Normal hearing of the participants was assessed using pure tone audiometry. The highest threshold found was at 25 dB hearing level (re. ISO 389-1, 1998) for one subject (M.L.) in one ear at two out of the ten audiometric frequencies between 250 and 8000 Hz.

Stimuli and apparatus

Two different sets of synthetic sounds constituted the two experimental conditions: In the first condition the stimuli consisted of four third-octave band Gaussian noises, having a center frequency of 500 and 2000 kHz, respectively, and either a rising (denoted by +) or falling (−) amplitude envelope; each noise had a duration of two seconds. In the second condition, six periodic complex tones served as stimuli, having respective fundamentals of 220 (*A*), 277 (*C#*), and 349 Hz (*F*) separated by at least a major third (400 cents), and being composed of either four or 20 harmonics in random phase. The amplitude of a given harmonic was proportional to the inverse of its number. Each complex tone had a duration of one second. All stimuli had cosine-shaped rise and fall times of 10 ms. Figure 2 depicts examples of the stimuli schematically. In the remainder of this paper, stimuli are labeled by their two components: 500+, for example, denotes the 500-Hz narrowband noise with rising envelope; *A*4 refers to the complex tone having the fundamental at 220 Hz and four harmonics, etc.

The stimuli were rendered digitally in Matlab at a sampling frequency of 44.1 kHz and exported as 16-bit Wave files. They were played back by a personal computer using a digital sound card (RME DIGI96/8 PST) connected to an external D/A converter (RME ADI-8 DS) and delivered to the headphones (Beyerdynamic DT990) by a power amplifier (Rotel RB-976 Mk II)[1]. The presentation software was implemented in LabView. The subjects entered their responses by clicking the respective buttons on a computer screen using the mouse. The experiment was conducted in a sound-insulated, double-walled listening cabin.

The instrumentally-measured loudness of the four noises and the six complex tones was aligned such that their mean loudness in sones matched approximately. In order to do so, the stimuli were recorded binaurally using a head and torso simulator (Brüel & Kjær 4128) and a measurement system (Brüel & Kjær PULSE 3560C). After loudness alignment, the noises varied within a range of 0.5 sone and the complex tones within a range of 1 sone. On average, the sound pressure level after loudness alignment was 59.3 dB L_{eq} for the noises and 60.6 dB L_{eq} for the complex tones.

Procedure

The experimental procedure consisted of two parts: In the familiarization part, the participants were presented with all four or six stimuli and asked to listen to them and try to recognize features of the sounds which they might share, or which distinguish them from other sounds. The sounds could be repeated as often as the subject desired. In the data-collection part, on each trial, the participants were presented with a stimulus triple together with the question: “Do sound A and B have something in common which makes them different from sound C?” The subjects were to answer “Yes” if they heard that the first and the second sound displayed a common feature which was not shared by the third one. Otherwise the answer was to be “No”. Any of the three sounds could be repeated as often as necessary to reach a decision.

Generally, in order to establish the relation $\mathcal{Q}$ for n stimuli, $3 \cdot \binom{n}{3}$ questions have to be asked. The 12 and 60 questions for the noises and tones

[1] The influence of the headphones on the signal was neglected in this study. It was expected that equalizing for the headphone transfer functions (which imitate a diffuse field) does not improve the identifiability of the auditory features elicited by the stimuli.

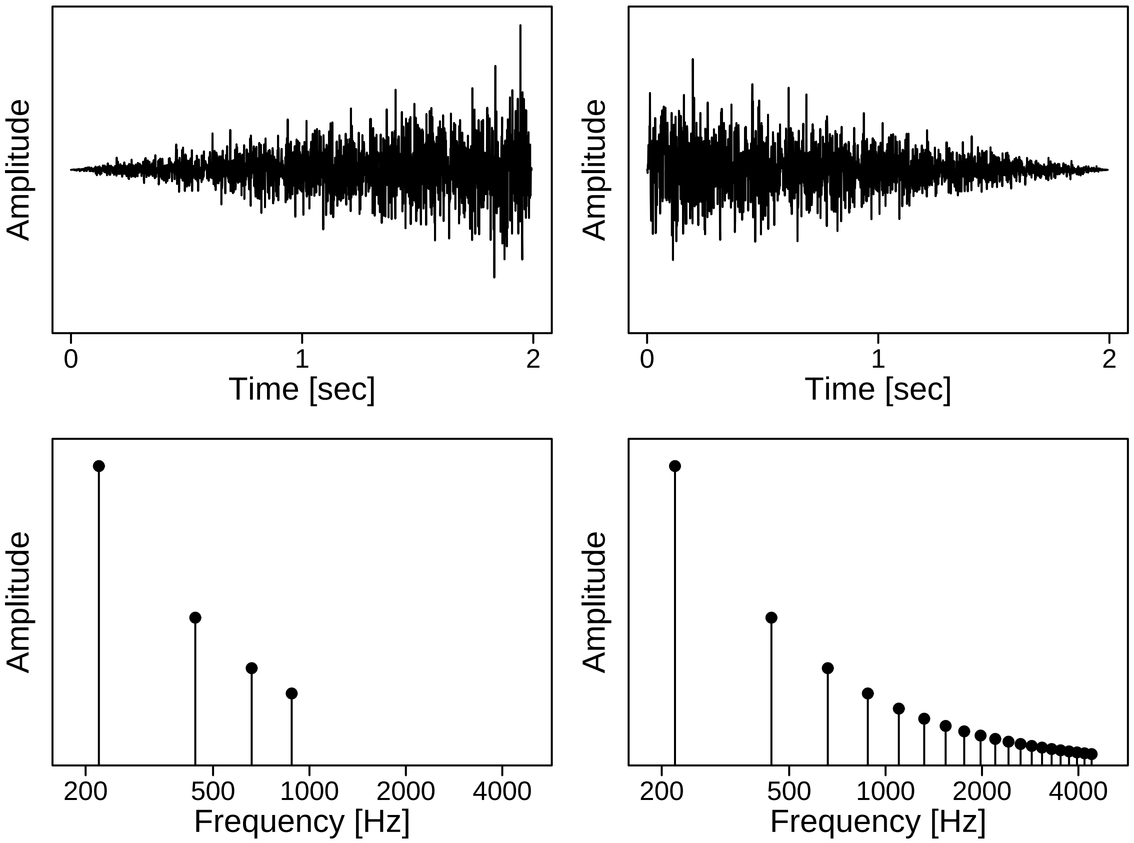


Figure 2: Illustration of the stimuli: Condition I: rising or falling third-octave band noise (upper panels). Condition II: periodic complex tones with four or 20 harmonics (lower panels). Only tones having the fundamental at 220 Hz are displayed.

were presented twice in two separate blocks. If the responses in the two blocks were identical, data collection was completed. In case of contradictory responses, a third block was presented consisting of only those questions that had been answered differently in the first and second block. The order of the triples was randomized in each block. The order of sound A and B was balanced across subjects, as was the order of the stimulus conditions: Half of the sample started with noises, the other half with complex tones.

The blocks were distributed over two sessions of approximately 45 minutes each on two separate days. A short training block of six trials preceded the data collection. The experiment was concluded by an informal debriefing in which the participants were asked to name the auditory features upon which they had based their answers.

Data analysis

The derivation of auditory feature structures from the collected triadic-comparison judgments rests on two conditions: First, the responses should reflect a certain degree of *reliability*, i. e. the number of answers changing between the blocks should be small. Second, the judgments have to be consistent enough to allow for a representation. *Consistency* is operationalized by the number of transitivity violations. In principle, only a single systematic violation prevents representability, but in an experiment it is well conceivable that a subject carelessly gives the wrong response, resulting in a random violation. Therefore, when reliability and consistency were judged to be sufficient, but (only few) transitivity violations were still present after the third block, an attempt was made to find as closely fitting a structure as possible. The discrepancy $\delta_{\mathcal{Q}}(\sigma)$ between the experimentally-determined relation $\mathcal{Q}$ and the proposed structure σ served as a lack-of-fit measure; it amounts to the number of response changes necessary to resolve all transitivity violations in $\mathcal{Q}$ in order to be consistent with σ.

RESULTS

Stimulus set I: Narrow-band noises

Reliability and consistency

Table 1 displays the indices of reliability and consistency in the narrow-band noise condition for the 18 subjects. Overall the number of responses changed between the blocks was low. In twelve cases no third block had to be presented because of perfect reliability (indicated by a dash in the third column). In the few cases in which a third block was necessary, at most one answer was changed back.

Transitivity was checked for the first, the second, and, if available, for all answers after the third block; in doing so, the contradictory answers in the first and second block were replaced by the ones given in the third block. Thus, the number of violations reported in the sixth column of Table 1 denotes the *residual* transitivity violations based on all collected responses. In 15 of the 18 subjects there were no violations after the last block; this means that the condition for representability of their judgments by an auditory feature structure was fulfilled without restriction. The remaining three subjects showed one (A.G. and V.H.) and three (A.M.) violations, respectively.

For these three subjects the violations had to be classified as random or systematic inconsistencies, the latter preventing representation by an auditory structure. The following example shall illustrate the procedure. From the analysis of A.G.'s responses it was assumed that sounds having a rising amplitude envelope (+) evoked an auditory feature salient to him[2]. One response, however, was not consistent with the hypothesis of a simple auditory structure including this feature: when asked whether sounds 500+ and 2000+ had a common feature not shared by 2000−, he had responded with "No". Considering the overall good reliability and consistency of his judgments, it seems likely that on this trial only he had missed the otherwise salient feature. Therefore, the response was changed from "No" to "Yes," resulting in a discrepancy of $\delta_{\mathcal{Q}}(\sigma) = 1$. A similar argument applies to V.R.'s data. For subject A.M., however, no representation was attempted. This is partly due to an overall lack of reliability and consistency,

[2] Heller (2000) and Choisel & Wickelmaier (2005) have developed software tools which provide assistance in finding a feature structure potentially underlying the responses in the presence of transitivity violations.

Table 1: Indices of reliability and consistency in the third-octave band noise condition. The rightmost column gives the discrepancy between the responses and the closest representing auditory structure. Note—Parentheses indicate no representation attempted.

	Response changes		Transitivity violations			
Subject	Block I–II	II–III	I	II	III	$\delta_Q(\sigma)$
A.M.	5	1	0	4	3	(2)
M.B.	0	–	0	0	–	0
M.L.	0	–	0	0	–	0
S.J.	0	–	0	0	–	0
N.C.	0	–	0	0	–	0
O.K.	0	–	0	0	–	0
A.G.	1	0	2	1	1	1
M.A.	0	–	0	0	–	0
G.B.	0	–	0	0	–	0
N.L.	0	–	0	0	–	0
S.R.	3	1	1	1	0	0
J.S.	0	–	0	0	–	0
C.G.	0	–	0	0	–	0
V.H.	3	0	2	1	1	1
F.M.	4	1	0	1	0	0
K.G.	4	0	2	0	0	0
A.K.	0	–	0	0	–	0
K.P.	0	–	0	0	–	0

and partly due to a discrepancy of $\delta_Q(\sigma) = 2$ with the closest structure, which seems high for a total of twelve questions.

Auditory structures

Based on the reported results concerning consistency and reliability, an auditory feature structure could be derived for 17 of the 18 subjects. The left panel in Figure 3 displays a lattice diagram of the structure obtained from the judgments of four subjects. The nodes in the graph denote features common to all stimuli connected to the node. The four noises (500+, 500−, 2000+, 2000−) have one unique feature each, represented by the lowest nodes in the graph. The top node represents a feature common to all four stimuli. These features, however, result already from the assumption that the sounds are discriminable and comparable, and are therefore included in any auditory structure. More interesting are the two additional features, one common to the 500-Hz stimuli and one to the 2000-Hz stimuli indicating that noises of the same center frequency share an auditory feature.

It is worth noting that one node in the lattice diagram can represent one or a combination of auditory features. More specifically, one node denotes a salient auditory category characterized by one or more features. The identifiability of single features will depend on the choice of the stimuli and on how the features covary in these stimuli. For convenience, *auditory feature* will be used interchangeably with *auditory category* bearing in mind that one feature might consist of several feature components.

The right panel of Figure 3 shows the auditory structure derived for twelve subjects. It contains four non-trivial features. In addition to the categories for noises of the same center frequency, two features are assigned to noises having the same amplitude envelope. Note that already such a relatively simple structure is too complex to be represented by a rooted tree graph which allows for only one possible pathway from the top to the terminal node.

Stimulus set II: Complex tones

Reliability and consistency

Table 2 displays the indices of reliability and consistency in the complex-tone condition. Generally, more within- and between-subjects variability was observed here than for the narrow-band noises. In one case (M.L.), however, there was perfect reliability, and six more subjects answered at most

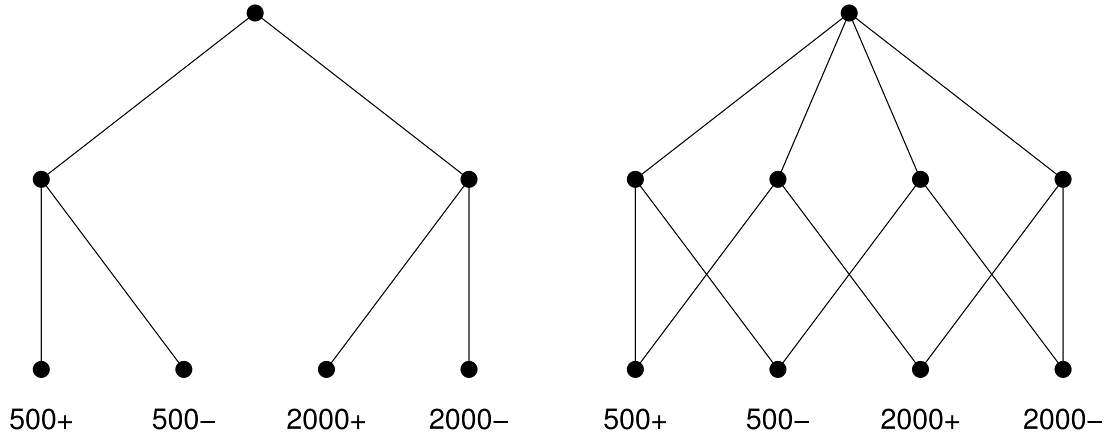


Figure 3: Auditory feature structures derived from triadic-comparison judgments of four (left) and twelve subjects (right). Stimuli consisted of narrow-band noises having the center frequencies [Hz] indicated by the numbers, and rising (+) or falling (−) amplitude envelopes.

four of the 60 questions differently when queried the second time, which indicates a high degree of reliability in their judgments. For three subjects (O.K., J.S., and V.H.) not only the number of changes between the first and second block, but also the fact that they reversed about half of these answers when queried again, suggests that their judgments are unreliable.

Overall transitivity was found to hold without violation for eight listeners, which corresponds to perfect representability. Seven more subjects displayed a discrepancy of at most four answers ($\delta_{\mathcal{Q}}(\sigma) \leq 4$) with an auditory structure, and a representation was therefore attempted. In doing so, the remaining transitivity violations were classified as random errors. This appeared to be justified when several violations were resolved by changing only few responses. For example, there were six violations left for subject F.M. after the last block; only a single response change was needed to resolve them all. This makes it likely that the subject had carelessly given the response. For the three subjects O.K., J.S., and V.H., consistency was not judged sufficient for a representation; the discrepancy of their judgments with the closest fitting structure was at least five answers.

Auditory structures

A representation was derived for 15 of the 18 participants. The left panel of Figure 4 displays an auditory structure representing the judgments of four of the participating subjects. It includes two nontrivial auditory categories, one for the 4-harmonic and one for the 20-harmonic complex tones, indicating that overtone content elicited a common auditory sensation. The right panel of Figure 4 shows a structure derived for six other subjects. It contains three additional features common to tones of the same fundamental frequency (A, C#, and F).

All auditory feature structures for which perfect representability held were found to be unique in the sense of the uniqueness theorem which means that $\varphi^{(2)}$ and $\sigma^{(2)}$ were equal (see Equation 5), i. e. all features in a given structure followed directly from the relation $\mathcal{Q}$ which was established by the triadic comparisons. This was true with one exception: Subject M.A. displayed a rather complex relation $\mathcal{Q}$. Consequently, the triadic comparisons did not provide enough information to decide whether two features were or were not included in his auditory structure. These two features were the ones characterizing the four- and 20-harmonic tones, respectively. The resulting non-uniqueness could be resolved in two ways: One possibility is to ask the questions which provide the necessary information. One such question could for example have been: "Do the sounds A4, C#4, and F4 share a feature that A20 does not have?" Thus, quadruple comparisons would have resolved the non-uniqueness. The second, less elegant, but more practical solution is to rely on the debriefing to provide the missing information, and in this case there was no doubt that both features were included in this participant's auditory structure.

Table 2: Indices of reliability and consistency in the complex-tone condition. The rightmost column gives the discrepancy between the responses and the closest representing auditory structure. Note—Parentheses indicate no representation attempted.

	Response changes		Transitivity violations			
Subject	Block I–II	II–III	I	II	III	$\delta_{\mathcal{Q}}(\sigma)$
A.M.	9	4	21	25	17	3
M.B.	6	5	17	17	14	4
M.L.	0	–	0	0	–	0
S.J.	3	1	20	6	9	3
N.C.	4	4	10	0	10	4
O.K.	20	10	36	37	24	(5)
A.G.	2	1	8	0	0	0
M.A.	8	0	12	0	0	0
G.B.	9	1	25	3	0	0
N.L.	11	0	39	0	0	0
S.R.	13	7	23	14	11	4
J.S.	11	6	31	26	20	(6)
C.G.	1	0	6	0	0	0
V.H.	32	17	24	56	33	(7)
F.M.	2	2	6	18	6	1
K.G.	3	1	8	6	0	0
A.K.	9	1	31	4	4	1
K.P.	18	0	0	0	0	0

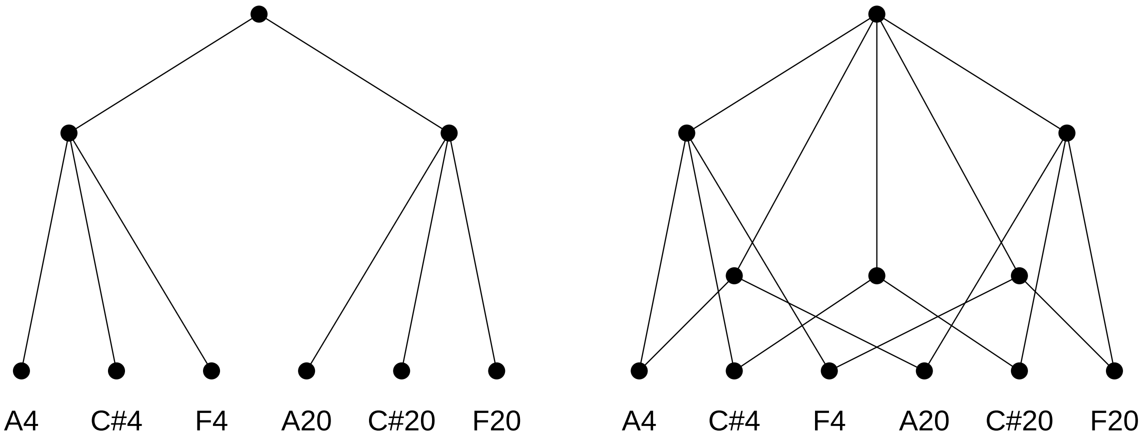


Figure 4: Auditory feature structures derived from triadic-comparison judgments of four (left) and six subjects (right). Stimuli consisted of complex tones having their fundamentals at 220 (A), 277 ($C\#$), and 349 Hz (F), and four or 20 harmonics (denoted by the numbers).

Comparing feature structures

So far, the results presented were strictly individual. Indeed, one of the strengths of the proposed method is that it does not rely on aggregated or averaged data, but allows for individual differences to become apparent. On the other hand, a researcher might be interested in questions like "How salient is a given auditory feature in a sample of subjects?" or "How (in)homogeneous with respect to auditory perception is the sample under study?" which require a certain level of aggregation. Such questions can be answered within the framework of auditory structures. In order to do so, the individual structures of all subjects were arranged in a common lattice graph in which each node represents a possible auditory structure.

Figure 5 shows the lattice graph of all extracted individual structures in the narrow-band noise condition. Solid circles denote structures which actually represent the judgments of one or more listeners while open circles indicate *potential* structures which were not implied by the actual judgments. The top node shows the simple structure displayed in the left panel of Figure 3; only the non-trivial features $\{500+, 500-\}$ and $\{2000+, 2000-\}$ are indicated (braces are omitted for convenience). To the left of the node are the initials of the four subjects for which this structure was derived. Lower level nodes represent structures including the features of *all* higher level nodes that can be reached by following ascending lines. Therefore, the lower the node, the more complex is the structure. For example, the node labeled $\{500-, 2000-\}$ also contains the three features from higher level nodes; it represents the structure shown in the right panel of Figure 3 and was derived for twelve subjects.

From Figure 5 it is obvious that the two features elicited by noises of the same center frequency ($\{500+, 500-\}$ and $\{2000+, 2000-\}$) were the most salient in the sample, because they are included in all auditory structures. Thirteen listeners had the feature $\{500+, 2000+\}$, twelve listeners the feature $\{500-, 2000-\}$. Only one subject (S.R.) displayed an extra feature shared by three noises ($\{500+, 2000+, 2000-\}$) which was therefore the least salient in the sample. The fact that only three different structures were derived argues for a strong agreement between the subjects about the auditory features emerging from this stimulus set.

