

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

Sylvain Choisel

Sound Quality Research Unit, Department of Acoustics, Aalborg University, 9220 Aalborg, Denmark  
Bang & Olufsen A/S, Peter Bangs vej 15, 7600 Struer, Denmark

Florian Wickelmaier

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

(Received 18 November 2005; revised 10 October 2006; accepted 11 October 2006)

A study was conducted with the goal of quantifying auditory attributes that underlie listener preference for multichannel reproduced sound. Short musical excerpts were presented in mono, stereo, and several multichannel formats to a panel of 40 selected listeners. Scaling of auditory attributes, as well as overall preference, was based on consistency tests of binary paired-comparison judgments and on modeling the choice frequencies using probabilistic choice models. As a result, the preferences of nonexpert 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 some generalizations within musical program genres regarding the perception of and preference for certain spatial reproduction modes, but for limited generalizations across selections from different musical genres. © 2007 Acoustical Society of America. [DOI: 10.1121/1.2385043]

PACS number(s): 43.66.Lj, 43.66.Ba, 43.38.Md, 43.38.Vk [AK]

Pages: 388–400

## 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 are 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 (ITU-R BS.775-1, 1994), or by various processing algorithms. More recent studies have addressed the problem of identifying and quantifying auditory attributes that are relevant to sound quality in the context of multichannel reproduced sound (Rumsey, 1998; Berg and Rumsey, 2006; 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*, defined in ITU-R BS.1116 (1997)—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 investigation, however, intended to seek explanations for such global differences in terms of more specific auditory attributes. It was part of a larger-scale study, the goals of which were to (1) identify the auditory attributes that 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 attributes relevant for this study has been reported elsewhere (Choisel and Wickelmaier, 2006a), and in the present paper emphasis is placed on the remaining three goals.

In all the earlier investigations cited above, auditory attributes and/or overall quality were directly estimated using rating scales with either numerical or verbal labels, or graphical (visual analog) scales. Such direct scaling procedures are the *de-facto* 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 (ITU-R BS.1116, 1997) and intermediate (ITU-R 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 two or three dimensions already 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 end points of the scale. Methods to deal with scale-usage heterogeneity exist (e.g., Rossi *et al.*, 2001), but they employ involved statistical procedures and are therefore rarely used in practice. 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). Such scaling methods have been successfully applied to sound quality evaluation, most notably to auditory unpleasantness (Ellermeier *et al.*, 2004; Zimmer *et al.*, 2004). In the present study, probabilistic choice models are employed both for determining the overall preference and for measuring the strength of more basic auditory attributes, thereby verifying that listeners could judge upon them in a consistent manner. Subsequently, the resulting scale values are applied to formulate an exploratory statistical model in which preference is related to the auditory attributes.

## II. METHOD

### A. Apparatus and stimuli

#### 1. Experimental setup

The listening tests took place in a 60 m<sup>2</sup> sound-insulated listening room complying with the ITU-R BS.1116 (1997) requirements. Seven loudspeakers (Genelec 1031A) were placed as shown in Fig. 1, at a distance of 2.5 m to the listening position. The height of the tweeters was 108 cm above the floor, which corresponds the average height of the entrance of the listeners' ear canals when seated. Five of the seven loudspeakers were arranged in accordance with the ITU-R recommendation BS.775-1 (1994); two additional speakers (LL and RR) were placed at  $\pm 45^\circ$  for the reproduc-