Figure 6 displays the common lattice graph for the complex-tone condition. The two structures shown in Figure 4 are denoted by the two top-left nodes in the graph. The most salient features were the ones assigned to tones with the same number of harmonics; fourteen of the 15 subjects for whom structures were derived used the two features $\{A4, C\#4, F4\}$ and $\{A20, C\#20, F20\}$. The features elicited by tones of the same fundamental frequency ($\{A4, A20\}$, $\{C\#4, C\#20\}$, and $\{F4, F20\}$) were included in the structures of eight subjects. Four subjects perceived an auditory feature when two tones with the same number of harmonics were not more than one third apart from each other, for example $\{A20, C\#20\}$ or $\{C\#4, F4\}$. This might indicate that fundamental frequency and number of harmonics interact in order to create a new feature. The additional features found by subject S.R. seem to be rather idiosyncratic, and the informal debriefing did not provide further information as to how they might be labeled appropriately. In general, however, the simple shape of Figure 6 indicates good agreement between the subjects.

Labeling auditory features

The labeling of the obtained features does not directly follow from the triple-comparison judgments. So far, by deriving feature structures, the stimuli have been organized into categories (or sets), which – due to the absence of other, e. g. semantic, information – may be assumed to be *auditory* categories. In the case of carefully designed synthetic stimuli (as in the present case) an educated guess might be attempted as to what forms the perceptual basis of these categories. In general, however, the category labels will have to be inferred from the information acquired in the debriefing session *after* the data collection.

In the narrow-band noise condition, descriptions like "crescendo"/"decrescendo" or "fade in"/"fade out" corroborate the hypothesis that noises of the same amplitude envelope (+ or −) elicited auditory sensations related to their dynamic loudness. Descriptions like "low"/"high" or "thick"/"thin" indicate that noises of the same center frequency (500 or 2000 Hz) shared the same pitch or brightness feature. In the complex-tone condition, subjects described the four- and 20-harmonic tones as being played on two different instruments or as "smooth" versus "scratchy," indicating that by manipulating the number of harmonics two timbral auditory features were elicited. The sensations evoked by sounds of the same fun-

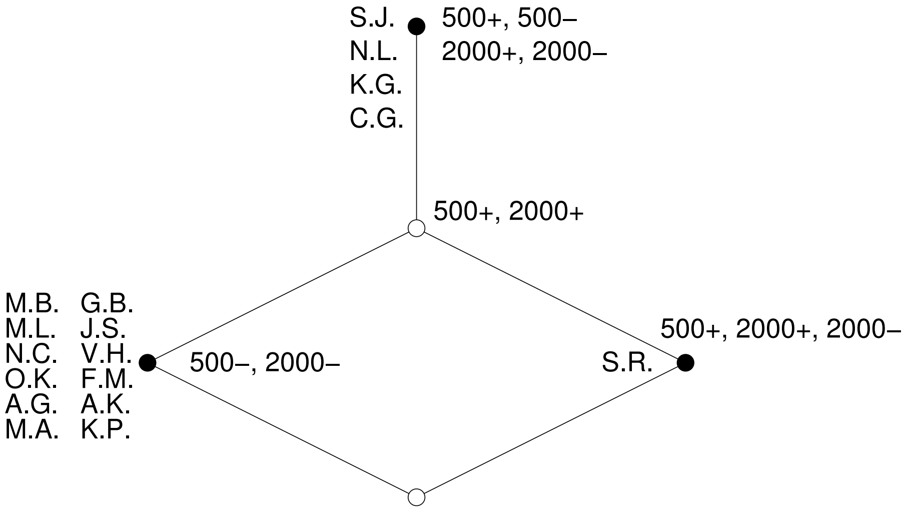


Figure 5: Lattice graph of the 17 individual auditory structures representing the narrow-band noises.

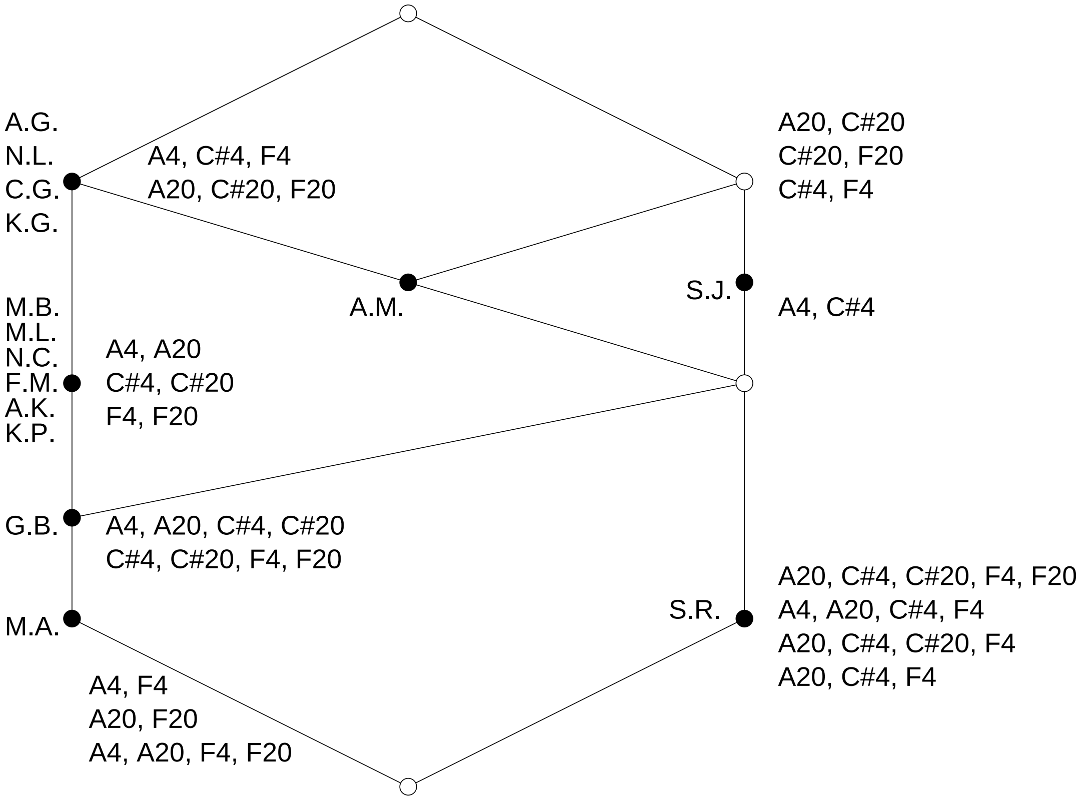


Figure 6: Lattice graph of the 15 individual auditory structures representing the complex tones.

damental frequency were described as different musical notes or as being "low," "medium," or "high" suggesting that the pitch was varied when manipulating the fundamental frequency.

DISCUSSION

The major advantage of the presented method is that it aims at a feature representation of auditory stimuli on the basis of simple qualitative judgments. In spite of the simplicity of the answer ("Yes"/"No") it should not be overlooked that the decision process to reach such an answer is potentially much more complex. In order to reduce the demand for the subject it is crucially important to include a familiarization part in the experimental session together with instructions for the subject to structure the stimuli, and to recognize, identify, and organize the auditory features. With more complex stimuli than the ones used in the present study, a more elaborate familiarization could be fruitful, potentially in form of a tutorial using pictures of simple geometric shapes and eye-catching visual features.

It is hypothesized that the success of the method—the highly consistent judgments, and the large number of representations obtained—was partly due to the simple acoustical structure of the stimuli. The high demands of the task for the subject will become more obvious as the sounds under study become more complex. In such a case, eventually labeling the auditory features would require more accurate information than could be obtained by the informal debriefing in the present study. A possible strategy for a more formal debriefing would be to present a listener with the sets of sounds from his or her own auditory structure together with the question: "Describe briefly what feature(s) these n sounds share with each other, but not with the remaining sounds." From the answers to such questions it should also be possible to resolve the problem of non-unique representations as reported in the results section.

A restriction of the presented approach is that it requires a certain independence from the *local* context provided by a given triad of sounds. Instead, the decision whether or not a feature belongs to a sound has to be based on the *global* context provided by all sounds under study. Features based on local context effects will most likely result in inconsistent judgments. It is, however, possible that a subject learns to appreciate new features during the data-collection part of the experiment which had not been identified during the familiarization part. Consider as an example subject K.P. in the complex-tone condition. Throughout the experiment, her judgments were perfectly consistent indicating that she could clearly identify the salient features. There were, however, 18 response changes between the first and the second block. Together with the perfect consistency this should not be taken as unreliable behavior, but rather as an evidence that learning of new features has occurred after the data-collection part of Block I, and these features were then consistently judged in the remaining blocks.

With more complex stimuli, the classification of the transitivity violations as random or systematic would also become more difficult. Unfortunately, as with other applications of axiomatic measurement theory, there are no simple criteria for such a classification. Rather, the indices of reliability and consistency, their development over time (blocks), and the discrepancy $\delta_{\mathcal{Q}}(\sigma)$ have to be considered together in order to decide whether there is enough evidence in the data that allows the violations to be classified as random, and consequently for the transitivity axiom holding. A statistical test would certainly remedy the problem, but such a test has not been developed yet. Lacking such a test, it is common practice to relate the number of violations of an axiom to the number of possible tests of that axiom, implying that a higher violation *ratio* is indicative of stronger evidence against that axiom to hold. Such a strategy, however, cannot be advocated for perceptual structures, since the more transitivity tests that are possible, the more frequently a subject must have responded with a "No," which in turn implies that only few features have to be considered. The more features a subject has in mind, the more complex $\mathcal{Q}$ will become, and the *less* transitivity tests are possible. For that reason, the number of possible tests can be misunderstood and was therefore not reported in the present study.

Concluding remarks

In summary, the following conclusions can be drawn based on the present study: (1) The subjects were able to produce reliable and consistent judgments about common auditory features of simple synthetic sounds. (2) The proposed approach, founded on measurement theory, for deriving auditory features was shown to have the advantage that (a) the representation can fail due to inconsistent judgments and is therefore falsifiable, (b) it provides an opportunity to test the identifiability of auditory features, and (c) it does not require labeling of the features encountered and therefore separates perceptual from verbal abilities of a subject. (3) The results from the present study encourage the application of the method to more complex auditory stimuli (Choisel & Wickelmaier, 2005), and potentially to investigating features in other perceptual modalities.

Author's Note

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 comes from the Ministry for Science, Technology, and Development (VTU), and from the Danish Research Council for Technology and Production (FTP). The authors are grateful to Sylvain Choisel for numerous helpful discussions of the theory and implementation of auditory feature structures and to Ville Sivonen for helping with the instrumental analyses of the stimuli. Correspondence concerning this paper should be sent to Florian Wickelmaier, Department of Acoustics, Aalborg University, Fredrik Bajers Vej 7 B5, 9220 Aalborg, Denmark, e-mail: fw@acoustics.dk

References

Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological Review*, **75**, 127–142.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: theory and applications.* New York: Springer.

Carrol, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of Echart-Young decomposition. *Psychometrika*, **35**, 283–319.

Choisel, S., & Wickelmaier, F. (2005). Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound. *The Audio Engineering Society 118th Convention, Barcelona, preprint 6369.*

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces.* Berlin: Springer.

Falmagne, J.-C., Koppen, M., Villano, M., Doignon, J.-P., & Johannesen, L. (1990). Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review*, **97**, 201–224.

Ganter, B., & Wille, R. (1999). *Formal concept analysis.* Berlin: Springer.

Grey, J. M. (1977). Multidimensional scaling of musical timbres. *Journal of the Acoustical Society of America*, **61**, 1270–1277.

Heller, J. (2000). Representation and assessment of individual semantic knowledge. *Methods of Psychological Research*, **5**, 1–37.

ISO 389-1 (1998). Reference zero for the calibration of audiometric equipment – Part 1: Reference equivalent threshold sound pressure levels for pure tones and supra-aural earphones. ISO, Geneva, Switzerland.

Iverson, P., & Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbre. *Journal of the Acoustical Society of America*, **94**, 2595–2603.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement I.* New York: Academic Press.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**, 1–27.

Lakatos, S., McAdams, S., & Caussé (1997). The representation of auditory source characteristics: Simple geometric form. *Perception & Psychophysics*, **59**, 1180–1190.

Sattath, S., & Tversky, A. (1987). On the relation between common and distinct feature models. *Psychological Review*, **94**, 16–22.

Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an an unknown distance function I. *Psychometrika*, **27**, 125–139.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, **17**, 401–419.

Tversky, A. (1977). Features of similarity. *Psychological Review*, **84**, 327–352.

Wille, R. (1982). Restructuring lattice theory: An approach based on hierarchies of concepts. In I. Rival (Ed.), *Ordered sets*, (pp. 445–470). Dordrecht: Reidel.

Selecting participants for listening tests involving multichannel reproduced sound[*]

Florian Wickelmaier[1] Sylvain Choisel[1,2]

[1] *Sound Quality Research Unit, Dept. of Acoustics, Aalborg University, 9220 Aalborg, Denmark*
[2] *Bang & Olufsen A/S, 7600 Struer, Denmark*

Abstract

A selection procedure was devised in order to select listeners for experiments in which their main task will be to judge multichannel reproduced sound. Ninety-one participants filled in a web-based questionnaire. Seventy-eight of them took part in an assessment of their hearing thresholds, their spatial hearing, and their verbal production abilities. The listeners displayed large individual differences in their performance. Forty subjects were selected based on the test results. Self-assessed listening habits and experience as obtained from the web questionnaire did not predict the results of the selection procedure. Further, absolute hearing thresholds did not correlate with the spatial-hearing test. This leads to the conclusion that task-specific performance tests might be the preferable means of selecting a listening panel.

1 INTRODUCTION

In many experiments involving human listeners the experience or expertise of the participants is of crucial importance. The experimenter has to make a decision as to whether the subjects should be naïve (unexperienced with respect to the task) or experts. Clearly, if generalizability of the experimental results was the only concern, one would randomly sample the level of experience rather than restrict the sample to only that, potentially small, part of the population which can be regarded as "expert listeners" by any given criterion. On the other hand, human behavior is always characterized by an intrinsic random component, which often makes it a difficult task to extract the systematic effects, unless certain sources of variation have been eliminated a priori. It is often assumed, and there is empirical evidence (Bech, 1992), that expert listeners display less variation in their judgments and are therefore more reliable.

One way of dealing with this dilemma between generalizability and reliability is to select the participants randomly and subsequently train them to become experts (Bech, 1993). While this strategy might be applicable to many experiments, great care has to be taken in order not to bias subjects' judgments by the training procedure. The risk of biasing listeners, however, is particularly high in studies having an exploratory character because subjects might be asked to make judgments about a variety of auditory sensations, possibly even new and as yet unlabeled ones. Since the present selection was made for such a study, training subjects was disregarded. Rather the strategy in the present study was to start from a random sample of participants and select the best ones according to specified criteria. In contrast to other procedures e. g. the Generalized Listener Selection procedure (*GLS*, Mattila and Zacharov, 2001), which are similar in spirit, but base the selection on general listening abilities, the current procedure presents the participants with specific tests which are related to the abilities required in later tasks.

The participants were selected for a series of experiments which aim at uncovering auditory attributes of multichannel reproduced sound. It will be important that the panelists can appreciate the differences between different reproduction modes

[*] Portions of this work have been presented at the 118th Convention of the Audio Engineering Society, Barcelona, Spain, 2005 May 28–31.

(like mono, stereo, 4- and 5-channel surround, etc.). It will also be important that they possess good verbal abilities, especially when it comes to promptly producing a description of their sensations. The challenge for the selection procedure is to assess these abilities without telling the subjects what to listen for or what to describe, and thereby irrevocably biasing their judgments. For the two desired abilities this problem was addressed in the following ways: A discrimination test of sounds varying in stereo width was conducted employing a three-interval forced-choice procedure in which the task was to choose the "odd" sound. In doing so, naming the involved attribute is circumvented, both by the experimenter and the subject. It was assumed that listeners with better discrimination could also differentiate the reproduction modes more easily in later experiments. The verbal production abilities were assessed via a standard verbal fluency test (Spreen and Strauss, 1998), assuming that participants with a high fluency score can describe their sensations more readily.

It is generally recommended (e. g. by the International Electrotechnical Commission, CEI IEC technical report 60268-13, 1998) that an audiometric test should not be the (only) means for selecting a listening panel. In this study, audiometry was used to assess normal hearing of the participants and to supplement the specific ability tests. As a further supplement, data about the listening habits and prior experience of the participants were collected by means of a questionnaire made available via the Internet before the experiment started.

2 METHOD

The selection was conducted in four steps. First, the candidates signed up for the tests by filling in a web questionnaire. Invitations to the experiments had been placed on the Internet, as well as in public places, such as libraries, cafeterias, music shops, pubs, and shopping centers. The requirements for participation in the study were (a) to be a native Danish speaker and (b) to be available for the duration of the project (about ten months). After signing up, the participants were invited to the tests proper, which included audiometry, a spatial hearing test, and a verbal fluency test. These tests were conducted in a double-walled sound-insulated chamber.

2.1 Web questionnaire

A web-based questionnaire, inspired by the one used by Mattila and Zacharov (2001), was used for registering the subjects' demographic variables and listening experience into a database. The entire questionnaire can be seen in the Appendix. Two questions were used to screen the participants for clinically relevant hearing problems or hearing damage. 91 persons filled in the questionnaire, of which four did not fulfill the language requirement, two were participants in parallel (and possibly biasing) experiments, and seven dropped out. The remaining 78 participated in the selection tests, none of them reported any known hearing problems or damage.

2.2 Audiometry

The next requirement for the 78 listeners was a maximum hearing loss of 20 dB HL (re. ISO 389-1, 1998) in any ear at any frequency between 250 Hz and 8 kHz. The audiometric test was performed using a Madsen (model OB 40) audiometer. Twenty of the subjects had already participated in earlier experiments, and recent audiometric data were available.

2.3 Stereo-width discrimination

The second test concerned the subjects' ability to discriminate between sounds which varied in stereo width. Stereo width was manipulated by decomposing the signal into a weighted sum of the sum $(L + R)$ and the difference $(L - R)$ of the left and right channels (Equation 1). This weighted sum is sometimes called the mid/side (MS) ratio, especially when the sound has been recorded with both an omnidirectional/cardioid and a bidirectional microphone. From an original stereo recording with left and right channel L and R, a new signal (L', R') varying in stereo width can be derived by

$$\begin{aligned} L' &= (1 - \tfrac{\beta}{2})(L + R) + \tfrac{\beta}{2}(L - R), \\ R' &= (1 - \tfrac{\beta}{2})(L + R) - \tfrac{\beta}{2}(L - R), \end{aligned} \qquad (1)$$

where the parameter β determines the stereo width: When β equals one, the left and right channel of the derived sound are identical to the original stereo channels; when β equals zero, L' and R' both amount to the sum of the stereo channels, i. e. mono. By varying β between zero and one, it is possible to create sounds having a different degree of stereo width, from mono to stereo.

It is apparent from Equation 1 that changing β will result in a change of the interchannel correlation. For the stimuli used in the present study, changing β from zero to one corresponds to a change of the interchannel correlation from one to 0.6. Blauert (1997, chap. 3) reports experimental results which indicate that the width of the auditory event associated with a signal presented over headphones increases when the interchannel correlation decreases.

2.3.1 Apparatus

A personal computer equipped with a sound card (RME Hammerfall HDSP) connected to an external D/A converter (RME ADI-8 DS) was used to play back the sounds in the MS-ratio test. The stimuli were of approximately 1.5 s duration and were presented over headphones (Beyerdynamic DT990) fed by a headphone amplifier (Behringer Powerplay 4400). The stereo recording was presented at an A-weighted equivalent level of 66.4 dB SPL to the left and 68.8 dB SPL to the right ear[1] as measured with an artificial ear (Brüel & Kjær 4153). The participants entered their responses by clicking one of three buttons presented on a computer screen.

2.3.2 Procedure

An adaptive procedure (3AFC, 2-up/1-down; Levitt, 1971; Jesteadt, 1980) was employed in order to assess the reduction in stereo width that was detected 71% of the time. A stereo recording of a piano chord (EBU document Tech. 3253, 1988, track 39, at 1'53) served as a standard, and a comparison was derived from it by changing the MS ratio according to Equation 1. The participants performed a forced-choice oddity task. On each trial, they had to identify which of the three sounds was *different* from the other two. According to the subject's response, the comparison varied adaptively from mono ($\beta = 0$) towards stereo ($\beta = 1$), converging on the 71%-discrimination threshold. After responding correctly in two successive trials, β was increased; it was decreased after every wrong answer. The step size decreased with increasing β-value by $0.3(1 - \beta)$. Thus, as the task became harder, the step size became smaller, and the upper bound of $\beta = 1$ could never be exceeded. The procedure stopped after eight reversals, and the discrimination threshold was estimated from the average β-value at the last four reversals. In order for the procedure to be less transparent to the listeners, each trial contained either two standards and one comparison, or one standard and two (identical) comparisons, in random order (i. e. the odd sound was either standard or comparison). The subjects were not told what kind of differences to listen for, and therefore were free to use any criterion. In order to remove loudness cues as much as possible, the mono stimulus was attenuated by 3 dB. This attenuation was decreased gradually to zero, as β approached a value of one. Figure 1 displays the stimulus generation schematically.

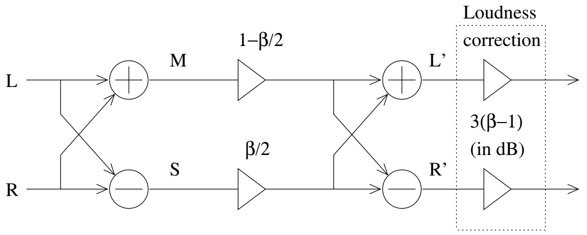


Figure 1: Stimulus generation in the MS-ratio test: The output is a mono signal when $\beta = 0$, and the original stereo when $\beta = 1$. The final stage applies a gain varying between -3 dB (mono) and 0 dB (stereo).

In a pilot experiment including four consecutive measurements of seven listeners, it was observed that the discrimination thresholds improved after the first measurement, and remained constant afterwards. Therefore, before the actual test, the subjects underwent a short familiarization, in which the procedure stopped after two reversals. The results from the familiarization were not incorporated into the threshold estimate. For none of the participants did the measurement last longer than 15 minutes, including familiarization.

2.4 Verbal fluency

In the last test, the subjects' verbal production abilities were assessed in an alternating verbal fluency test (Spreen and Strauss, 1998). At the beginning of the test the following written instruction (in Danish) was handed out:

> In this task you should within one minute name as many different Danish words as possible which belong alternately to the categories "animals" and "fruits". First name

[1] This level difference is inherent in the recording as a result of the asymmetry of the source (a piano) with respect to the microphones.

> an animal then a fruit, then again an animal, etc. Please do not repeat a word you have already said before.
>
> Please try to say as fast as possible as many different words as you can. Start with an animal.

The participants were seated in the listening booth in which a microphone, connected to the computer in the control room, was used to record the word list. The sound files were saved, and were analyzed at a later point in time. A fluency score was assigned to each word list by counting the correct responses. Incorrect were words not belonging to the categories "animals" or "fruits", newly created words, proper nouns or given names, word repetitions, and category perseverations (naming, e. g., two animals in a row, as in "mouse – goat – nut"). A familiarization session preceded the test, which was identical, but had the two different semantic categories "professions" and "items which can be found in a supermarket". After having checked that the instructions were understood, the familiarization ended. Results from the familiarization were not included in the fluency score.

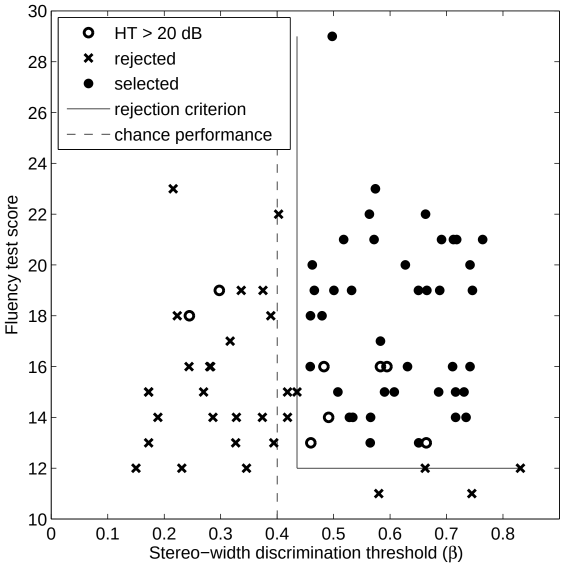


Figure 2: Results of the selection procedure. Listeners above and to the right of the rejection criterion (solid lines) were selected on the basis of the results in the discrimination and fluency test (solid circles). A priori rejected were participants with a hearing threshold (HT) of more than 20 dB (open circles), or with chance performance in the discrimination test (dashed line).

3 RESULTS

Figure 2 displays the data obtained in the selection procedure. On the abscissa the stereo-width discrimination thresholds are displayed, on the ordinate the fluency test scores. The discrimination thresholds ranged from 0.15 to 0.83 (corresponding to MS ratios between 93:7 and 59:41) with a mean of 0.50 and a standard deviation of 0.18. Note that higher thresholds indicate higher sensitivity to reduced stereo width, since $\beta = 1$ leaves the original stereo sound unchanged, and $\beta < 1$ reduces the stereo width. The fluency scores ranged from 11 to 29 with a mean of 16.6 and a standard deviation of 3.4. Eight out of the 78 participants (marked by open circles) had a mild hearing loss of between 25 and 40 dB in either ear at at least one of the audiometric frequencies between 250 and 8000 Hz, and were therefore rejected.

A further criterion for a-priori rejection was a stereo-width discrimination threshold below chance level. The chance level was determined by means of a Monte-Carlo simulation, in which the outcome of the adaptive procedure was recorded when a virtual subject responded randomly. On each simulation run, 1000 simulated thresholds were generated. Figure 3 shows a typical example of the distribution of the resulting simulated thresholds. The median of this distribution is close to zero (0.08). The chance level was adopted as the 95% percentile of the distribution, which lies at 0.4. In order to take variations due to sampling into account, 100 simulation runs were performed and each time the 95% percentile was estimated. The 95% percentile of the 100 estimates again was found to be 0.4 and consequently set to the criterion of chance performance. The 24 participants having a lower sensitivity than 0.4 were excluded, because it cannot be ruled out that they were guessing while performing the discrimination task.

In order to select the final listening panel, the following decision rule was applied to the remaining subjects: A subject was removed from the list of candidates, if he or she performed worst at either the stereo-width discrimination or the verbal fluency task. In doing so, both tasks were weighted equally. This elimination process stopped when 40 subjects were left. The selected listeners lie in the upper right quadrant of Figure 2 (marked by solid circles). They are separated from the rejected sub-

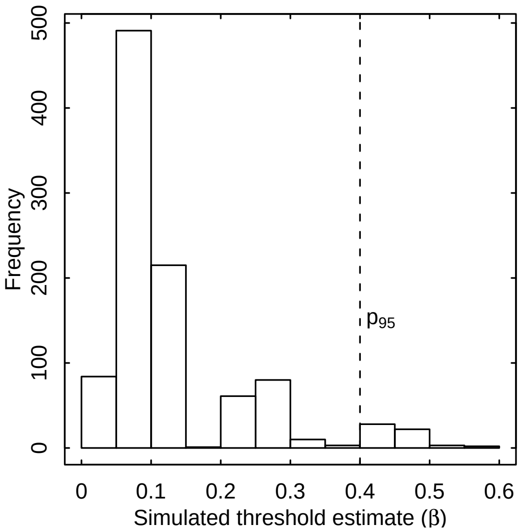


Figure 3: Typical result of the Monte-Carlo simulation of discrimination thresholds, when the responses were given at random. Displayed are the frequencies of 1000 simulated thresholds. The 95% percentile is indicated by the dashed line.

jects (crosses) by a horizontal and a vertical line. These lines correspond to rejecting the worst cases both with respect to verbal production abilities and sensitivity to changes in stereo width. Another eight listeners were excluded on the basis of this criterion over and above those previously excluded due to hearing loss, or below-chance performance in the spatial hearing test. The remaining 40 subjects were selected for participating in later listening experiments.

3.1 Stereo-width discrimination and demographic variables

The influence of the demographic variables acquired via the web questionnaire (see Appendix) on the sensitivity to stereo width was investigated. In particular, sex, occupational background of the participants, habits concerning listening to music, attending concerts, or going to the cinema, playing an instrument, owning a hi-fi or a surround sound system, considering oneself as a critical listener, and being professionally involved in music or audio were included in the analysis. Frequently, such variables are considered potential predictors of listening abilities. Table 1 shows the average discrimination thresholds stratified by those variables together with the sample size and standard deviation.

Table 1: Estimated stereo-width discrimination thresholds stratified by demographic variables, self-assessed experience, and listening habits. Only the difference between males and females is statistically significant. Note—n sample size, M mean, SD standard deviation.

		Est. threshold	
Category	n	M	(SD)
Sex			
male	43	0.56	(0.18)
female	27	0.41	(0.16)
Background			
music	13	0.59	(0.15)
engineering	27	0.51	(0.18)
languages	9	0.51	(0.22)
social science	15	0.42	(0.17)
others	6	0.45	(0.20)
Professional experience			
yes	20	0.56	(0.17)
no	50	0.48	(0.19)
Listening to music			
daily	59	0.50	(0.19)
weekly	11	0.50	(0.18)
Attending concerts			
weekly/monthly	32	0.54	(0.18)
rarely or not	38	0.47	(0.18)
Playing instrument			
daily	25	0.55	(0.20)
rarely or not	45	0.48	(0.17)
Critical listener			
yes	63	0.50	(0.18)
no	7	0.49	(0.22)
Going to cinema			
monthly	43	0.49	(0.20)
less than monthly	27	0.53	(0.16)
Own hi-fi system			
yes	54	0.52	(0.18)
no	16	0.45	(0.19)
Own surround system			
yes	14	0.44	(0.18)
no	56	0.52	(0.18)
Participated in tests			
yes	25	0.48	(0.17)
no	45	0.52	(0.19)

Contrary to the expectations, however, only the variable *sex* turned out to have a significant influence on the discrimination threshold, as confirmed by a two-sample t-test [$t(68) = 3.66$; $p < .001$]. On average the male subjects were by about 0.9 of a standard deviation more sensitive than the females. In order to investigate interactions with the occupational background of the participants, they were assigned according to their profession the five categories *engineering*, *languages and communication*, *music and music therapy*, *social sciences*, and *others*. The majority of the participants were students of the respective fields. A two-factor analysis of variance revealed no significant interaction between background and discrimination threshold [$F(4, 60) = 1.71$; $p = .159$], nor a significant main effect of the background [$F(4, 60) = 2.31$; $p = .068$], but a highly significant gender effect [$F(1, 60) = 12.94$; $p < .001$]. A similar result (main effect of sex only) was obtained when analyzing an interaction with the self-assessed professional experience. Since there was no reason to a priori expect better performance of the male participants in the stereo-width discrimination tests, no gender correction was applied to the tests results. As a consequence, proportionally more females than males were rejected based on this criterion.

3.2 Semantic fluency and demographic variables

No significant differences between groups based on sex, education and age were found in semantic fluency. Table 2 shows the mean fluency scores, standard deviations and sample sizes stratified by sex, years of education and age. A three-factor analysis of variance revealed no significant influence of the interaction of sex, education and age [$F(1, 65) = 0.72$; $p = .398$], nor significant two-way interactions or main effects. The variation of the fluency scores can therefore be attributed to the individual differences in the sample.

Sample percentiles of the 78 native Danish speakers are displayed in Table 3. Participants having a score of less than 13 were rejected according to the rejection criterion (cf. Figure 2). This corresponds to excluding the subjects in the lower 10% of the distribution.

Table 2: Semantic fluency scores (animals–fruits) stratified by sex, years of education and age. Note—n sample size, M mean, SD standard deviation.

		Fluency score	
Category	n	M	(SD)
Sex			
female	28	17.1	(3.1)
male	50	16.3	(3.6)
Education, years			
less than 13	5	15.8	(3.6)
13–16	36	17.1	(3.8)
more than 16	37	16.2	(3.0)
Age, years			
20–24	41	16.8	(3.8)
25–29	29	16.5	(3.1)
30–44	8	15.6	(2.2)

Table 3: Sample percentiles of the semantic fluency test.

Percentile	10	25	50	75	90
Fluency score	13	14	16	19	21

4 DISCUSSION

The participants displayed considerable variation in the results of the specific ability tests. Therefore, these two tests are especially suited for selection purposes. The web-based questionnaire, however, failed to explain the differences in sensitivity to changes in stereo width entirely. Especially, the questions related to (self-assessed) prior experience or to listening habits of the subjects provide no good means for predicting the results. From these findings it might be concluded that such investigations into the *attitudes* of potential panelists should have little priority for their selection, whereas the main focus should be put on their *behavior* observed in specific tests.

An impression of the range of the fluency test scores observed in the present study might be obtained by comparing them to the scores in similar investigations. Table 4 displays the results of other studies on alternating word fluency. The fluency scores, both in terms of mean value and standard deviation, are comparable with results from studies involving healthy native English speakers (Zec et al., 1999; Baldo et al., 2001; Phillips et al., 2002; Bouquet et al., 2003). The results in Dujardin et al. (2001) are based on a sample of Parkinson's patients, who obtained the lowest fluency scores among the listed studies. Note, however, that neither the age structure nor the semantic (or lexical) categories exactly match those of the present study.

Finally, it was found that the discrimination thresholds in the spatial hearing test cannot be predicted by the absolute thresholds measured in the audiometry. The correlation between the maximum hearing threshold per subject obtained at any of the frequencies at either ear with the spa-

Table 4: Results of alternating word fluency tests among native English speakers. The bottom line shows the present sample (native Danish speakers). The sample in Dujardin et al. (2001) consisted of Parkinson's patients, remaining results are of healthy control subjects. Note—n sample size, M mean, SD standard deviation, age and education in years.

		Fluency score		Age		Education	
Semantic/lexical categories (study)	n	M	(SD)	M	(SD)	M	(SD)
Boy's names–fruits (Bouquet et al., 2003)	20	17.1	(4.0)	63.5	(10.1)	9.9	(3.5)
M-words–vegetables (Phillips et al., 2002)	60	14.5	(3.0)	29.1	(6.4)	–	–
Fruits–furniture (Baldo et al., 2001)	11	16.0	(3.3)	68.1	–	14.6	–
L-words–R-words (Dujardin et al., 2001)	9	9.7	(2.5)	54.8	(8.2)	11.7	(2.8)
Colors–occupations (Zec et al., 1999)	45	14.0	(3.8)	63.1	(10.6)	13.6	(3.1)
Animals–states (Zec et al., 1999)	45	17.5	(5.4)	63.1	(10.6)	13.6	(3.1)
C-words–P-words (Zec et al., 1999)	45	10.2	(4.6)	63.1	(10.6)	13.6	(3.1)
Animals–fruits	78	16.6	(3.4)	25.8	(5.0)	–	–

tial discrimination threshold was not significant [$r = .06$; $p = .578$]. Toole (1985) reported a positive correlation between hearing level below 1 kHz and the variability of fidelity ratings. He hypothesized that a high absolute sensitivity at the middle and lower frequencies is an important criterion when selecting listeners for judging sound quality. Figure 4 shows the spatial discrimination thresholds as a function of hearing level below 1 kHz. In the present study, there is no systematic relation between these two parameters, which argues that the results from audiometry should be given little priority in the selection procedure, when the panelists' task is to judge supra-threshold stimuli. In contrast to Toole who observed mean hearing levels up to 30 dB at middle and lower frequencies, however, in the present study mean hearing levels were only up to 8.75 dB in this frequency range. Furthermore, the performance criterion is different: the present study focuses on discrimination sensitivity, whereas Toole focused on judgment reliability.

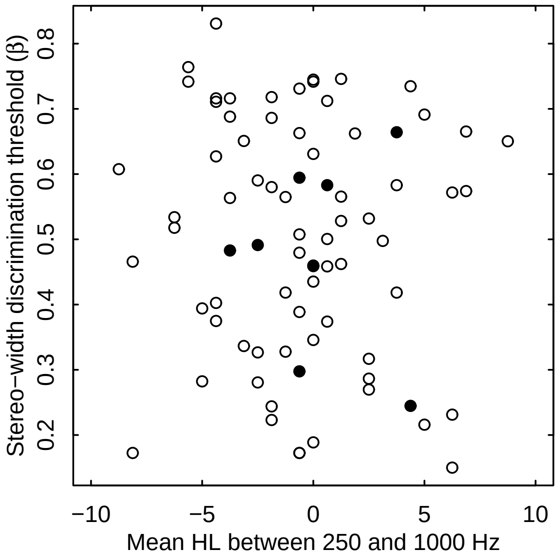


Figure 4: Spatial discrimination (β) as a function of mean hearing level (HL) below 1 kHz. Filled circles denote the eight subjects having a hearing threshold of more than 20 dB HL at *any* frequency.

Can the selection procedure predict the performance of the panelists?

A major assumption underlying the selection procedure is that the selected listeners would outperform the non-selected ones in later experiments, be it by their superior ability to discriminate the sounds, by their better verbalization skills, or generally by an increased reliability of their judgments. While it was not within the scope of the present study to perform a rigorous validation of the selection procedure, data from the main experiments might be taken as an indicator for the relevance of the tests performed.

Among the tasks of the selected panel in the main experiments was to judge overall preference between audio reproduction formats (mono, stereo, and several multichannel formats). Details of these experiments are reported in Choisel and Wickelmaier (2005a). Eight reproduction formats were compared to each other by making all pairwise comparisons. From the choice data, a preference scale was derived using suitable models. The preference scaling was repeated with four different musical excerpts. Figure 5 shows the preference scales for the eight reproduction formats for one

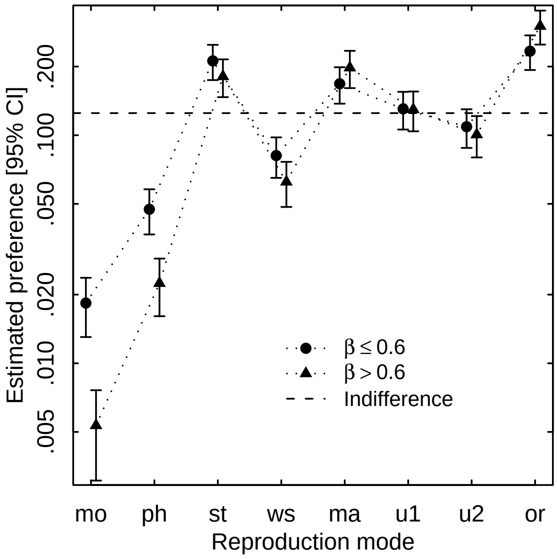


Figure 5: Preference scales for eight reproduction modes derived from paired-comparison judgments. Circles denote the group of 19 subjects having lower sensitivity to stereo width, triangles denote the group of 20 subjects having higher sensitivity. The reproduction modes were mono (*mo*), phantom mono (*ph*), stereo (*st*), wide-angle stereo (*ws*), four- (*ma*) and five-channel upmixing (*u1* and *u2*), and the original five-channel material (*or*). The musical excerpt was Steely Dan. Error bars show 95%-confidence intervals.

excerpt (Steely Dan). If all pairwise choice frequencies were 50%, that is, if the subjects were indifferent in their preferences between the reproduction modes, the scale values would lie on the line of indifference (at 1/8). Consequently, the distance from this line denotes how pronounced the preferences for the reproduction modes are. Subjects were grouped according to their discrimination threshold (β) into a group ($n = 19$) of less sensitive, and a group ($n = 20$) of more sensitive listeners. The cutoff value for this classification was at $\beta = 0.6$.

In Figure 5, subjects with higher sensitivity ($\beta > 0.6$) display on average greater distances from the indifference line, which suggests that they are more sensitive in the preference task (less likely to make an indifferent choice). Table 5 shows the mean absolute distance from the indifference line for the two groups of subjects. This distance is greater for the more sensitive listeners for all musical excerpts, which implies that they have more pronounced preference patterns. A χ^2-test (detailed in Choisel and Wickelmaier, 2005a) indicates that the preference scale values for the less sensitive listeners were significantly different from those of the more sensitive subjects in three of the four excerpts. In summary, this suggests that listeners with a higher sensitivity (β) as measured in the selection procedure, displayed also an increased sensitivity in the preference task in the main experiments, which corroborates the hypothesis that a relevant ability was measured by this selection test.

Table 5: Comparison of preference scales obtained for subjects grouped by discrimination threshold (β). Displayed are the mean absolute distance from the indifference line and a χ^2-test for equality of the scale values in the two groups.

	Mean distance		Equality	
Excerpt	$\beta \leq .6$	$\beta > .6$	$\chi^2(7)$	p
Beethoven	.056	.065	14.21	.048
Rachmaninov	.057	.062	4.60	.709
Steely Dan	.061	.077	37.45	<.001
Sting	.054	.060	35.99	<.001

Only a limited amount of data from the main experiments was available in order to assess the validity of the verbal fluency test. A group of 20 listeners performed a verbal elicitation task (Choisel and Wickelmaier, 2005b) in which they were to describe the perceptual differences among the sounds. The correlation between the average number of provided descriptors per subject and the fluency score was not significant ($r = .20$; $p = .390$). It should, however, be noted that these results are based on only a portion of the selected panel, thus the power to detect a significant correlation is small. Furthermore, the sample consisted largely of university students, and it is to be expected that fluency differences would be larger in a sample which is more heterogeneous with respect to age and education. Finally, just as the audiometric test might be considered a means to screen the sample for hearing damages, the fluency test might help to identify subjects having insufficient verbalization skills.

Concluding remarks

In this study two performance tests, a spatial discrimination test and a verbal fluency test, were proposed for the purpose of selecting listeners for experiments on the evaluation of multichannel reproduced sound. The advantages of the two tests chosen are, that they allow for an efficient assessment of listeners, give rise to sufficient variance between them, and are easily analyzed and reported in quantitative indices.

Results of the selected panel of listeners in the main experiments indicate that subjects with a better performance in the discrimination test were also more sensitive to differences between (multichannel) reproduced sounds. Data which clearly support the validity of the verbal fluency test are yet to be provided by future studies.