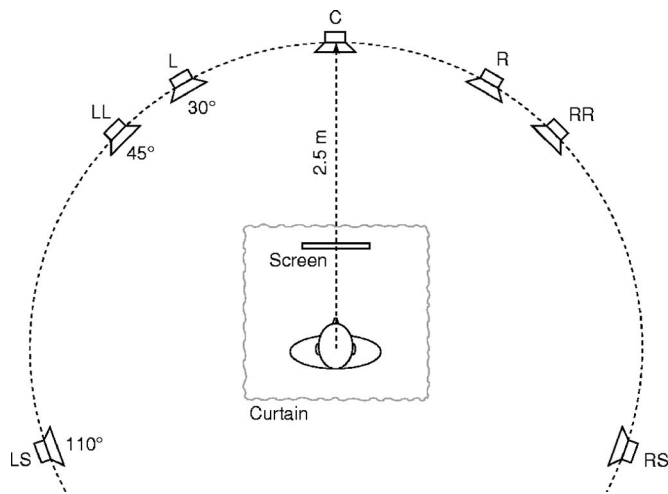


FIG. 1. Playback setup consisting of seven loudspeakers: left (L), right (R), center (C), left-of-left (LL), right-of-right (RR), left surround (LS), and right surround (RS). This setup was symmetrically placed with respect to the width of the room and was hidden from the subject by an acoustically transparent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

tion 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.

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 eight-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 in. flat screen placed in front of the listener, below the loudspeaker level (45 cm above the floor) in order to limit interactions with the sound field. A headrest 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.

The seven loudspeakers were matched in sensitivity and minimum-phase frequency response based on impulse response measurements carried out in an anechoic chamber using a 14th order maximum length sequence (MLS) at 48 kHz. The equalization was implemented as FIR filters applied to the corresponding channels in the sound files. In order to verify that the interchannel level alignment was preserved in the listening room, the A-weighted sound-pressure level of bandpassed pink noise (200 Hz–2 kHz) was measured at the listening position for each channel. As a result, the interchannel level differences were within 0.3 dB, and the differences between left/right pairs did not exceed 0.1 dB.

#### 2. Program material

Four musical excerpts (two pop, two classical) were selected from commercially available multichannel material (Table I). Their different musical contents (genre, instrumental versus vocal) as well as the various spatial information present in the multichannel mix (natural room reverb in the

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

classical recordings, or distributed instruments in pop music) made this selection suitable for eliciting different spatial sensations. The two classical recordings were made with five omnidirectional microphones placed in a circular array, and the two pop recordings were mixed with standard surround panning technique. These excerpts were transferred from their original medium—Super Audio Compact Disc (SACD) or Digital Versatile Disc—Audio (DVD-A)—onto a computer (48 KHz, 24 bit) using a Denon 2200 player connected to an eight-channel A/D converter (RME ADI-8 DS), and carefully cut to include a musical phrase, their duration ranging from 4.7 to 5.4 s.

### 3. Reproduction modes

From the original five-channel program material (or), seven additional formats were derived, as summarized in Table II. This selection of reproduction modes was made in order to create a wide range of perceptual changes typically encountered in home audio applications. First, the original was mixed down to stereo (st) according to the ITU-R recommendation BS.775-1 (1994):

$$L_{st} = L_{or} + \frac{1}{\sqrt{2}}C_{or} + \frac{1}{\sqrt{2}}LS_{or},$$

$$R_{st} = R_{or} + \frac{1}{\sqrt{2}}C_{or} + \frac{1}{\sqrt{2}}RS_{or}. \quad (1)$$

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

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

TABLE II. Reproduction modes: full name, abbreviation, and loudspeakers used for playback (see Fig. 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	— <sup>a</sup>	L,R,C,LS,RS
DTS Neo:6	— <sup>a</sup>	L,R,C,LS,RS
Original 5.0	or	L,R,C,LS,RS

<sup>a</sup>Referred to as u1 and u2 (in no specific order) in the rest of this paper.