5 Acknowledgments

This manuscript was completed while the authors were employed at the Sound Quality Research Unit at Aalborg University. This unit receives financial support from Delta Acoustics & Vibration, Brüel & Kjær, and Bang & Olufsen, as well as from the Danish National Agency for Industry and Trade (EFS) and the Danish Technical Research Council (STVF). The authors would like to thank Wolfgang Ellermeier for his helpful comments on an earlier draft of the manuscript.

References

Baldo, J. V., Shimamura, A. P., Delis, D. C., Kramer, J., and Kaplan, E. (2001). Verbal and design fluency in patients with frontal lobe lesions. *Journal of the International Neuropsychological Society*, 7:586–596.

Bech, S. (1992). Selection and training of subjects for listening tests on sound-reproducing equipment. *Journal of the Audio Engineering Society*, 40:590–610.

Bech, S. (1993). Training of subjects for auditory experiments. *Acta acustica*, 1:89–99.

Blauert, J. (1997). *Spatial Hearing.* MIT Press, Cambridge, USA.

Bouquet, C. A., Bonnaud, V., and Gil, R. (2003). Investigation of supervisory attentional system functions in patients with parkinson's disease using the hayling task. *Journal of Clinical and Experimental Neuropsychology*, 25:751–760.

CEI IEC technical report 60268-13 (1998). Sound system equipment – Part 13: Listening tests on loudspeakers. International Electrotechnical Commission.

Choisel, S. and Wickelmaier, F. (2005a). Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. *In preparation.* (This thesis).

Choisel, S. and Wickelmaier, F. (2005b). Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound. *118th Convention of the Audio Engineering Society, Barcelona, Spain, May 28–31.* Preprint 6369.

Dujardin, K., Defebvre, L., Krystkowiak, P., Blond, S., and Destée, A. (2001). Influence of chronic bilateral stimulation of the subthalamic nucleus on cognitive function in parkinson's disease. *Journal of Neurology*, 248:603–611.

EBU document Tech. 3253 (1988). Sound quality assessment material. Recordings for subjective tests. Users' handbook for the EBU-SQAM compact disc. European Broadcasting Union.

ISO 389-1 (1998). Reference zero for the calibration of audiometric equipment – Part 1: Reference equivalent threshold sound pressure levels for pure tones and supra-aural earphones. ISO, Geneva, Switzerland.

Jesteadt, W. (1980). An adaptive procedure for subjective judgments. *Perception & Psychophysics*, 28:85–88.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, 49:467–477.

Mattila, V. V. and Zacharov, N. (2001). GLS – A generalised listener selection procedure. *Journal of the Audio Engineering Society (Abstracts)*, 49:546. Preprint 5405.

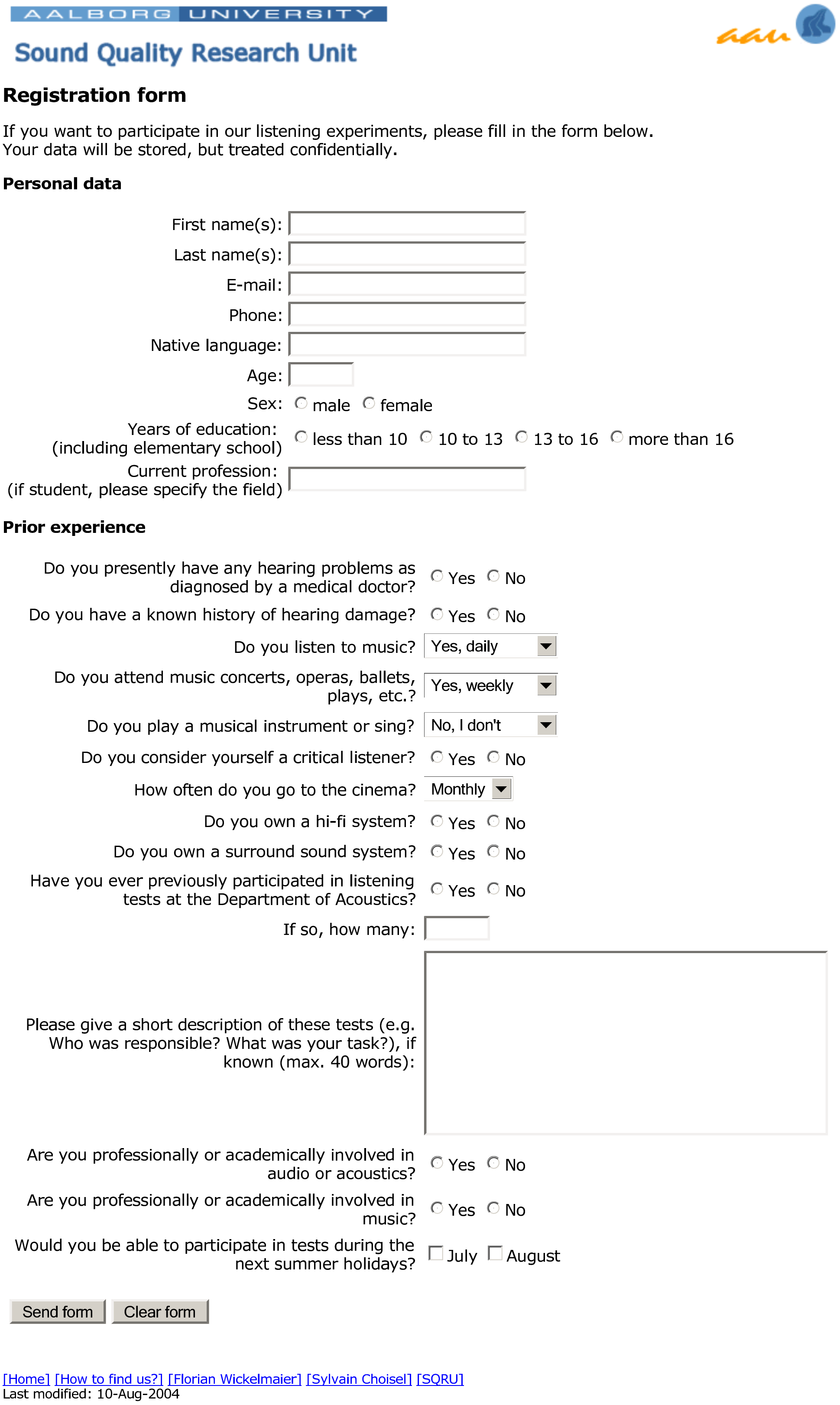
Phillips, L. H., Bull, R., Adams, E., and Fraser, L. (2002). Positive mood and executive function evidence from stroop and fluency tasks. *Emotion*, 2:12–22.

Spreen, O. and Strauss, E. (1998). *A compendium of neuropsychological tests.* Oxford University Press, New York.

Toole, F. E. (1985). Subjective measurements of loudspeaker sound quality and listener performance. *Journal of the Audio Engineering Society*, 33:2–32.

Zec, R., Landreth, E., Belman, J., Fritz, S., Hasara, A., Fraiser, W., Wainman, S., McCool, M., Grames, E., O'Connell, C., Harris, R., Robbs, R., Elble, R., and Manyam, B. (1999). A comparison of phonemic, semantic, and alternating word fluency in parkinson's disease. *Archives of Clinical Neuropsychology*, 14:255–264.

Appendix: Web-based questionnaire



AALBORG UNIVERSITY

Sound Quality Research Unit

aau

Registration form

If you want to participate in our listening experiments, please fill in the form below.
Your data will be stored, but treated confidentially.

Personal data

First name(s): ____

Last name(s): ____

E-mail: ____

Phone: ____

Native language: ____

Age: ____

Sex: ○ male ○ female

Years of education: (including elementary school) ○ less than 10 ○ 10 to 13 ○ 13 to 16 ○ more than 16

Current profession: (if student, please specify the field) ____

Prior experience

Do you presently have any hearing problems as diagnosed by a medical doctor? ○ Yes ○ No

Do you have a known history of hearing damage? ○ Yes ○ No

Do you listen to music? Yes, daily ▼

Do you attend music concerts, operas, ballets, plays, etc.? Yes, weekly ▼

Do you play a musical instrument or sing? No, I don't ▼

Do you consider yourself a critical listener? ○ Yes ○ No

How often do you go to the cinema? Monthly ▼

Do you own a hi-fi system? ○ Yes ○ No

Do you own a surround sound system? ○ Yes ○ No

Have you ever previously participated in listening tests at the Department of Acoustics? ○ Yes ○ No

If so, how many: ____

Please give a short description of these tests (e.g. Who was responsible? What was your task?), if known (max. 40 words): ____

Are you professionally or academically involved in audio or acoustics? ○ Yes ○ No

Are you professionally or academically involved in music? ○ Yes ○ No

Would you be able to participate in tests during the next summer holidays? □ July □ August

Send form | Clear form

[Home] [How to find us?] [Florian Wickelmaier] [Sylvain Choisel] [SQRU]
Last modified: 10-Aug-2004

Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound*

Sylvain Choisel[1,2] Florian Wickelmaier[1]

[1] *Sound Quality Research Unit, Dept. of Acoustics, Aalborg University, 9220 Aalborg, Denmark*
[2] *Bang & Olufsen A/S, 7600 Struer, Denmark*

Abstract

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 directly assign verbal labels to the features when encountering them, and to subsequently rate the sounds on the scales thus obtained. The second method 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. Based on the individual results from both methods, a common list of eight attributes was obtained.

1 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 field of reproduced sound [1, 2, 3, 4, 5, 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, which can then be used by either experts or non-experts. A more individual approach, applicable to naïve subjects, is the *repertory grid technique* (RGT) [2, 8]. Listeners are instructed to label the differences they can identify between the sounds, and subsequently rate the sounds on the descriptors thus obtained. Various verbalization methods have been used in other studies [e. g., 3, 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 hand. This is problematic in at least two ways: First, the elicitation of auditory attributes will be dependent upon 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 an elicitation procedure. Second, it cannot be ensured

[*] Portions of this work have been presented at the 118th Convention of the Audio Engineering Society, Barcelona, Spain, 2005 May 28–31.

that, when a listener provides a verbal expression, it is related to an actual sensation at all. That the listener had the sensation of, e. g., 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) [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 [e. g., 10], and only requires a judgment of perceived distances between stimuli, typically in the form of dissimilarity ratings. The outcome is a map of the stimuli in a multidimensional space. The interpretation of the dimensions, however, is not straightforward, but often requires additional knowledge about the stimuli, for instance rating scales as obtained from the RGT.

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. This method, which is based on *knowledge space theory* [12] and *formal concept analysis* [13], was adapted to the extraction of auditory features, and experimentally tested with synthetic sounds by Wickelmaier and Ellermeier [14]. In addition to its mathematical foundation, its major advantage lies in the fact that it strictly separates the identification of auditory sensations from their labeling. One of the research questions in this work is whether this method is applicable to more complex auditory stimuli, typically encountered in sound quality assessment of multichannel systems.

In the present study, both RGT and PSA were employed as methods to elicit auditory attributes. The two techniques 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.

2 METHOD

2.1 Setup and stimuli

2.1.1 Program material

Four musical excerpts (two pop, two classical) were selected from commercially available multichannel recordings (Table 1), and recorded on a computer (48 kHz, 24 bit) from their original medium—Super Audio Compact Disc (SACD) or Digital Versatile Disc–Audio (DVD-A)—using a Denon 2200 player connected to an 8-channel A/D converter (RME ADI-8 DS). The excerpts were carefully cut to include a musical phrase, their duration ranging from 4.7 to 5.4 s.

2.1.2 Experimental setup

The loudspeaker configuration (represented in Fig. 1) consisted of a 5-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 speakers 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 between 0.25 and 0.45 s. The setup was hidden from the subject by a curtain. The sounds

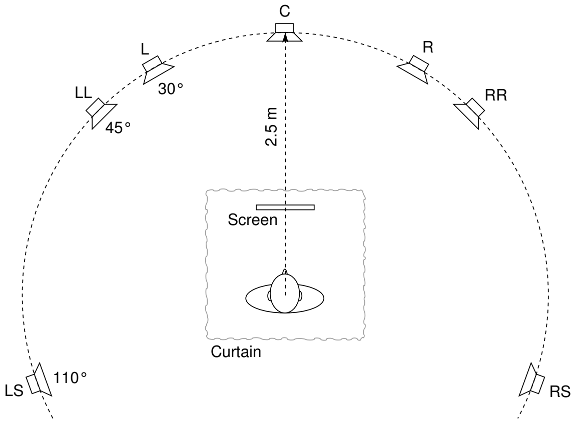


Figure 1: Playback setup consisting of seven loudspeakers: left (L), right (R), center (C), left-of-left (LL), right-of-right (RR), left surround (LS) and right surround (RS). This setup was symmetrically placed with respect to the width of the room and was hidden from the subject by an acoustically 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 – 1'56
Rachmaninov: Vespers – St. Petersburg Chamber Choir/Korniev	Blazen Muzh	SACD	03	2'04 – 2'09
Steely Dan: Everything Must Go	Everything Must Go	DVD-A	09	0'52 – 0'57
Sting: Sacred Love	Stolen Car	SACD	06	1'55 – 2'00

were played back by a PC placed in the control room, equipped with an RME Hammerfall HDSP sound card connected to an 8-channel D/A converter (RME ADI-8 DS).

The response interface consisted of a 15" flat screen placed in front of the listener, at a height of 45 cm 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.

2.1.3 Downmixing and upmixing

From the original 5-channel recordings, several formats were derived as shown in Table 2. When present (only in the pop recordings), the low-frequency effect (LFE) channel was disregarded.[1]

The original 5-channel program material was mixed down to stereo using Equation 1, as recommended in [15].

$$\begin{aligned} L_{st} &= \tfrac{1}{\sqrt{2}}\left(L + \tfrac{1}{\sqrt{2}}C + \tfrac{1}{\sqrt{2}}LS\right) \\ R_{st} &= \tfrac{1}{\sqrt{2}}\left(R + \tfrac{1}{\sqrt{2}}C + \tfrac{1}{\sqrt{2}}RS\right) \end{aligned} \qquad (1)$$

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

$$C_{mo} = \tfrac{1}{\sqrt{2}}\left(L_{st} + R_{st}\right) \qquad (2)$$

$$L_{ph} = R_{ph} = \frac{1}{2}\left(L_{st} + R_{st}\right) \qquad (3)$$

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°.

Table 2: Reproduction modes: full name, abbreviation and loudspeakers used for playback (see Figure 1).

Name	Abbr.	Speakers
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.

Finally, three upmixing algorithms were used to re-construct 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 settings shwon 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 a 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 3 dB (Equation 4). After informal listening, it was decided not to use the center channel, because of obvious timbral differ-

[1] According to ITU-R BS.775-1 [15, p. 10], the LFE channel is optional and "should not [...] be used for the entire low frequency content of the multi-channel sound presentation [but] only carry the additional enhancement information."

ences from the other reproduction modes.

$$\begin{aligned} L_{ma} &= L \\ R_{ma} &= R \\ LS_{ma} &= \tfrac{1}{\sqrt{2}}(L-R) \\ RS_{ma} &= \tfrac{1}{\sqrt{2}}(R-L) \end{aligned} \tag{4}$$

Table 3: Parameters for the Pro Logic II upmix on the 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

2.1.4 Equalization and calibration

The on-axis 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 (Brüel & Kjær 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 speaker responses and the calculated filters can be seen in Fig. 2 (upper panel). The resulting equalized responses are shown in the bottom panel.

Each channel of the stimuli was equalized using the corresponding filter. This equalization was based on the 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 inter-channel 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–2 kHz) was employed, and recorded for 10 s at the listening position using a Brüel & Kjær 4134 pressure-field microphone pointing upwards. This signal was equalized for each channel based on the anechoic loudspeaker responses, in the same way as all the musical excerpts. The A-weighted sound pressure level was then calculated from the recordings. The inter-channel level differences were within 0.3 dB, but the differences between left/right pairs did not exceed 0.1 dB.

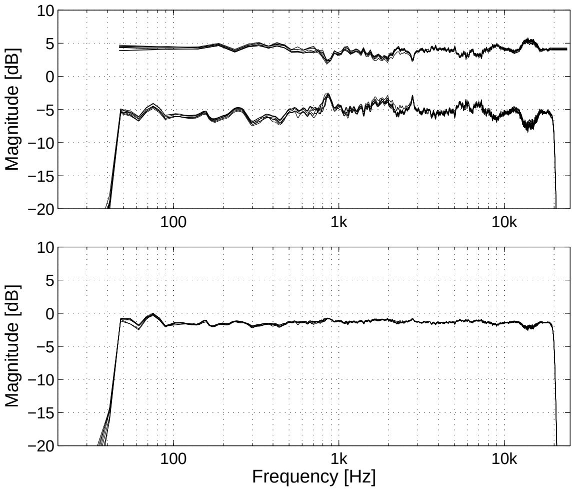


Figure 2: Top panel: Measured frequency responses of the seven loudspeakers (lower curves) and their corresponding equalization filters (upper curves). Bottom panel: Equalized loudspeaker responses.

2.1.5 Loudness matching

After the inter-channel 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 different reproduction modes of the same musical excerpts. 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 employing 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 being the comparison, in random order. When the listener indicated that the comparison was louder than the standard, the level of the comparison was reduced, and 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) 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 randomly interleaved in a single block, with a probability proportional to the number of remaining reversals. Each track had a random starting level between ±3 dB. On average one block lasted 12.7 min for Beethoven, 14.0 min for Rachmaninov, 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.

2.2 Subjects

Thirty-nine listeners (27 males, 12 females) were selected among 78 candidates, according to their listening abilities and verbal fluency—see [21] for details on the selection procedure[2]. They were all native Danish speakers, and their age 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 randomly assigned to one of three groups, two of which took part in the repertory grid technique, and the third one in the perceptual structure analysis.

2.3 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.

2.3.1 Elicitation of verbal descriptors

A triadic elicitation procedure was implemented following Berg and Rumsey [2]. On each trial, the subject was presented with triples 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 re-use 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 subjects took part in the triadic elicitation (referred to as RGT-3), while 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 program materials (one pop and one classical, balanced across subjects), while the second group completed the task for all four types of music.

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

2.3.2 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. Eight lines were displayed on the screen, 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.

2.3.3 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 one involves classifying the verbal data into semantic categories [e. g., 3]. The second one 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 defined 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 tree-like 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 experimenter. In the present study, a cut-off level was chosen, above which the clusters were disregarded, and below which sounds clustering together were combined into a single attribute.

2.4 Perceptual structure analysis

Perceptual structure analysis (PSA) attempts to extract auditory features identified by the subjects in a set of sounds. 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 employed in this study are presented.

2.4.1 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], $\langle X, \sigma \rangle$ 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.

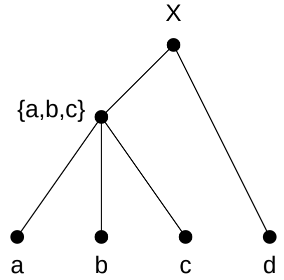


Figure 3: Lattice graph of a hypothetical perceptual structure. The sounds are denoted by a, b, c and d; $\{a,b,c\}$ represents a feature shared by a, b and c, but not by d; $X = \{a,b,c,d\}$ is the domain.

A relation $\mathcal{Q}$ which 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 $A\mathcal{Q}x$, if and only if the subject answers "No" to the question:

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

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

$\mathcal{Q}$ is said to be *transitive* if

$$A\mathcal{Q}b\ (\forall b \in B) \text{ and } B\mathcal{Q}c \Rightarrow A\mathcal{Q}c \tag{5}$$

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

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

$$\{a,b\}\mathcal{Q}c \text{ and } \{b,c\}\mathcal{Q}d \Rightarrow \{a,b\}\mathcal{Q}d,$$

and consequently, a perceptual structure cannot be derived.

If and only if transitivity holds, $\mathcal{Q}$ can be represented by a perceptual structure on X such that:

$$A\mathcal{Q}b \text{ if and only if } b \in \sigma(A) \tag{6}$$

for all nonempty $A \subseteq X$ and $b \in X$.

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

$$A\mathcal{Q} = \{x \in X : A\mathcal{Q}x\},$$

then it follows from Equation 6 that $A\mathcal{Q} = \sigma(A)$. 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 subsets A is usually too large to be accommodated in an experimental session. For that reason, in the present experiment, the subsets are restricted to pairs only. The consequence of such an incomplete design is a potentially non-unique representation [11], 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 have at least one characteristic feature. Finally, $\mathcal{Q}$ is assumed to hold for both sounds in each pair, i. e. $\{a,b\}\mathcal{Q}a$ and $\{a,b\}\mathcal{Q}b$, so that each pair will only be presented together with the remaining $|X| - 2$ sounds.

2.4.2 Experimental procedure

Let n be the number of sounds under study in an experiment. For each of the $n(n-1)/2$ 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 which 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. It is therefore of major importance that the subject already has a clear idea of the features before proceeding with this task. For that purpose, a short 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. In order to clarify the concept of *feature*, the subject 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.

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 high working load required by this method, only $n = 7$ reproduction modes were included in the experimental design; the matrix-upmixed format 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.

2.4.3 Fitting perceptual structures

In principle, a single violation of transitivity (Equation 5) in the responses of a subject prevents their representation by a perceptual structure. It is possible, however, to visually inspect the violations and attempt to resolve them by changing as few responses as possible from "Yes" to "No" or vice versa. Often a single response change 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 changes, or evaluate several solutions to resolve the violations.

Therefore, a computer program was developed, which searches for the best solution allowing a feature representation, i. e. the structure that best fit 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 discrepency with the responses—there are $2^{(2^n-n-2)} = 2^{119} = 6.6 \times 10^{35}$ possible structures. Testing all possible response changes could also be a possible approach, but the complexity of such an algorithm quickly rises with the number of response changes. The proposed method attempts to infer the features which potentially underlay the subject's responses. 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 stimulus triple.

From Equation 5 it follows that a violation occurs when:

$$B \subseteq A\mathcal{Q} \text{ and } B\mathcal{Q} \nsubseteq A\mathcal{Q},$$

where A and B are two pairs in X. The violations can be resolved in several ways:

(i) replace $A\mathcal{Q}$ by $A\mathcal{Q} \cup B\mathcal{Q}$,

(ii) remove from $B\mathcal{Q}$ the elements which are not in $A\mathcal{Q}$,

(iii) remove from $A\mathcal{Q}$ one or both elements of B, with the restriction that $A\mathcal{Q}$ must still contain A.

This can be interpreted as follows: Generally, adding elements to either $A\mathcal{Q}$ or $B\mathcal{Q}$ 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 $A\mathcal{Q}$ 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 changed answers. When too many violations were observed, such a simulation was not attempted.

2.4.4 The uniqueness problem

Because the relation $\mathcal{Q}$ 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 can be answered 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 debriefing session has to clarify whether or not all proposed features had been identified.

2.4.5 Labeling of the features

When a perceptual structure was obtained, the subject was asked to label each feature during a debriefing session. All seven sounds were arranged in a playlist, some of which—those sharing a common feature according to the structure obtained—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.

3 RESULTS

3.1 Repertory grid technique

The number of verbal descriptors elicited using RGT are shown in Table 4. Although the maximum number of descriptors was higher in the pairwise (RGT-2) than in the triadic (RGT-3) elicitation, on average, the difference between the two methods was rather subtle.

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

	RGT-2			RGT-3		
Prog. mat.	min	max	mean	min	max	mean
Beethoven	3	12	7.9	3	9	5.4
Rachman.	3	13	8	5	13	9.2
Steely Dan	2	19	8.8	2	13	8.6
Sting	4	15	9.9	3	11	8.2

In order to reduce the number of descriptors per subject, cluster analysis was performed on the ratings associated with each descriptor. This analysis was done individually for each subject, in a similar way as proposed by Berg and Rumsey [2]. The result of the cluster analysis for one subject (92) is depicted in Fig. 4 as an example. A cut-off level was chosen, below which all scales connected together were grouped into the same cluster. A higher cut-off level will result in fewer clusters. A value of 0.3 was found appropriate for most subjects, resulting in an average number of clusters of 3.6 (Beethoven), 4.2 (Rachmaninov), 5.1 (Steely Dan) and 5.1 (Sting).

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 one minus the correlation coefficient 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.

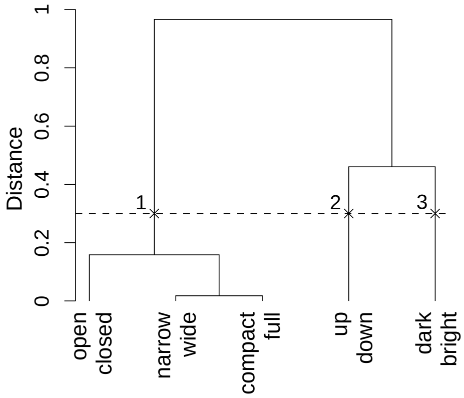


Figure 4: Cluster analysis of the RGT constructs for Sting, rated by subject 92. The distances are derived from the absolute correlations. The dashed line indicates the cut-off level of 0.3, resulting in three clusters. Verbal descriptors are translated from Danish.

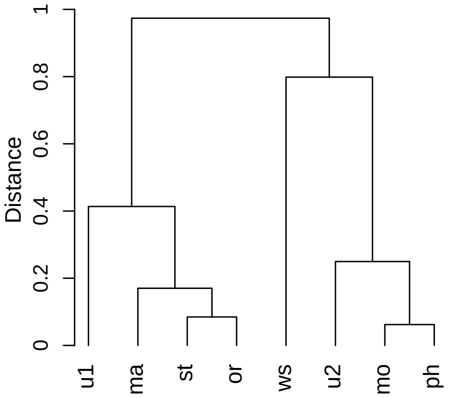


Figure 5: Cluster analysis of the reproduction modes for Sting, rated by subject 92. The distances are derived from the correlations.

3.2 Perceptual structure analysis

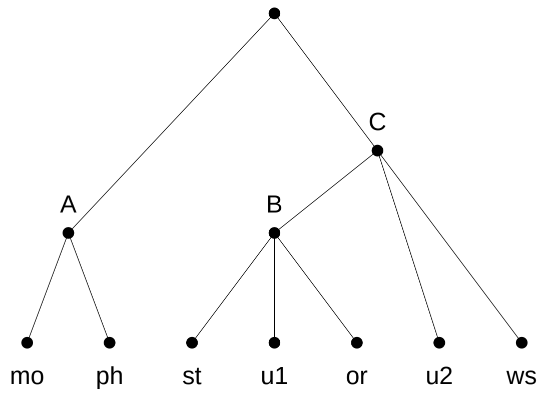
From the subjects binary (yes/no) responses to each triple of sounds, the simulation attempted to find the best fitting perceptual structure. A measure of fit $\delta_Q(\sigma)$ is reported in Table 5 for all cases where the fitting of a structure had been attempted. It is calculated as the relative number of response alterations necessary for $\mathcal{Q}$ to be consistent with the structure σ: $\delta_Q(\sigma) = c/T$, where c is the number of response changes and T the total number of triples (105 in the present case). In Table 5, the number of response changes between the first and the second session (I-II) and between the second and the third (II-III) are also reported. This number is an indicator of the subject's reliability: a value close to 50% would suggest that

Table 5: Reliability and consistency of the judgments collected in PSA. Displayed are, for each subject, the response changes from session to session, the number of transitivity violations in the three sessions, and a measure of fit between structures and data. Roman numerals indicate session numbers.

Subj.	Resp. changes		Transitivity			$\delta_Q(\sigma)$
	I-II	II-III	I	II	III	
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

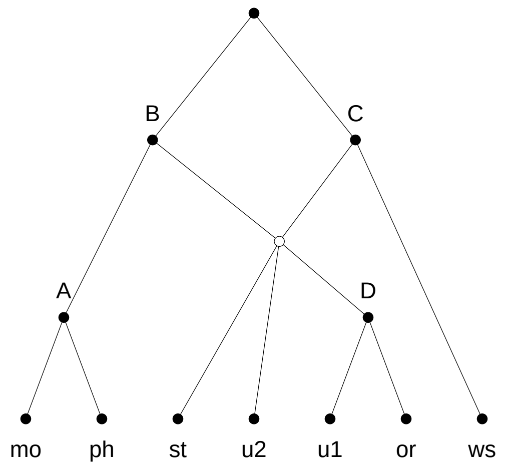
the subject was guessing. Finally, the number of transitivity violations is an indicator of how consistently the features were identified in different contexts (different triples of sounds).

Fig. 6 and 7 show the structures obtained for two subjects, with the corresponding labels (translated from Danish) 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 non-elevated sound, C referred to the width, and the 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 as a new feature. From the 19 subjects assigned to this elicitation method, 11 structures were obtained, fitting from 90 to 96% of their answers, i. e., from 4 to 10 changes out of 105 answers. The number of distinct features per



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

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



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."

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

structure ranged from 3 to 8.

3.3 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), *presence* (pr); four categories reflecting the timbral aspects—*brightness* (br), *spectral balance* (sb), *sharpness* (sh), *bass* (ba); and four categories not belonging specifically to spatial or timbral aspects—*naturalness* (na), *clarity* (cl), *loudness* (lo), *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 classifying either the RGT constructs or the clusters, are very similar, suggesting that the choice of approach has little impact on the outcome of the attribute selection.

Table 6: Number of occurences (Freq.) of PSA features, individual RGT constructs, and RGT clusters in the semantic categories (Cat.) which are explained in the text. The eight selected attributes, i. e. those having the highest positions in all three rankings, are indicated in boldface.

PSA		RGT		Clusters	
Cat.	Freq.	Cat.	Freq.	Cat.	Freq.
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	–

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.*

4 DISCUSSION

4.1 Repertory grid technique

Two groups of 20 subjects took part in the RGT, using either triadic (RGT-3) and pairwise (RGT-2) elicitation. While these two methods yielded a comparable 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 the 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 (i. e. 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 classify the heterogeneous clusters.[3] A reason for heterogeneous clusters might be the low reliability of the subjects' ratings, which is likely to be improved by including several repetitions of the ratings.

It should be noted that the constructs, reported here in English for the reader's convenience, were elicited in Danish. Such translations can be problematic, as the English word might not as accurately describe the actual sensation. There is empirical evidence [10] that adjectives might be used differently in describing the same sensation by native speakers of different languages. This problem is related to the above-mentioned assumption of a direct correspondence between the sensations and the verbal descriptors, which is addressed by PSA.

4.2 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 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*, while the other sounds could be identified as *wide*. In 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 which 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 which 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 changes 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 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.

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 sessions in which they were explicitly asked to identify features of the sounds arranged in a playlist. It can be concluded that a more controlled training session might be necessary, such as completing the exercise with drawings on their own, rather than it being demonstrated to them.

4.3 Concluding remarks

In search of a method to elicit auditory attributes in the context of multichannel reproduced sound,

[3] For instance, if one descriptor was not semantically related to the other descriptors in the same cluster, the cluster was classified so as to represent most of the descriptors.

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. For half of the subjects who performed in PSA, a feature structure could be obtained, suggesting that this method can be applied to the identification of features in complex stimuli. When the differences between the stimuli are subtle, however, the strong restrictions imposed on the judgments are often violated. A method to assess the validity of a structure in the presence of violations was proposed.

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.

5 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 comes 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 his helpful comments on an earlier version of the manuscript, and Jody Ghani for helping improve the English style.

References

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

[2] J. Berg and F. Rumsey, "Spatial Attribute Identification and Scaling by Repertory Grid Technique and Other Methods," in *Proceedings of the 16th AES International Conference on Spatial Sound Reproduction*, pp. 51–66 (1999).

[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.), preprint 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.*, 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).

[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.), preprint 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.*, in press (2005).

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

[16] ITU-R Rec. BS.1116, "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Telecommunication Union, Geneva, Switzerland (1997).

[17] AESTD1001.1.01-10, "Multichannel Surround Sound Systems and Operations," AES Technical Council document (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), preprint 6483.

Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference

Sylvain Choisel[1,2] Florian Wickelmaier[1]

[1] *Sound Quality Research Unit, Dept. of Acoustics, Aalborg University, 9220 Aalborg, Denmark*

[2] *Bang & Olufsen A/S, 7600 Struer, Denmark*

Abstract

Three experiments were conducted with the goal of quantifying auditory attributes which underlie listener preference in the context of multichannel reproduced sound. Short musical excerpts were presented in mono, stereo and several multichannel formats to a panel of 40 selected listeners. The first experiment aimed at an assessment of the overall preference between the reproduction modes by means of paired comparisons, and an exploratory analysis of the salient perceptual dimensions using multidimensional scaling. In the second experiment, individual auditory attributes were elicited and selected. In the third experiment, the selected attributes were quantified, and their contribution to overall preference was investigated. Scaling of preference and of the attributes was based on consistency tests of the paired-comparison judgments and on modeling the choice frequencies using probabilistic choice models. As a result, the preferences of non-expert listeners could be measured reliably at a ratio scale level. Principal components derived from the quantified attributes predict overall preference well. The findings allow for a careful generalization regarding the perception of and preference for certain spatial reproduction modes across musical program materials.

PACS numbers: 43.66.Lj, 43.66.Ba, 43.38.Md, 43.38.Vk

I. INTRODUCTION

One of the goals of research in sound quality is to understand the mechanisms underlying listener preference. Complex stimuli are typically involved in sound quality assessments, giving rise to various sensations, or *auditory attributes*, which potentially contribute to perceived overall quality. The identification and quantification of these sensations is necessary before their relation to preference can be established.

Apart from pioneering studies on multichannel recording and playback (Nakayama *et al.*, 1971), most work on quality of reproduced sound has focused on timbral aspects of monophonic reproduction (e. g., Gabrielsson and Sjögren, 1979). As multichannel audio formats are growing in popularity, the question arises how the various reproduction modes influence the listener's perception. Of particular interest is how spatial auditory sensations are affected by the introduction of center and surround loudspeakers in a multichannel setup, or by various processing algorithms. More recent studies have addressed the problem of identifying and quantifying auditory attributes which are relevant to sound quality in the context of multichannel reproduced sound (Rumsey, 1998; Berg and Rumsey, 1999; Zacharov and Koivuniemi, 2001; Guastavino and Katz, 2004). The first three employed combinations of recording and playback techniques to evoke various auditory sensations, and the latter used Ambisonics (Gerzon, 1985), a versatile recording and playback technique in which the sound signals are optimally decoded for each loudspeaker configuration.

By contrast, the present study aimed at investigating more specifically the perceptual differences between reproduction modes typically encountered in home audio systems: Selected musical excerpts—originally produced for five-channel reproduction—were reproduced in various formats (mono, stereo and several multichannel formats). In a recent study, Zieliński *et al.* (2003) have focused on the overall perceptual evaluation (the so-called *basic audio quality*) of reproduction modes similar to the ones used in the present work. Rumsey *et al.* (2005) investigated the influence of timbral, frontal and surround fidelity changes on basic audio quality. The present study, however, intended to seek explanations for such global differences in terms of more specific auditory attributes. More precisely, the goals of this study were to (1) identify the auditory attributes which are relevant in the context of multichannel music reproduction, (2) verify that listeners can judge upon them in a consistent manner, (3) quantify them on meaningful scales, and (4) determine their relation to overall preference.

The identification of the relevant auditory attributes has been reported elsewhere (Choisel and Wickelmaier, 2005), and the present paper places emphasis on the scaling of these attributes as well as overall preference. In the earlier investigations cited, sensation magnitudes were directly estimated using rating scales with either numerical or verbal labels, or graphical (visual analog) scales. Such direct scaling procedures are the defacto standard in sound quality assessments. As an example, consider the ITU-T recommendation P.800 (1996) for transmission quality, or the ITU-R recommendation for small (BS.1116, 1997) and intermediate (BS.1534, 2003) impairments in audio systems. The validity of such scales, however, relies on many implicit and untested assumptions.

First, it is usually assumed that the order of the scale values corresponds to an order of the sounds along the investigated attribute. This is problematic at least for multidimensional stimuli, because subjects might not be able to combine the different dimensions into a single one (e. g., overall quality). Classical studies on human choice behavior (May, 1954; Tversky, 1969) have demonstrated that already two or three dimensions lead to predictable inconsistencies. Focusing on different aspects depending on the stimuli being compared can result in intransitive judgments, such as preferring stimulus A over B, B over C, but C over A. It is evident that a preference order of the stimuli cannot be established in this case. While paired comparisons easily reveal these intransitivities, in direct scaling procedures problems associated with multidimensionality will go unnoticed, which casts doubt on the validity of such directly obtained scales. Very often researchers are interested not only in an order of the stimuli, but also in information about their differences or ratios, which requires measurements on higher scale types (interval or ratio scales; Stevens, 1946). The higher the scale level, the more restrictive forms of transitivity (as will be defined later in this paper) must be fulfilled.

Another assumption is the subjects' ability to map their sensation magnitude onto a scale. Often the freedom to choose among the many response categories in direct scaling procedures will result in an idiosyncratic strategy of scale usage. Some subjects might display a bias for certain response categories, for example the center or the endpoints of the scale. Binary paired comparisons, on the other hand, require nothing but simple comparative judgments, and thereby eliminate response biases due to scale usage.

Therefore, a major methodological objective of the present work was to use well-founded scaling techniques based on paired comparisons (so-called *probabilistic choice models*; Luce, 1959; Tversky, 1972), both for determining overall preference, and for measuring the strength of more basic auditory attributes. Such scaling methods have been successfully applied to sound quality evaluation, most notably to auditory unpleasantness (Ellermeier *et al.*, 2004; Zimmer *et al.*, 2004). The present study consists of three experiments. The first, exploratory experiment, aimed at uncovering perceptual dimensions by means of multidimensional scaling (MDS) based on dissimilarity judgments, as well as quantifying overall preference between the different reproduction modes. In the second experiment, auditory attributes relevant in the context of the selected sounds were identified using three different elicitation methods. The details of these methods and the main results have been reported elsewhere (Choisel and Wickelmaier, 2005). In the third experiment, selected attributes were quantified using paired comparisons, thereby verifying that listeners could map them on a unidimensional scale. Finally, their relation to overall preference was formulated in a multiple-regression model.

II. METHOD

A. Subjects

Forty listeners (28 males, 12 females) took part in this study. They were selected among 78 candidates, according to their auditory and verbal abilities. The selection procedure (detailed in Wickelmaier and Choisel, 2005) consisted of pure-tone audiometry, a stereo-width discrimination task and a verbal fluency test. This was done in order to ensure that the listeners selected could (1) appreciate spatial differences in sound and (2) readily produce a description of their sensations. All candidates were native Danish speakers, without any known hearing problems. Eight listeners showing a hearing threshold of more than 20 dB HL (re. ISO 389-1, 1998) in any ear at any frequency between 250 Hz and 8 kHz were rejected based on this criterion. From the remaining 70, the 40 listeners performing best in the other two tests (stereo-width discrimination and verbal fluency) were selected to participate in the main experiments. Their age ranged from 21 to 39 (median = 24). One of the participants dropped out during the first experiment, the remaining 39 took part in the complete study which extended over approximately six months.

B. Apparatus and stimuli

1. Program material

Four musical excerpts (two pop, two classical) were selected from commercially available multichannel material (Table I), and recorded from their original medium, Super Audio Compact Disc (SACD) or Digital Versatile Disc Audio (DVD-A), onto a computer, with a sampling frequency of 48 kHz and 24-bit resolution. The excerpts were carefully cut to include a musical phrase, their duration ranging from 4.7 to 5.4 s. The two classical recordings were made with the same microphone technique (five omnidirectional microphones placed in a circular array), and the two pop recordings were mixed with standard surround panning techniques.

2. Playback setup

The listening tests took place in a 60-m^2 sound-insulated listening room complying with the ITU-R BS.1116 (1997) requirements. Seven loudspeakers (Genelec 1031A) were placed as shown in Figure 1, at a distance of 2.5 m to the listening position. The height of the tweeters was 108 cm above the floor, which was the average height of the entrance of the listeners' ear canals. Five of the seven loudspeakers were arranged in accordance with the ITU-R recommendation BS.775-1 (1994); two additional speakers were placed at ±45° for the reproduction of stereo over a wider base angle (defined as the bearing angle between the loudspeaker pair, as seen from the listening position). The setup was hidden from the subject by an acoustically transparent curtain.

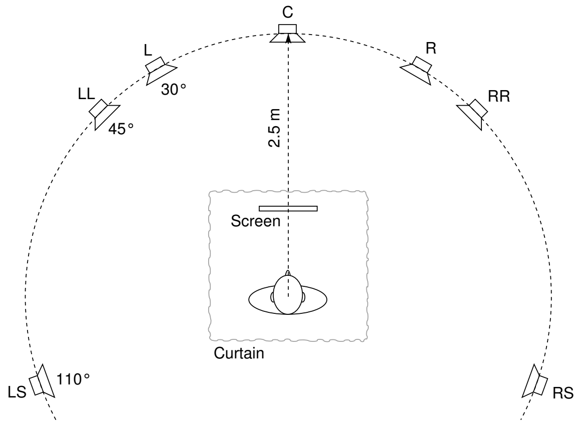


Figure 1: Playback setup consisting of seven loudspeakers: left (L), right (R), center (C), left-of-left (LL), right-of-right (RR), left surround (LS) and right surround (RS). This setup was symmetrically placed with respect to the width of the room and was hidden from the subject by an acoustically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

The sounds were played back by a computer placed in the control room, equipped with a multichannel sound card (RME Hammerfall HDSP) connected to an 8-channel D/A converter (RME ADI-8 DS) having a flat frequency response from 5 Hz to 21.5 kHz.

The response interface consisted of an optical mouse and a 15" flat screen placed in front of the listener, below loudspeaker level (45 cm above the floor) in order to limit interactions with the sound field. A head-rest fixed to the armchair ensured that the subject's head was always centered during the listening test. The head position could be monitored from the control room, via a camera attached to the ceiling above the listener.

For each of the seven loudspeakers, the free-field frequency response was measured with a 14th order maximum length sequence (MLS) at a sampling frequency of 48 kHz. Based on these measurements, seven FIR filters were calculated to equalize the minimum phase component of the frequency responses (Oppenheim and Schafer, 1989, pp. 781–785). The deviations of the resulting frequency responses from a flat response were within ±1.5 dB across the frequency range from 50 Hz to 20 kHz. The deviations from the mean of the seven equalized loudspeaker responses were within ±0.3 dB at frequencies above 70 Hz, slightly larger differences being observed at lower frequencies. In order to verify the inter-channel level alignment in the listening room after equalization, the A-weighted sound-pressure level of band-passed pink noise (200 Hz–2 kHz) was measured at the listening position for each channel. As a result, the inter-channel level differences were within 0.3 dB, and the differences between left/right pairs did not exceed 0.1 dB.