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

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

Finally, three upmixing algorithms were used to reconstruct 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 surround processor (Meridian 861) that was 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, an audio/video receiver (Yamaha RX-V 640) was used to generate the DTS Neo:6 upmix. The matrix upmixing (ma) was inspired by matrix decoding systems that are typically applied to encoded stereo tracks (cf. Rumsey, 2001). In this study, however, it was applied to a “regular” stereo downmix [Eq. (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 6 dB. The front ( $L$  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-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 in Choisel and Wickelmaier (2006a). The resulting loudness 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 bits ( $\pm 1$  LSB triangular probability density function) and with a sampling frequency of 48 kHz. Further information regarding acoustical and psychoacoustical characteristics of the reproduction modes may be found in Choisel and Wickelmaier (2006b).

## B. Subjects

Forty listeners (28 male, 12 female) took part in this study. They were mostly university students, had no or little prior experience with this type of experiment, and were naïve with respect to the research questions. In contrast to expert listeners trained to identify subtle differences in a reliable manner, the selected sample was closer to a consumer population. Such a sample was chosen in order to avoid possible biases, such as an excessive display of knowledge during the identification of auditory attributes. It was, however, desired that they possess the ability to perform the tasks required from them. For that purpose, these participants were selected among 78 candidates, according to their auditory and verbal aptitudes. The selection procedure (detailed in Wickelmaier and Choisel, 2005) consisted of pure-tone audiometry, a stereo-width discrimination task, and a verbal-fluency test. These tests were performed 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 the 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 years (median=24 years). One of the participants dropped out during the first part of the experiment, the remaining 39 took part in the complete study that extended over approximately six months.

## C. Procedure

The study was organized in several larger experimental parts, throughout which the same sample of subjects and the same set of stimuli were used. First, an overall preference was determined for the reproduction modes. Next, auditory attributes salient in the context of these sounds were identified (elicited) using the same sample of subjects; this part is reported in Choisel and Wickelmaier (2006a); the outcome was a set of eight attributes: *width*, *elevation*, *spaciousness*, *envelopment*, *distance*, *brightness*, *clarity*, and *naturalness*. Subsequently, the strength of these attributes was quantified. Finally, the preference was reevaluated.

### 1. Quantification of auditory attributes

Quantification of the attributes was carried out by asking the subjects (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,

generated by the authors so as to represent as much as possible the subjects’ own descriptors, can be found in the Appendix .

For each of the eight attributes and for four musical excerpts, all possible pairs of reproduction modes were presented to the subjects. Two buttons on a computer screen, labeled A and B, were visually emphasized in turn (by changing their size) during playback to indicate which sound was played. The response was made by clicking the button corresponding to the chosen sound. Each pair was judged only once. The within-pair order was balanced across subjects (David, 1988, Chap. 5) and the between-pair order was random. Each attribute was evaluated for all four program materials in a single block lasting for about 25 min. Each subject evaluated two attributes in a session lasting for one hour, 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  $8 \times 8$  Graeco-Latin squares. Each subject gave 28 judgments per program material and auditory attribute.

### 2. Quantification of overall preference

It was hypothesized that as the study proceeded (especially by taking part in the attribute elicitation, Choisel and Wickelmaier, 2006a) participants would gain experience with the sounds, which potentially influenced their perception. In order to investigate the influence of experience, preference was measured at two points in time: once at the beginning (first measurement) and once at the end of the study (second measurement, about six months after the first data collection). For each pair of reproduction modes the subjects were instructed to indicate which one they preferred. In the first data collection, each pair was presented in both within-pair orders (AB and BA), and a third time in one within-pair order, counterbalanced across subjects. The second data collection on preference only included two judgments per pair (both within-pair orders). Thus, each subject gave 84 (respectively, 56) preference judgments per program material in the first (respectively, second) data collection.

## D. Analysis of choice frequencies

Both, for 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 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.

The 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 transitivity, which imply that if  $P_{xy} \geq 0.5$  and  $P_{yz} \geq 0.5$ , then

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

for all sounds  $x$ ,  $y$ , and  $z$ . Whenever the premise holds, but the implication in Eq. (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 subjects were not able to integrate several stimulus dimensions into one (common) 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 (Carroll and De Soete, 1991).