3. Reproduction modes

From the original five-channel program material (*or*), several formats were derived, as summarized in Table II. First, the original was mixed down to stereo (*st*) according to the ITU-R recommendation BS.775-1 (1994):

$$\begin{aligned} L_{st} &= L_{or} + \tfrac{1}{\sqrt{2}}\, C_{or} + \tfrac{1}{\sqrt{2}}\, LS_{or} \\ R_{st} &= R_{or} + \tfrac{1}{\sqrt{2}}\, C_{or} + \tfrac{1}{\sqrt{2}}\, RS_{or} \end{aligned} \qquad (1)$$

From the stereo version, mono (*mo*) and phantom mono (*ph*) were computed as described by Equation 2 and 3, respectively.

$$C_{mo} = \tfrac{1}{\sqrt{2}}\,(L_{st} + R_{st}) \qquad (2)$$

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

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

Finally, three upmixing algorithms were used to re-construct a multichannel sound from the stereo downmix; two commercially available algorithms, Dolby Pro Logic II and DTS Neo:6—later referred to as upmixing 1 and 2 (*u1* and *u2*), in no specific order—and a simple matrix upmixing algorithm. Dolby Pro Logic II was implemented on a Meridian 861 surround processor. The processors were fed with a digital signal (S/PDIF) from the RME sound card, and the five analog output signals were recorded through the RME converter, using 24-bit resolution. In a similar fashion, a Yamaha RX-V 640 receiver was used to generate the DTS Neo:6 upmix.

The matrix upmixing (*ma*) was inspired by matrix decoding systems which are typically applied to encoded stereo tracks (cf. Rumsey, 2001). In this study, however, it was applied to a "regular" stereo downmix (Equation 1). The upmixing was implemented in Matlab in the following way: 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 3 dB. The L and R channels were left unchanged.

The eight reproduction modes were matched in loudness by eight subjects (not taking part in the main experiments) using a forced-choice adaptive procedure (2AFC, 1-up/1-down, cf. Levitt, 1971; Jesteadt, 1980). On each trial, the task was to decide which of the two presented sounds was louder, one being the standard, the other one being the comparison, in random order. For all four types of program material (Beethoven, Rachmaninov, Steely Dan and Sting), the standard was chosen to be the stereo reproduction mode. Its playback level was adjusted beforehand to a comfortable level by the experimenters, and measured in the listening position to have A-weighted, energy-

Table I: 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 – 1'56
Rachmaninov: Vespers – St. Petersburg Chamber Choir conducted by Korniev	Blazen Muzh	SACD	03	2'04 – 2'09
Steely Dan: Everything Must Go	Everything Must Go	DVD-A	09	0'52 – 0'57
Sting: Sacred Love	Stolen Car	SACD	06	1'55 – 2'00

Table II: Reproduction modes: full name, abbreviation and loudspeakers used for playback (see Figure 1).

Name	Abbr.	Speakers
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 as *u1* and *u2* (in no specific order) in the rest of this paper.

equivalent sound pressure levels of 65.8, 59.4, 66.5 and 67.7 dB, respectively (averaged over the duration of the stimuli). The loudness matching procedure was reported in more details by Choisel and Wickelmaier (2005).

The resulting matches were averaged across subjects, and appropriate gains were applied to the 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.

C. Procedure

1. Experiment I: Dissimilarity and preference

The first experiment aimed at uncovering the perceptual dimensions in the context of multichannel sound by means of multidimensional scaling (Borg and Groenen, 1997), and at quantifying listeners' preference for the various reproduction modes.

In the first part of this experiment, dissimilarity ratings were collected for every pair of reproduction modes within each type of program material. Two sounds were presented with an inter-stimulus interval of 500 ms, and the subject was asked "How dissimilar are these two sounds?" A response was given by making a mark on a line on the computer screen. The end-points of this line were labeled "ens" (Danish for *similar*) and "forskellig" (Danish for *dissimilar*).[1] All possible pairs were presented once to each subject, in only one order. The position of each sound in the pair was balanced for each subject, in such a way that each reproduction mode was presented an equal number of times in the first and second position. After a dissimilarity judgment was made, the next stimulus pair was played. It was not possible for the subject to listen to the sounds more than once. For each of the four types of program material, 28 pairs were presented in random order to each subject within one experimental block. The order of the blocks (and thus of the program materials) was balanced across subjects. Each block was preceded by a short familiarization in which the subject could listen to all the sounds as often as needed in order to have an impression of the range of variation between the different reproduction modes.

In the second part of this experiment, an attempt was made to quantify the overall preference for each reproduction mode using a paired-comparison procedure. For each pair of reproduction modes the subjects were instructed to indicate which one they preferred. Two buttons, labeled A and B, were visually emphasized in turn (by changing their size) during playback to indicate which sound was playing. The response was made by clicking the button corresponding to the preferred sound. Each pair was presented three times: twice in opposite order (AB BA), and a third time in a balanced order across subjects, in order to ensure that each pair was presented equally often in both orders. The between-pair order was random. In total, 84 preference judgments were given by each subject for each type of program material.

The data collection was split into two sessions taking place on two different days. In the first session, the subjects performed the dissimilarity task for the four types of program material with short breaks between the blocks. After a longer break (5 to 10 minutes), they performed the preference task on one type of program material. The remaining three program materials constituted the three blocks of the second session. The duration of a session was approximately one hour. The order of the program materials was balanced across subjects for both tasks.

2. Experiment II: Elicitation of auditory attributes

In order to identify the relevant auditory attributes for the sounds under study, three elicitation methods were employed: two based on the repertory grid Technique (RGT, Berg and Rumsey, 1999; Kelly, 1955) and an indirect elicitation method which strictly separates the identification of auditory features from their labeling (Wickelmaier and Ellermeier, 2005; Heller, 2000).

The details of the experimental procedures and the main results of the elicitation are reported in Choisel and Wickelmaier (2005). The auditory attributes were elicited individually and combined across subjects into 20 categories common to the three elicitation methods. The eight categories occurring most often in all three methods were selected as the list of attributes to be quantified in Experiment III. Those were *width, brightness, spaciousness, elevation, distance, envelopment, naturalness* and *clarity*.

3. Experiment III: Quantification of selected attributes and preference

For each of the selected attributes, a paired-comparison procedure was used, identical to that employed in Experiment I. For each pair of reproduction modes, the subject was asked (in Danish) "Which of the two sounds is more..." followed by one of the following adjectives: *wide (bred), elevated (høj oppe), spacious (rummelig), enveloping (omsluttende), far ahead (langt foran), bright (lys), clear (tydelig)* and *natural (naturlig)*. Definitions of these attributes (see Appendix) were generated by the experimenters so as to represent as much as possible the subjects' descriptors elicited in Experiment II. Each pair was judged only once. Each attribute was evaluated for all four program materials in a single block lasting for about 25 minutes. Two attributes were evaluated per day and subject, in a one-hour session including a break in the middle. Thus, four sessions were required for all eight attributes. The order of the attributes and program materials was balanced across subjects using five different 8×8 Graeco-Latin squares.

Finally, preference was quantified again in order to investigate a possible influence of training since the first experiment. The procedure was identical to that in Experiment I, except that the number of replications was reduced to two. The 56 preference judgments per program material and subject were collected in a fifth session.

D. Statistical analysis

1. Multidimensional scaling (MDS)

The pairwise dissimilarity ratings of the reproduction modes made by each subject for each type of program material were organized in dissimilarity matrices which served as input for a multidimensional scaling algorithm (INDSCAL, Carrol and Chang, 1970). MDS is widely used in the context of sound quality evaluation to uncover perceptual dimensions (e. g., Martens and Zacharov, 2000). INDSCAL maps the sounds into a common multidimensional space, such that the inter-point distances in that space match the perceived dissimilarities as closely as possible. The number of dimensions of such a space does not follow readily from an INDSCAL analysis, but was determined by the correlation between dissimilarities and distances (R^2) and the degree of interpretability of the obtained dimensions. The INDSCAL analysis was repeated for each type of program material.

2. Analysis of the choice frequencies

Both, for the overall preference and for the selected auditory attributes, the pairwise choices among the eight reproduction modes were aggregated across all listeners, resulting in matrices of the choice frequencies. In such a matrix it can be seen how often, for example, mono (*mo*) reproduction was chosen to be more *spacious* than stereo (*st*) and vice versa. From these frequencies the probability, P_{xy}, of choosing sound x over sound y according to a given criterion was estimated.

Derivation of scales from the choice frequencies crucially depends on the consistency of the judgments given by the subjects. Consistency was analyzed by testing weak (WST), moderate (MST), and strong (SST) stochastic transitivities, which imply that if $P_{xy} \geq 0.5$ and $P_{yz} \geq 0.5$, then

$$P_{xz} \geq \begin{cases} 0.5 & \text{(WST)} \\ \min\{P_{xy}, P_{yz}\} & \text{(MST)} \\ \max\{P_{xy}, P_{yz}\} & \text{(SST)} \end{cases} \qquad (4)$$

for all sounds x, y and z. Whenever the premise holds, but the implication in Equation 4 does not hold (for any permutation of the triple x, y, z), a transitivity violation is observed. Violations of the different transitivities are of different severity. A systematic violation of WST indicates that the subject was not able to integrate several stimulus dimensions into one percept, and it is therefore impossible to even derive a meaningful ordering of the sounds. Less severe are violations of SST which suggest a certain context dependency of the choices made. Such a context dependency usually comes into play when there are subgroups of similar sounds based on multiple perceptually salient aspects or features (Carrol and De Soete, 1991).

Counting the number of transitivity violations in a matrix of choice frequencies only yields a descriptive measure of (in)consistency. In an experiment with a limited number of observations, it is conceivable that violations occur at random; a statistical test is therefore required to classify such violations as either systematic, and thus critical, or random.

Two kinds of probabilistic choice models were considered for representing the choice frequencies, with the goal of (1) providing statistical evaluation of the transitivity violations encountered and (2) in the presence of only random violations, quantifying the attribute in question. The first model used was the Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 1959), which predicts P_{xy} as a function of parameters associated with each sound

$$P_{xy} = \frac{u(x)}{u(x) + u(y)}, \qquad (5)$$

where $u(\cdot)$ is a ratio scale of the criterion. Since Equation 5 implies SST, systematic violations of SST preclude a BTL representation.

The second, less restrictive, model was the so-called *elimination-by-aspects* (EBA) model (Tversky, 1972; Tversky and Sattath, 1979), which is a generalization of the BTL model. According to EBA, one sound is chosen over a second one because of a certain *aspect* which belongs to the first but not to the second sound. EBA predicts P_{xy} by

$$P_{xy} = \frac{\sum\limits_{\alpha \in x' \setminus y'} u(\alpha)}{\sum\limits_{\alpha \in x' \setminus y'} u(\alpha) + \sum\limits_{\beta \in y' \setminus x'} u(\beta)}, \qquad (6)$$

where $\alpha, \beta, \ldots$ are the aspects (or features) of the sounds, and $x' \setminus y'$ denotes the set of aspects belonging to sound x but not to sound y. As for the BTL model, $u(\cdot)$ is a ratio scale of the criterion. EBA only implies MST, and can therefore to some extent cope with multiple-aspect criteria.

The goodness of fit of the choice models was evaluated by comparing the likelihood L_0 of a given (restricted) model to the likelihood L of a saturated (unrestricted) binomial model which perfectly fits the choice frequencies, under the assumption of independent choices. The test statistic, $-2 \log L_0/L$, is approximately χ^2-distributed with as many degrees of freedom as the difference in parameters of the two models. A significant likelihood ratio test indicates lack of fit of the restricted choice model, and thereby that the violations of the corresponding stochastic transitivity have been systematic rather than random. If the fit was adequate, scale values for the reproduction modes were derived. Parameter estimation and model testing were performed using software described in Wickelmaier and Schmid (2004).

Probabilistic choice models provide a powerful method for scaling supra-threshold sensations, not only because they allow for *testing* the validity of a scale of a certain attribute (rather than *assuming* it when using direct scaling procedures), but also because these models enable the investigator to test hypotheses about perceived magnitudes in the framework of standard statistical theory. In order to test whether there was a significant change in the scale values of the reproduction modes in different conditions, for example whether the preference changed between the two times of data collection (Experiment I and III), standard likelihood ratio tests were performed. The logic of these tests is to investigate if restricting the parameters to be equal in both conditions entails a significant lack of fit, which implies that the conditions have a significant effect on the scale values. This would mean in the example that the preferences have changed from the first to the third experiment.

III. RESULTS

A. Scaling listener preference

Table III displays the evaluation of the stochastic transitivities (Equation 4) of the preference judgments collected in the first experiment. For the evaluation, data were aggregated over all subjects and the three repetitions, within each type of program material. Thus, the choice probabilities were estimated based on $N = 120$ observations per stimulus pair for Steely Dan, and $N = 117$ for the other program materials, since one subject had left the panel after the first session. Weak and moderate stochastic transitivity were found to be violated either in none or in very few of the 56 possible tests, indicating that the participants were able to integrate their various sensations into a unidimensional preference judgment. Consequently, at least an ordinal preference scale may be derived from the choice frequencies.

In order to evaluate the more frequent violations of SST, and to test whether a preference *ratio* scale could be obtained, a BTL model (Equation 5) was fitted to the paired-comparison data. Table III shows the results of the goodness-of-fit tests which

Table III: Transitivity violations and goodness-of-fit test of the BTL model for preference judgments in Experiment I. Columns 2 to 4 show the number of violations of weak, moderate, and strong stochastic transitivity in 56 possible tests. Columns 5 and 6 show test statistic and p-value of a likelihood ratio test (see text).

Excerpt	WST	MST	SST	$\chi^2(21)$	p
Beethoven	0	2	14	9.13	.988
Rachmaninov	2	4	19	16.96	.714
Steely Dan	0	0	12	18.13	.640
Sting	0	0	13	10.72	.968

support the validity of the model in each of the four program material conditions. Accordingly, the SST violations were classified as random, and preference scales were extracted.

Preference was re-assessed in the final part of Experiment III, approximately six months after data collection for Experiment I, in order to investigate the stability of the listeners' evaluation. The choice probability estimates were based on $N = 78$ observations per stimulus pair (each subject contributed two judgments per cell). Table IV shows the number of transitivity violations and the goodness-of-fit test of the BTL model. As in the first experiment, there was no compelling evidence in the data against the BTL model to hold; therefore the occasional transitivity violations may be considered unsystematic. Consequently, listener preference can be measured on a ratio scale level using the very simple, but very restrictive BTL model.

The reliability of the judgments was assessed by testing whether there were any changes of preference between the three and two repetitions, respectively, *within* each experiment. Likelihood ratio tests were devised to compare a BTL model which allows for preference changes to one with a fixed set of parameters across repetitions. Neither in the first nor in the third experiment, however, did the fixed-parameter model fit significantly worse than the model having variable parameters; this was true for all types of program material. Therefore, the preference values of the reproduction modes can be regarded constant throughout the repetitions within each experiment. This indicates a high degree of reliability of the preference judgments.

Figure 2 displays the parameter estimates of the BTL model, i. e., the *preference scales*, for the four program materials obtained in the two experiments, together with the 95%-confidence intervals. The preference ratio scales are plotted on logarithmic y-axes in order to facilitate the comparison among reproduction modes. For example, two-channel phantom mono (*ph*) was preferred about twice as much as single channel-mono (*mo*) for the Beethoven excerpt. About the same ratio was observed between wide-angle stereo (*ws*) and one of the upmixing algorithms (*u2*). Since the BTL parameters are unique up to multiplication by a positive constant, they were normalized to sum to unity. Consequently, the distance from the line of indifference ($u = 1/8$, which would be the location of the scale values if all pairwise choice frequencies were 0.5) indicates how pronounced the preferences are between the reproduction modes. In all conditions, equality of the scale values can be rejected, which suggests that listeners were far from indifferent, but had rather strong preferences for certain reproduction modes.

Table IV: Transitivity violations and goodness-of-fit test of the BTL model for preference judgments in Experiment III. See Table III.

Excerpt	WST	MST	SST	$\chi^2(21)$	p
Beethoven	0	1	12	9.06	.989
Rachmaninov	0	0	18	8.44	.993
Steely Dan	0	0	11	17.74	.666
Sting	0	0	9	13.66	.884

Across the program materials and the two experiments it was observed that mono reproduction (*mo* and *ph*) was inferior to the other formats. Stereo on the other hand, was generally among the most preferred, whereas the original five-channel material outperformed stereo only once (Steely Dan). A further interesting relation was observed between *ws* and matrix upmixing (*ma*) when comparing classical and pop music. While it was beneficial for the classical music to increase the stereo base angle from 60° to 90°, this had adverse effects for the pop excerpts. Conversely, while *ma* was less preferred than *ws* for the classical music, it was favored over *ws* for the pop music. The results so far indicate common preference patterns at least within a musical genre, but also excerpt-specific effects.

In a further set of likelihood ratio tests it was investigated to what extent the preference scales were generalizable across programs materials. In spite of the obvious similarities within the classi-

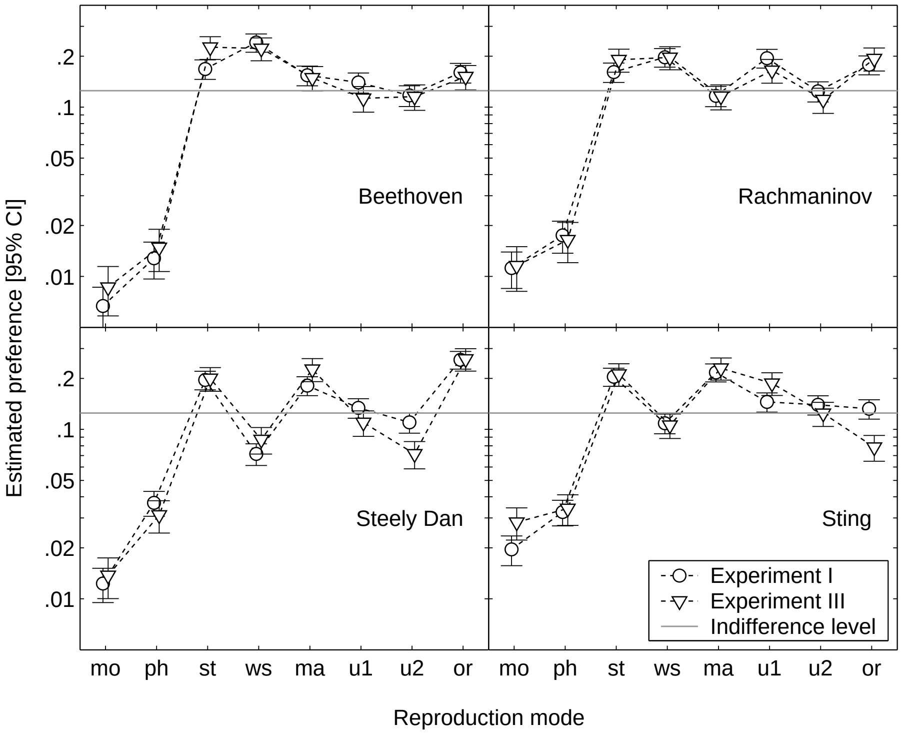


Figure 2: Ratio scale of preference for eight reproduction modes derived from paired-comparison judgments using the BTL model. The reproduction modes were mono (*mo*), phantom mono (*ph*), stereo (*st*), wide-angle stereo (*ws*), four- (*ma*) and five-channel upmixing (*u1* and *u2*), and the original five-channel material (*or*). Error bars show 95%-confidence intervals.

cal and the pop genres (see the rows in Figure 2), the excerpt-specific differences were statistically significant. In the first experiment, a common model for Beethoven and Rachmaninov fared significantly worse [$\chi^2(7) = 30.81$; $p < .001$] than a model having two sets of parameters, one for each program material. The same was true for Steely Dan and Sting [$\chi^2(7) = 78.15$; $p < .001$]. Analog results were obtained in Experiment III for classical [$\chi^2(7) = 21.14$; $p = .004$] and pop music [$\chi^2(7) = 146.01$; $p < .001$], respectively. From the magnitudes of the test statistics it seems that the differences between the classical excerpts were not as striking as between the pop excerpts, and the difference between Steely Dan and Sting even increased in the third experiment. Therefore, the generalizability of the results concerning the preference for certain reproduction modes should not be overestimated, since the dependence on the program material is evident.

Since preference data were collected twice for the same listeners, once *before* the elicitation and quantification of the more specific attributes and once *after* that, the effect of experience with the sounds on preference may be examined. Figure 2 suggests that there is a close correspondence between the preference scales obtained at the two points in time, indicating that preference was relatively stable even over a period of about six months. Again, likelihood ratio tests were employed for the statistical analyses. This time, it was tested for each type of program material whether the preference scale had changed between the first and the third experiment. No significant changes were observed for Beethoven [$\chi^2(7) = 12.33$; $p = .090$] and Rachmaninov [$\chi^2(7) = 5.90$; $p = .551$], whereas for Steely Dan [$\chi^2(7) = 25.37$; $p = .001$] and Sting [$\chi^2(7) = 35.80$; $p < .001$] the changes were significant. These differences might be attributed to listeners becoming more sensitive to subtle differences between the reproduction modes. For example, there were no significant preference differences between the two upmixing algorithms (*u1* and *u2*) and the original five-channel Sting material *or* in the first experiment (see bottom right panel in Figure 2). In the third experiment, however, the ratio between *u1* and *or* extended to about 3:1. A similar argument holds for the *ma* and *u2* reproduction modes of the Steely Dan excerpt (see bottom left panel in Figure 2).

B. Scaling auditory attributes

The same logic of consistency checks, model evaluation, and scaling was applied to the more elementary auditory attributes elicited in Experiment II, and evaluated in Experiment III. Table V displays the violations of the stochastic transitivities for each auditory attribute and program material. Since the pairwise probability estimates were based on 39 observations (every listener judged each pair only once) it was expected to see more (random) violations than for the preference judgments. From the low number of WST violations it follows that at least an ordinal scale of sensation magnitude can be derived in each condition. In order to test for systematic SST violations, a BTL model was applied and evaluated in each case. As shown in Table V, in general the model fit is adequate which suggests that consistency in the judgments was sufficiently high for extracting *ratio* scales. Additional likelihood ratio tests were devised to confirm that each scale was significantly different from the case where all scale values are equal. These tests indicated that for no attribute-excerpt combination did listeners show indifference with respect to the reproduction modes.

In only two cases (Steely Dan: *envelopment* and *width*) was there a significant lack of fit of the BTL model. This should not compromise the overall conclusion that the listeners' choice behavior could be described by a simple model, since one might expect about two tests out of 32 to become significant by chance alone on an α-level of 5%. The original Steely Dan material, however, was different from the other three excerpts in that it not only contained reverberation but clearly distinct sound sources (e. g., a guitar playing a staccato single-note line) in the surround channels, which might have given rise to a more complex decision strategy. Potentially, the emergence of such a new feature might be more adequately described by an elimination-by-aspects (EBA) model (Equation 6). Among the EBA models with only one additional

Table V: Transitivity violations and goodness-of-fit test of the BTL model for selected attributes. See Table III. Note: *$p < .05$

Attribute	WST	MST	SST	$\chi^2(21)$	p
Beethoven					
width	0	1	19	24.55	.267
elevation	1	11	25	24.63	.263
spaciousness	0	2	18	17.80	.661
envelopment	0	3	23	22.16	.391
distance	3	9	32	22.83	.353
brightness	2	3	19	12.25	.933
clarity	4	5	27	25.55	.224
naturalness	3	5	24	15.41	.802
Rachmaninov					
width	1	1	14	21.20	.447
elevation	2	7	23	16.08	.765
spaciousness	2	7	19	7.35	.997
envelopment	2	4	27	16.82	.722
distance	2	11	37	21.74	.414
brightness	4	4	27	14.49	.848
clarity	2	6	21	8.86	.990
naturalness	0	2	14	16.46	.744
Steely Dan					
width	0	3	14	36.01	.022*
elevation	0	2	24	30.64	.080
spaciousness	2	2	19	26.66	.182
envelopment	0	2	23	39.40	.009*
distance	3	13	30	15.89	.776
brightness	0	0	15	20.39	.496
clarity	0	2	18	14.05	.867
naturalness	0	2	18	14.35	.854
Sting					
width	0	2	24	29.47	.103
elevation	0	0	16	27.30	.161
spaciousness	0	4	16	22.60	.366
envelopment	1	3	16	15.04	.821
distance	0	1	19	21.40	.435
brightness	0	0	23	31.54	.065
clarity	2	3	16	19.24	.570
naturalness	1	1	18	11.72	.947

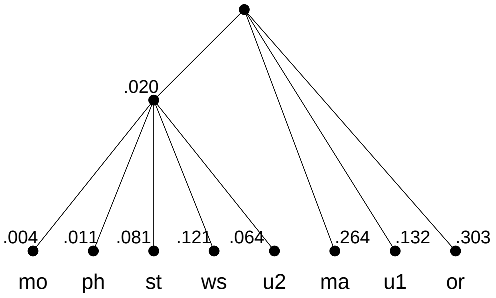


Figure 3: Elimination-by-aspects (EBA) model structure and parameter estimates for *envelopment* (Steely Dan). Nodes represent features shared only by the connected reproduction modes.

parameter, the best fitting one is depicted in Figure 3. The nodes in the graph denote the features, or *aspects*, of the reproduction modes. Apart from the top node (the aspect shared by all sounds) and the bottom nodes (the unique, individual aspects), the model includes one extra feature shared by all reproduction modes which do *not* reproduce any discrete source at the side of or behind the listener.[2] This simple EBA model was found to fit the data [$\chi^2(20) = 26.55$; $p = .148$] being significantly better than the BTL model [$\chi^2(1) = 12.85$; $p < .001$]. The parameter estimates are also displayed in Figure 3. In order to derive *envelopment* scale values from the model, the parameters belonging to each reproduction mode were added up. Similarly, an EBA model was found for the *width* attribute, which accounted for the data [$\chi^2(20) = 27.27$; $p = .128$] and outperformed the BTL model [$\chi^2(1) = 8.74$; $p = .003$]. Here, four reproduction modes (*st*, *ws*, *ma*, and *or*) shared a common aspect, the interpretation of which is not so straightforward. It is worth noting that, even though these EBA models provided a better fit than the BTL model, the differences in the actual scale values were rather subtle.

Figure 4 shows the derived ratio scales for each auditory attribute and the four types of program material. Within each attribute, considerable similarity of the scales was observed across program materials, which was even more pronounced within musical genre (classical and pop music). For example, *ws* was perceived to be strongly elevated in comparison with the other reproduction modes in the pop material (Steely Dan and Sting); the effect was less distinct, but still visible, for the classical material. The stimuli showed the smallest perceptual differences with respect to *distance*; the mono sounds (*mo* and *ph*) were perceived to be near-

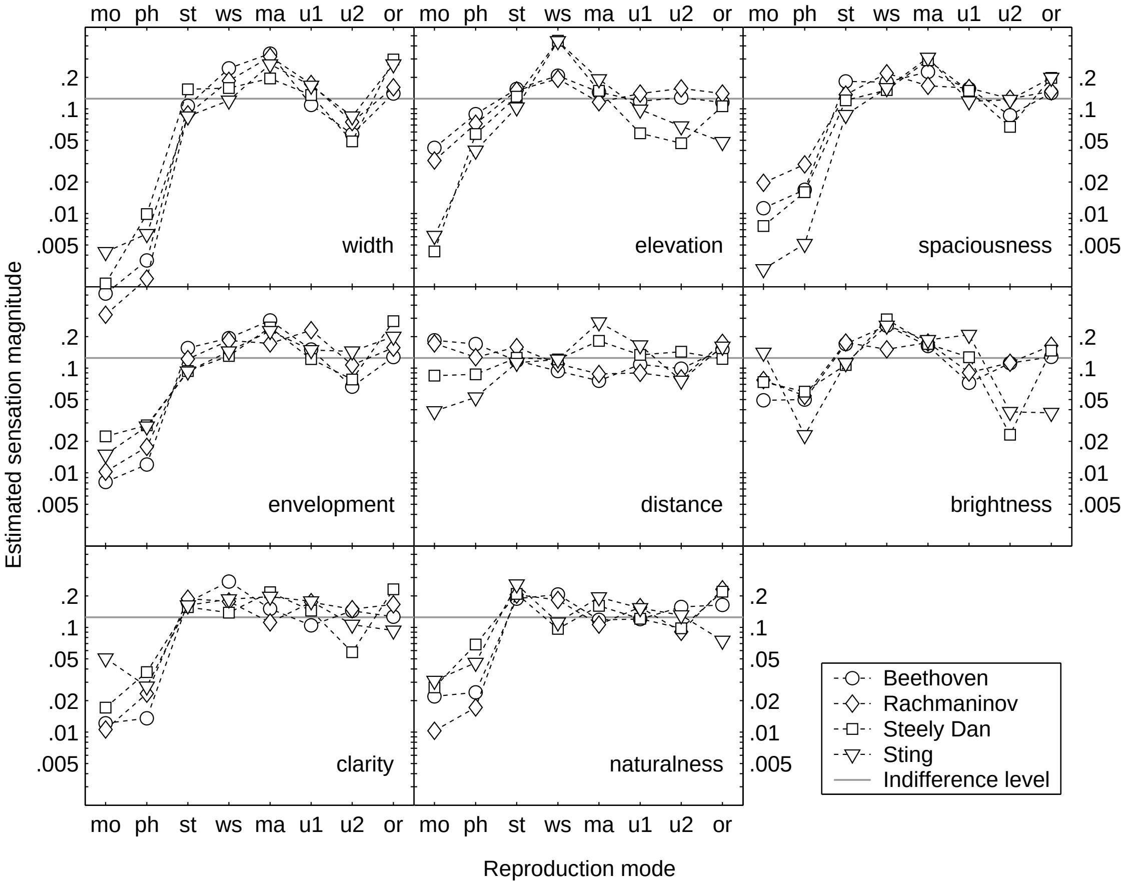


Figure 4: Ratio scales of eight auditory attributes estimated using BTL and EBA models for four types of program material.

est to the listener only for the pop music, for the classical music they were further away than most of the other reproduction modes. Except for distance and brightness, *mo* and *ph* were located at the lower end of the sensation scales, which induces correlation also *across* the attributes. Especially the correspondence between *spaciousness* and *envelopment* is striking. Clearly, these attributes did not vary independently in the stimuli under study.

C. Identification of perceptual dimensions

In order to uncover salient sensations which played a role in judging the similarity of the various reproduction modes, multidimensional scaling was performed—based on the data collection in Experiment I—using INDSCAL. As a representation of the perceptual space, a two-dimensional solution was chosen for each type of program material (Figure 5). The amount of variance explained in the dissimilarity ratings (ranging from 43% for the Sting excerpt to 77% for Rachmaninov) did not increase noticeably with more than two dimensions. The relatively poor fit of the MDS solutions is likely due to the fact that the dissimilarity rating was the first experimental task the participants encountered, thus they had little experience with the sounds and produced "noisy" judgments (each listener judged each stimulus pair only once). In addition, INDSCAL takes the individual differences in the ratings into account and is therefore known to result in lower R^2 values than one would obtain when averaging the dissimilarity judgments before performing an MDS analysis (Carrol and Chang, 1970). Most importantly, the attribute scales obtained did not provide an explanation of more than two MDS dimensions (see below). There-

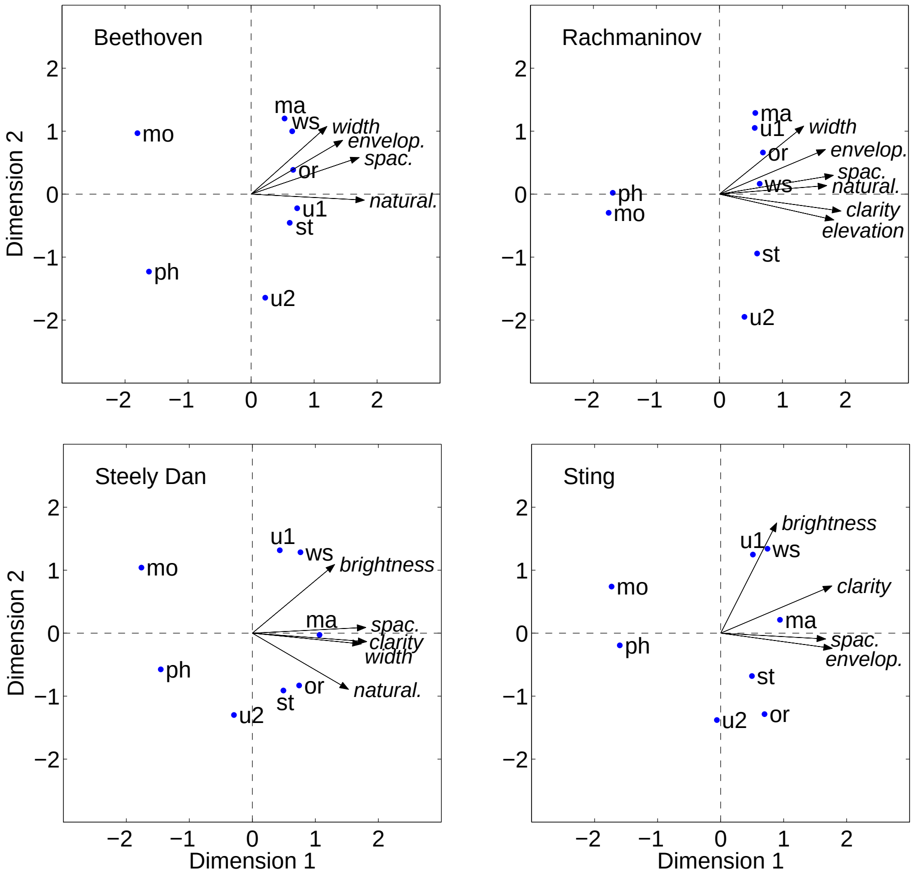


Figure 5: Multidimensional scaling of the eight reproduction modes. A two-dimensional INDSCAL solution was chosen for each of the four types of program material. The attributes significantly related ($p < .05$) to the 2D MDS solution are represented by arrows, with endpoints defined by two times their standardized regression coefficient in each dimension.

fore, in absence of a strong reason to introduce a third dimension, the two-dimensional solutions were judged sufficient for a first exploratory data analysis.

Some similarities can be observed between the programs of the same musical genre (classical and pop), and also to some extent between all four types of program material (Figure 5). Note that some of the MDS axes have been flipped to reveal these similarities. In all four maps, the two mono sounds (*mo* and *ph*) are separated from the non-mono stimuli along Dimension 1, while Dimension 2 differentiates upmixing 2 and stereo on the one hand, from wide-stereo on the other hand. The position of the original (*or*) and the matrix upmixing (*ma*) formats seems to be affected by the type of musical genre, but is almost identical between Beethoven and Rachmaninov, and between Steely Dan and Sting.

In order to attempt an interpretation of the two dimensions, the attribute scales obtained in Experiment III were projected in the MDS space (Figure 5). This was done by multiple regression of each attribute scale on the two MDS dimensions (cf. Hollins *et al.*, 2000). The coordinates of the arrows are calculated as the standardized regression coefficients of the corresponding attributes, multiplied by two for better readability. Whether an attribute is significantly related to the MDS space was determined from the overall F-test of the regression model; only attributes for which $p < .05$ are represented on Figure 5. Although there is not always a direct correspondence between attributes and dimensions, for the pop music the first dimension seems to be concerned with the spatial character of the sound (*width*, *spaciousness* and *envelopment*) as well as *clarity*, while the second dimension correlates most with *brightness*. For Beethoven and Rachmaninov, however, such a clear distinction cannot be made because of the high correlation between the attributes. The attribute *width* appears to "load" equally on Dimension 1 and 2, while *naturalness* and *spaciousness* are more closely related to Dimension 1. For the two classical excerpts, *brightness* does not relate significantly to the MDS space.

D. Relation between specific sensations and overall preference

Overall quality (or preference) has often been related to perceptual dimensions, specific subjective attributes, or to objective parameters using a multiple regression approach (Nakayama *et al.*, 1971; Susini *et al.*, 1999; Zacharov and Koivuniemi, 2001; Mattila, 2002; Rumsey *et al.*, 2005). In the present case, however, the low number of reproduction modes (only eight data points to be predicted) compared to the number of possible predictors (eight attributes), makes such modeling trivial and of questionable generality. An additional concern is the high correlation between some of the attributes; collinearity of the independent variables in a regression yields unstable and therefore unreliable results. To circumvent these problems, principal component analysis (PCA) with varimax rotation was used to reduce the attribute space to fewer independent factors (or *components*). In order to increase the generalizability of the model, the data were aggregated within musical genre, i. e., classical music (Beethoven and Rachmaninov) and pop music (Steely Dan and Sting), thereby doubling the number of data points to be predicted. This was justified given the similarities observed in the attribute scales across program materials (see Figure 4). *Naturalness* was excluded from the analysis because it was considered more global than the other (specific) attributes and not sufficiently separate from preference, the correlation coefficients between *naturalness* and preference ranging from 0.94 (Steely Dan) to 0.98 (Rachmaninov).

The PCA was performed on the remaining seven attributes. In the case of the classical music, 87% of the variance in the scale values was explained by the first two factors which, after rotation, accounted for 48 and 39% of the variance, respectively. For the pop music, the first two compo-

Table VI: Attribute loadings on the factors ($F1$ and $F2$) obtained from principal component analysis, and variance explained by these factors after varimax rotation. Loadings higher than 0.6 are indicated in boldface.

	Classical		Pop	
Attribute	F1	F2	F1	F2
width	.50	**.75**	**.94**	.17
elevation	**.83**	.41	.15	**.93**
spaciousness	**.68**	**.68**	**.93**	.26
envelopment	.56	**.77**	**.94**	.17
distance	−.16	**−.88**	**.84**	.13
brightness	**.91**	.24	.24	**.92**
clarity	**.90**	.35	**.78**	.47
Var. explained (%)	48	39	58	30

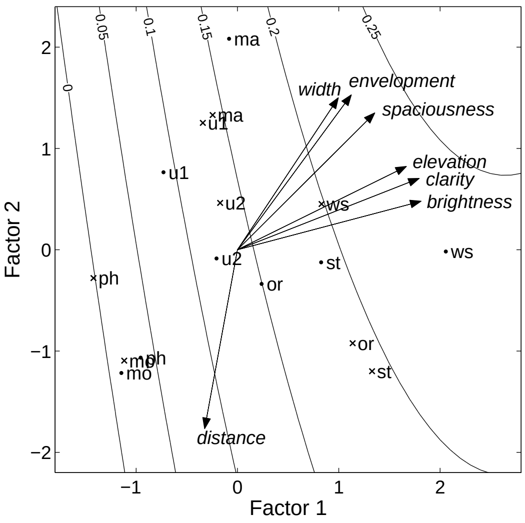


Figure 6: Graphical representation of the factor space obtained from principal component analysis of the attribute scales, and predicted preference (Equation 7) for the classical music material. Factor loadings of the attributes are shown as arrows, and the scores of the reproduction modes along the two factors are represented as dots (Beethoven) or crosses (Rachmaninov). The preference estimated from the two factors is represented by contour lines.

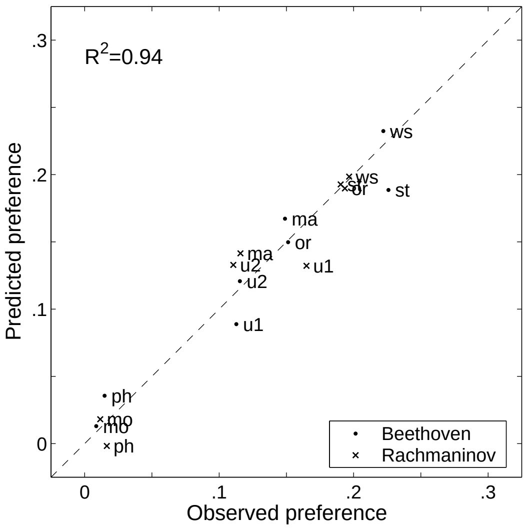


Figure 7: Predicted (Equation 7) vs. observed preference for the classical music material (Beethoven and Rachmaninov).

nents accounted for 58 and 30% (88% cumulated) after rotation. The loadings of the attribute scales on the first two factors, calculated as correlation coefficients, are reported in Table VI. Although the relationship between the attributes and the two factors is more clearcut for the pop music (because the intercorrelation between the attributes is not as strong as for the classical music), similarities can be observed between the two genres: *brightness* and *elevation* load on the same factor, while the other factor is closely related to *width*, *spaciousness*, *envelopment* and *distance* (note that *distance* loads negatively for the classical music; see also Figure 4). Thus, an analogy can be made between Factor 1 in the PCA for classical music and Factor 2 for the pop music, and vice-versa, with the following exceptions: *clarity* which loads on Factor 1 in both cases, and *spaciousness* which loads equally on both factors for the classical material. Figures 6 and 8 show a graphical representation of the attribute loadings and stimulus scores in the two-dimensional factor spaces. The coordinates of the arrow endpoints are calculated as two times the factor loadings.

A multiple regression was performed on the two factors (F_1 and F_2) obtained from PCA in order to predict the preference scale values (P) obtained in Experiment III. The resulting regression equations are

$$\hat{P} = .138 + .075F_1 + .017F_2 - .014F_1^2 \quad \text{(Classical)} \quad (7)$$

$$\hat{P} = .155 + .057F_1 + .058F_2 - .032F_2^2 \quad \text{(Pop)} \quad (8)$$

all three terms in each equation being significant. In both genres, the quadratic term refers to the factor correlating with *brightness* and *elevation*, and is mainly due to *ws* which was both bright and elevated, but only moderately preferred. This gives rise to an inverse u-shaped relation between this factor and preference which was modeled by the quadratic term. The predicted preferences are illustrated by contour lines in Figures 6 and 8 for classical and pop music, respectively; the values written along the equal-preference contours follow from Equation 7 and 8. In Figure 6, for example, predicted preference increases when moving from the left to the upper right part of the panel. Generally, the two models were found to predict preference quite well (Figures 7 and 9), with a total explained variance of 94% (classical) and 84% (pop). The largest prediction errors were obtained for *u1* in the classical music, and *st* in the pop music, both being underestimated.