Transitivity violations might result from either individually inconsistent choice behavior or from subjects disagreeing in their choices. The probabilistic choice models reported in this paper were applied to aggregate data, and it is therefore not possible to separate these two sources of inconsistencies. Models that account for individual differences have been developed (e. g., Böckenholt, 2001), but are not within the scope of the present study.

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 a 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,

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

where  $u(\cdot)$  is a ratio scale of the criterion. Since Eq. (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* that belongs to the first but not to the second sound. EBA predicts  $P_{xy}$  by

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

where  $\alpha, \beta, \dots$ , are the aspects (or features) of the sounds,  $x'$  indicates the set of aspects belonging to sound  $x$ , and  $x' \setminus y'$

denotes the set of aspects belonging to sound  $x$  but not to sound  $y$ . As in 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  $\chi^2$  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 suprathreshold 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 (before and after elicitation and scaling of the attributes), standard likelihood ratio tests (McCullagh and Nelder, 1989) 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 second measurement. A likelihood ratio test is possible whenever two models are *nested*, that is, one model results from the other one by applying restrictions on its parameters. A significant likelihood ratio test denotes that the restricted model is to be rejected.

### III. RESULTS

#### A. Scaling listener preference

Table III displays the evaluation of the stochastic transitivities [Eq. (4)] of the preference judgments collected before and after the subjects went through the elicitation and scaling of specific auditory attributes. For the evaluation, data were aggregated over all subjects and repetitions, within each type of program material. Thus, the choice probabilities in the first measurement were estimated based on  $N=40 \times 3=120$  observations per stimulus pair for Steely Dan, and on  $N=39 \times 3=117$  for the other program materials, since one subject left the experiment after the first session. In the second measurement, where two replicates were collected, the choice probabilities were based on  $N=39 \times 2=78$  observations. Weak and moderate stochastic transitivities 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

TABLE III. Transitivity violations and goodness-of-fit test of the BTL model for preference judgments at two points in the study: before and after elicitation and scaling of attributes. Displayed are the number of violations of weak, moderate, and strong stochastic transitivity [Eq. (4)], and the test statistic and p-value of a likelihood ratio test with the null hypothesis that the BTL model holds.

Excerpt	First measurement					Second measurement				
	WST	MST	SST	$\chi^2(21)$	$p$	WST	MST	SST	$\chi^2(21)$	$p$
Beethoven	0	2	14	9.13	0.988	0	1	12	9.06	0.989
Rachmaninov	2	4	19	16.96	0.714	0	0	18	8.44	0.993
Steely Dan	0	0	12	18.13	0.640	0	0	11	17.74	0.666
Sting	0	0	13	10.72	0.968	0	0	9	13.66	0.884

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 [Eq. (5)] was fitted to the paired-comparison data. Table III shows the results of the goodness-of-fit tests which 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. Consequently, it was possible to measure listener preference at 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 (respectively, two) repetitions, *within* each measurement. 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 second measurement, 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 measurement. 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 measurements, 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 the 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.

Across the four program materials and the two points of measurement, 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