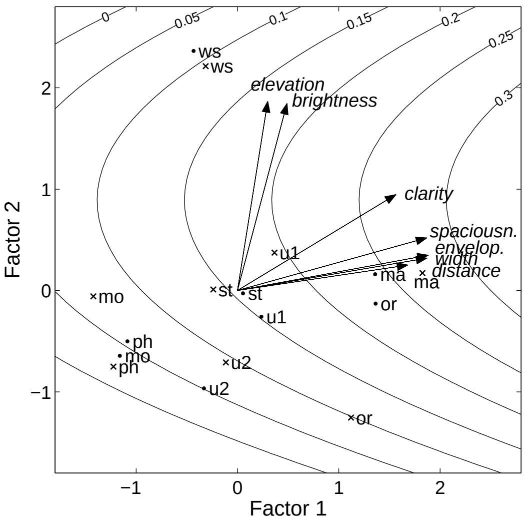


Figure 8: Graphical representation of the factor space obtained from principal component analysis, and predicted preference (Equation 8) for the pop music material. Factor loadings of the attributes are shown as arrows, and the scores of the reproduction modes along the two factors are represented as dots (Steely Dan) or crosses (Sting). The preference estimated from the two factors is represented by contour lines.

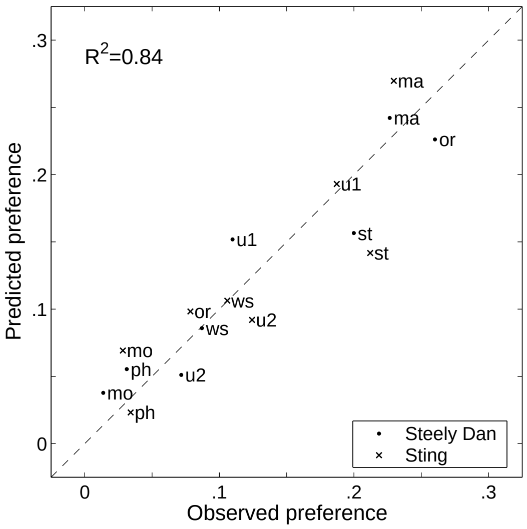


Figure 9: Predicted (Equation 8) vs. observed preference for the pop music material (Steely Dan and Sting).

IV. DISCUSSION

A. Scaling auditory attributes using probabilistic choice models

The quantification of attributes which play a role in the context of multichannel reproduced sound is a non-trivial problem because of the complex nature of the stimuli which typically give rise to several timbral, and spatial sensations simultaneously. From the outset, it is by no means clear that the endeavor of deriving a representation of even a single attribute (like, e. g., *spaciousness*) from listener judgments will be successful at all; inconsistent, intransitive behavior might render any numerical scale meaningless. Hence, the present study goes beyond previous work in that scales of both overall preference *and* the underlying—more basic—attributes were obtained using well-founded methodologies. Paired-comparison judgments were collected in order to allow inconsistencies to reveal themselves (which would have been impossible using direct scaling procedures). Subsequently, probabilistic choice models were employed to statistically evaluate the intransitivities encountered, and, whenever possible, to derive scales of sensation magnitude. It was demonstrated that listeners can consistently judge both upon their global preference and on more specific auditory attributes. Although the preference judgments might reasonably be assumed to be based—at least unconsciously—on many different aspects, listeners were evidently able to integrate them into a unidimensional judgment. This agrees with evidence from other fields of sound quality research, where global auditory attributes, for example the *overall unpleasantness*, have been thoroughly investigated with respect to whether listeners can make transitive judgments about heterogeneous sets of environmental sounds (Ellermeier *et al.*, 2004; Zimmer *et al.*, 2004). In the former study, a BTL model was found to represent the choice frequencies, while in the latter one, a simple EBA model was required to account for the complex stimuli. Taken together, these results suggest that it will strongly depend on the context to what extent the multiple aspects of complex stimuli pose a problem for deriving a meaningful sensation scale. To simply *assume* its unidimensionality, however, is hard to justify.

It is an encouraging result of the present study that the overall preference of the listeners was measurable at a high scale level, and that it was stable over a period of six months. The experience which the listeners had gained during their participation in the experiment had the beneficial effect that subtle differences between the reproduction modes became more salient to them in the course of time. From the preference data collected at two points in time in this study it can be concluded that non-expert listeners have a clear and stable concept of what they would like to listen to.

The highly restrictive BTL model which implies strong stochastic transitivity was found to be less adequate for some of the rather "simple" auditory attributes—especially for *envelopment* and *width* for the Steely Dan excerpt—than for the "complex" overall preference. Therefore, it cannot always be assumed that a seemingly simple question like "How wide is the sound event?" would yield a unidimensional evaluation for any kind of stimulus. In the present case, however, it was possible to find less restrictive EBA models which accounted for the few situations in which the BTL model was violated. The model structures as well as the hypothesis that discrete sound sources in the surround channels might have been responsible for the BTL model to fail should be confirmed in further studies.

It is conceivable that inconsistencies resulting from multidimensional stimuli could be eliminated by training the listeners and breaking up the problematic attribute in several unidimensional "subattributes."[3] Probabilistic choice models therefore constitute a valuable diagnostic tool to reveal such problems, even if they are difficult to point out directly by the listeners (or even the experimenter).

B. Generalizability across program materials

For all attributes as well as for overall preference, the type of program material had a significant effect, suggesting that perceptual effects evoked by the selected reproduction modes depend on the musical signals they are applied to. Nevertheless, certain similarities can be observed across programs. For instance, it appears clearly from Figure 4 that the effect of reproduction mode on *width*, *envelopment* and *spaciousness* is preserved across programs.

For other attributes (e. g., *elevation* and *distance*), certain patterns can be observed which distinguish the classical from the pop music selections. This is also true for preference (Figure 2): While matrix-upmixing (*ma*) was preferred over stereo (*st*) for pop music, it made it worse for the classical programs. Conversely, while increasing the stereo base angle (*ws*) was beneficial for classical music, it was detrimental for pop music. Bech (1998) showed that wider base angles yield higher perceived quality; however, this investigation only included angles up to ±30°. Increasing the angle to ±45° in the present study resulted in a perceived elevation of the sound sources (cf. Figure 4) which could be the reason for the lower preference for *ws* in the pop music. Such an elevation effect as a function of loudspeaker base angle has been studied by Damaske (1969), and can be explained by the spectral changes introduced (Bloom, 1977), a phenomenon closely related to Blauert's (1997, Chap. 2) "boosted bands". This constitutes a plausible explanation for the high correlation observed between the attributes *elevation* and *brightness* (Figure 4).

Similarities within musical genres can also be seen in the MDS maps (Figure 5). The similarities between the Beethoven and Rachmaninov excerpts, and those between Steely Dan and Sting might be attributed to the recording techniques (which were identical within musical genres) rather than to the genres per se; however, recording techniques and musical genres are likely to be confounded in many recordings available today, thus the origin of the similarities observed cannot be unequivocally revealed based on the present data.

Finally, two observations can be made across musical genres. First, mono and phantom mono were the least preferred formats for all four types of program material. This is likely to be due to the low values on most of the spatial attributes: *width*, *envelopment* and *spaciousness*, as well as *clarity* and *naturalness*. Second, overall preference for stereo reproduction was quite high in all four types of program material: For only one of the excerpts (Steely Dan) was the original five-channel reproduction preferred over stereo. This is surprising, because the stereo downmix does not necessarily yield an optimal stereo reproduction; a dedicated two-channel mix or recording would presumably have resulted in an even better evaluation. This observation is in contradiction with Zieliński *et al.* (2003) who found stereo to have a clearly reduced audio quality compared to the five-channel reference. Several factors, however, will moderate the perceptual effects of downmixing to stereo, such as the contents of the surround channels in the original five-channel recording, and whether the listening position is optimal, i. e., centered with respect to the loudspeaker setup (as in the present study), or off-centered.

C. Predicting preference

Predicting listener preference from specific subjective attributes and, ultimately, from objective measures, is one of the ongoing challenges in research on sound quality. It was not the ambition of this exploratory study to develop a general sound quality model; however, the relation between specific auditory attributes and overall preference established in this paper provides some insight in which sensations could play a role when assessing the overall quality of reproduced sound.

In order to deal with the collinearity of the elicited, and subsequently scaled attributes, this relation was obtained by regression of the preference scale values on two (orthogonal) principal components extracted from the attribute scales. It is not possible from this study to determine whether this collinearity results from a common underlying sensation, or whether distinct sensations are involved but co-vary in the context of the selected stimuli. Therefore, the relation between single attributes and overall preference must be interpreted with care. It should be seen as an indicator of the possible contribution of each specific attribute, which should be confirmed in further studies.

The four recordings were grouped into two musical genres, resulting in two models, one for classical music (Equation 7) and the other one for pop music (Equation 8), which accounted for 94 and 84% (respectively) of the variance in the preference scale values. The similarities between the classical and pop genres in Table VI and in Equations 7 and 8 are encouraging, as they suggest that similar sensations might have played a similar role in the preference judgments across program material. From the quadratic term in the regression equations, the tentative conclusion might be drawn that, for the two attributes *elevation* and *brightness*, there exists an optimal value above which preference starts to decrease. However, considering the exploratory nature of this study, and the limited number of stimuli, it will be incumbent upon future research to gain a clearer picture of the functional relations between preference in multichannel sound and the underlying auditory attributes.

ACKNOWLEDGMENTS

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 comes 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 Flemming Christensen and Søren Legarth for their help in translating the instructions to the subjects into Danish, Geoff Martin for his qualified advice on the program material selection, and Wolfgang Ellermeier and Søren Bech for their comments on an earlier version of the manuscript.

Appendix: Attribute definitions

These definitions of the attributes were part of the written instructions (in Danish) given to the test subjects prior to Experiment III.

Which of these two sounds is *wider*? Imagine the area occupied by the sound sources (e. g. instruments). For every pair of sounds, you should indicate for which of the sounds this area is wider.

Which of these two sounds is *more elevated*? Some sounds might appear to be positioned at the same level as your ears. Some others might be lower (closer to the floor) or higher (towards the ceiling). Indicate which of the two sounds you perceive as being higher in space. If they seem to be equally elevated (i. e. at the same height), then make your best guess.

Which of these two sounds is *more spacious*? A sound is said to be spacious when you have a good impression of the space in which it is played. Try to imagine this space, it can be a small room for example, or a large hall. Select the sound in which the impression of space is greater.

Which of these two sounds is *more enveloping*? A sound is enveloping when it wraps around you. A very enveloping sound will give you the impression of being immersed in it, while a non-enveloping one will give you the impression of being outside of it.

Which of these two sounds is *further ahead*? Some sounds might appear to be closer to you, whereas others seem further away. If one of the sounds appears to be behind you, then choose the one which is further ahead (in front).

Which of these two sounds is *brighter*? A sound is bright when it has emphasized treble, and dark when the emphasis is on the bass (or lacking treble). As an example, a female voice is usually brighter than a male voice.

Which of these two sounds is *clearer*? The clearer the sound, the more details you can perceive in it. Choose the sound which appears clearer to you.

Which of these two sounds is *more natural*? A sound is natural if it gives you a realistic impression, as opposed to sounding artificial.

Notes

[1] While in principle the criticism previously raised against direct scaling procedures also applies to such directly obtained dissimilarity ratings, their implied ordinal information was assumed to be valid for this first exploratory analysis.

[2] Zieliński *et al.* (2003) make the distinction between foreground/foreground and foreground/background material in order to denote whether or not there are distinct sources in the surround channels.

[3] This is consistent with Rumsey's (2002) proposal that a "macro-attribute" (such as *envelopment*) consists of several "micro-attributes" (such as *individual-source envelopment* and *ensemble envelopment*).

References

Bech, S. (**1998**). "The influence of stereophonic width on the perceived quality of an audiovisual presentation using a multichannel sound system," *J. Audio Eng. Soc.* **46**, 314–322.

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: Spatial Sound Reproduction*, pp. 51–66.

Blauert, J. (**1997**). *Spatial Hearing* (MIT Press, Cambridge, USA).

Bloom, P. J. (**1977**). "Creating source elevation illusions by spectral manipulation," *J. Audio Eng. Soc.* **25**, 560–565.

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

Bradley, R. A., and Terry, M. E. (**1952**). "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika* **39**, 324–345.

Carrol, J. D., and Chang, J. J. (**1970**). "Analysis of individual differences in multidimensional scaling via an n-way generalization of Echart-Young decomposition," *Psychometrika* **35**, 283–319.

Carrol, J. D., and De Soete, G. (**1991**). "Toward a new paradigm for the study of multiattribute choice behavior," *Am. Psychologist* **46**, 342–351.

Choisel, S., and Wickelmaier, F. (**2005**). "Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound," *118th Convention of the Audio Engineering Society, Barcelona, Spain, May 28–31*. Preprint 6369.

Damaske, P. (**1969**). "Richtungsabhängigkeit von Spektrum und Korrelationsfunktionen der an den Ohren empfangenen Signale [Directional dependence of the spectrum and the correlation function of the signals received at the ears]," *Acustica* **22**, 191–204.

Ellermeier, W., Mader, M., and Daniel, P. (**2004**). "Scaling auditory unpleasantness according to the BTL model: Ratio-scale representation and psychoacoustical analysis," *Acust. Acta Acust.* **90**, 101–107.

Gabrielsson, A., and Sjögren, H. (**1979**). "Perceived sound quality of sound-reproducing systems," *J. Acoust. Soc. Am.* **65**, 1019–1033.

Gerzon, M. A. (**1985**). "Ambisonics in multichannel broadcasting and video," *J. Audio Eng. Soc.* **33**, 859–871.

Guastavino, C., and Katz, B. F. G. (**2004**). "Perceptual evaluation of multi-dimensional spatial audio reproduction," *J. Acoust. Soc. Am.* **116**, 1105–1115.

Heller, J. (**2000**). "Representation and assessment of individual semantic knowledge," *Meth. Psychol. Res.* **5**, 1–37.

Hollins, M., Bensmaïa, S., Karlof, K., and Young, F. (**2000**). "Individual differences in perceptual space for tactile textures: Evidence from multidimensional scaling," *Percept. Psychophys.* **62**, 1534–1544.

ISO 389-1 (**1998**). "Reference zero for the calibration of audiometric equipment – Part 1: Reference equivalent threshold sound pressure levels for pure tones and supra-aural earphones," ISO, Geneva, Switzerland.

ITU-R BS.1116 (**1997**). "Methods for the subjective assessment of small impairment in audio systems including multichannel sound systems," International Telecommunications Union, Geneva, Switzerland.

ITU-R BS.1534 (**2003**). "Method for the subjective assessment of intermediate quality level of coding systems," International Telecommunications Union, Geneva, Switzerland.

ITU-R BS.775-1 (**1994**). "Multichannel stereophonic sound system with and without accompanying picture," International Telecommunication Union, Geneva, Switzerland.

ITU-T P.800 (**1996**). "Methods for subjective determination of transmission quality," International Telecommunications Union, Geneva, Switzerland.

Jesteadt, W. (**1980**). "An adaptive procedure for subjective judgments," *Percept. Psychophys.* **28**, 85–88.

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

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Am.* **49**, 467–477.

Luce, R. D. (**1959**). *Individual choice behavior: A theoretical analysis* (Wiley, New York).

Martens, W. L., and Zacharov, N. (**2000**). "Multidimensional perceptual unfolding of spatially processed speech I: Deriving stimulus space using INDSCAL," *109th Convention of the Audio Engineering Society, Los Angeles, USA, September 22–25*. Preprint 5224.

Mattila, V.-V. (**2002**). "Descriptive analysis and ideal point modelling of speech quality in mobile communication," *113th Convention of the Audio Engineering Society, Los Angeles, USA, October 5–8*. Preprint 5704.

May, K. O. (**1954**). "Intransitivity, utility, and the aggregation of preference patterns," *Econometrica* **22**, 1–13.

Nakayama, T., Miura, T., Kosaka, O., Okamoto, M., and Shiga, T. (**1971**). "Subjective assessment of multichannel reproduction," *J. Audio Eng. Soc.* **19**, 744–751.

Oppenheim, A. V., and Schafer, R. W. (**1989**). *Discrete-Time Signal Processing* (Prentice-Hall, Upper Saddle River).

Rumsey, F. (**1998**). "Subjective assessment of the spatial attributes of reproduced sound," in *Proceedings of the AES 15th International Conference: Audio, Acoustics & Small Spaces*, pp. 122–135.

Rumsey, F. (**2001**). *Spatial Audio* (Focal Press, Oxford).

Rumsey, F. (**2002**). "Spatial quality evaluation for reproduced sound: terminology, meaning, and a scene-based paradigm," *J. Audio Eng. Soc.* **50**, 651–666.

Rumsey, F., Zieliński, S. K., Kassier, R., and Bech, S. (**2005**). "On the relative importance of spatial and timbral fidelities in judgements of degraded multichannel audio quality," *J. Acoust. Soc. Am.* **118**, 968–976.

Stevens, S. S. (**1946**). "On the theory of scales of measurement," *Science* **103**, 677–680.

Susini, P., McAdams, S., and Winsberg, S. (**1999**). "A multidimensional technique for sound quality assessment," *Acust. Acta Acust.* **85**, 650–656.

Tversky, A. (**1969**). "Intransitivity of preferences," *Psychol. Rev.* **76**, 31–48.

Tversky, A. (**1972**). "Elimination by aspects: a theory of choice," *Psychol. Rev.* **79**, 281–299.

Tversky, A., and Sattath, S. (**1979**). "Preference trees," *Psychol. Rev.* **86**, 542–573.

Wickelmaier, F., and Choisel, S. (**2005**). "Selecting participants for listening tests of multichannel reproduced sound," *118th Convention of the Audio Engineering Society, Barcelona, Spain, May 28–31*. Preprint 6483.

Wickelmaier, F., and Ellermeier, W. (**2005**). "Deriving auditory features from triadic comparisons," *Percept. Psychophys.* In press.

Wickelmaier, F., and Schmid, C. (**2004**). "A Matlab function to estimate choice model parameters from paired-comparison data," *Behav. Res. Meth. Instr. Comp.* **36**, 29–40.

Zacharov, N., and Koivuniemi, K. (**2001**). "Audio descriptive analysis & mapping of spatial sound displays," in *Proceedings of the 2001 International Conference on Auditory Displays, Espoo, Finland*, pp. 95–104.

Zieliński, S. K., Rumsey, F., and Bech, S. (**2003**). "Effects of down-mix algorithms on quality of surround sound," *J. Audio Eng. Soc.* **51**, 780–798.

Zimmer, K., Ellermeier, W., and Schmid, C. (**2004**). "Using probabilistic choice models to investigate auditory unpleasantness," *Acust. Acta Acust.* **90**, 1019–1028.

Appendix: Tutorial to the task in perceptual structure analysis

The following tutorial (originally in Danish) was presented to the subjects who performed the perceptual structure analysis (PSA) as reported in Manuscript D.

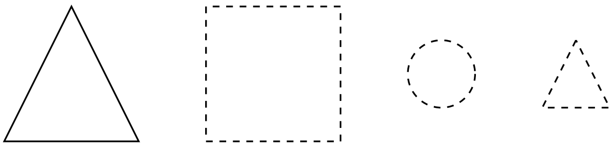
After having completed this tutorial, the listeners were introduced to their actual task which involved triadic comparisons among sounds.

Dear participant,

In this second experiment we want to find out what features characterize the sounds you are presented with. As in the first experiment, there will be no right or wrong answers, but we are interested in the features as you perceive them.

In order to make clear what your task will be, we will introduce you to the procedure by means of examples with pictures.

Consider the following pictures. What features do you discover?

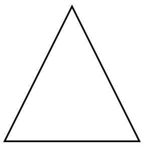
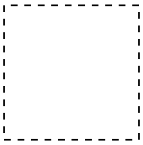
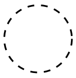
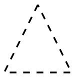


(continued on next page)

Part I:

Task: Look at all the pictures and try to recognize the characterizing features.

For example:

Picture	Potential features
	triangular, solid, big
	square, dashed, big
	circular, dashed, small
	triangular, dashed, small

Note:

You don't have to give names to the specific features you recognize. It will, however, help you in the second part, if you can identify them clearly.

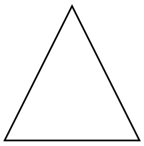
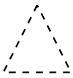
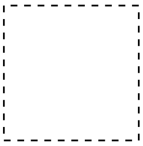
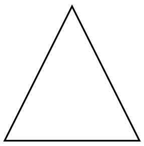
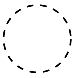
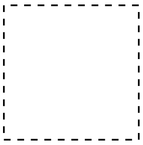
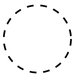
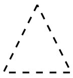
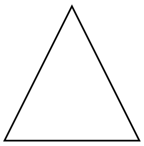
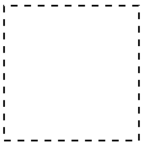
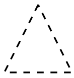
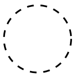
(continued on next page)

Part II:

Task: For each row, look at the three pictures and answer the following question based on the features you have recognized in the first part:

Do the first and second picture share a feature, which the third picture does not have?

For example:

1	2	3	Answer (feature)
			Yes (triangular)
			No
			Yes (small, dashed)
			No (all three are dashed)

Note:

You only have to answer "Yes," or "No." You don't have to name the features that determined your decision. It will help you, however, in making the decision, if you identify the features before you answer.

It might happen, that you discover new features in the second part, which you haven't recognized in the first part. If this is the case, try to also take the new features into account in future answers.

Please ask now if you have any questions about your task.