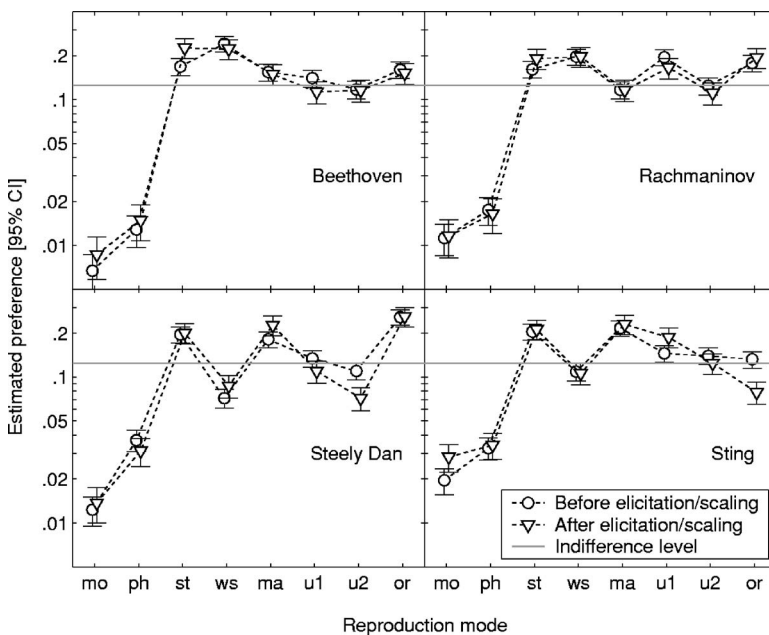


FIG. 2. Ratio scale of preference derived from eight reproduction modes for four musical excerpts. Scale values represent parameter estimates of the BTL model fitted to paired-comparison judgments. Preference was measured at two points in the study (see text), represented by two symbol styles. The reproduction 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.

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 program materials. In spite of the obvious similarities within the classical and the pop genres (see the rows in Fig. 2), the excerpt-specific differences were statistically significant. In the first measurement, a common model for Beethoven and Rachmaninov fared significantly worse [ $\chi^2(7)=30.81; p<0.001$ ] than a model having two sets of parameters: one for each program material. The same was true for Steely Dan and Sting [ $\chi^2(7)=78.15; p<0.001$ ]. Analogous results were obtained in the second measurement for classical [ $\chi^2(7)=21.14; p=0.004$ ] and pop music [ $\chi^2(7)=146.01; p<0.001$ ], 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 second measurement. 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* 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 second measurement. No, significant changes were observed for Beethoven [ $\chi^2(7)=12.33; p=0.090$ ] and Rachmaninov [ $\chi^2(7)=5.90; p=0.551$ ], whereas for Steely Dan [ $\chi^2(7)=25.37; p=0.001$ ] and Sting [ $\chi^2(7)=35.80; p<0.001$ ] 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 measurement (see the bottom right panel in Fig. 2). In the second measurement, 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 the bottom left panel in Fig. 2).

TABLE IV. Transitivity violations and goodness-of-fit test of the BTL model for selected attributes. See Table III. Note: \* $p<0.05$ .

Attribute	WST	MST	SST	$\chi^2(21)$	$p$
Beethoven					
Width	0	1	19	24.55	0.267
Elevation	1	11	25	24.63	0.263
Spaciousness	0	2	18	17.80	0.661
Envelopment	0	3	23	22.16	0.391
Distance	3	9	32	22.83	0.353
Brightness	2	3	19	12.25	0.933
Clarity	4	5	27	25.55	0.224
Naturalness	3	5	24	15.41	0.802
Rachmaninov					
Width	1	1	14	21.20	0.447
Elevation	2	7	23	16.08	0.765
Spaciousness	2	7	19	7.35	0.997
Envelopment	2	4	27	16.82	0.722
Distance	2	11	37	21.74	0.414
Brightness	4	4	27	14.49	0.848
Clarity	2	6	21	8.86	0.990
Naturalness	0	2	14	16.46	0.744
Steely Dan					
Width	0	3	14	36.01	0.022*
Elevation	0	2	24	30.64	0.080
Spaciousness	2	2	19	26.66	0.182
Envelopment	0	2	23	39.40	0.009*
Distance	3	13	30	15.89	0.776
Brightness	0	0	15	20.39	0.496
Clarity	0	2	18	14.05	0.867
Naturalness	0	2	18	14.35	0.854
Sting					
Width	0	2	24	29.47	0.103
Elevation	0	0	16	27.30	0.161
Spaciousness	0	4	16	22.60	0.366
Envelopment	1	3	16	15.04	0.821
Distance	0	1	19	21.40	0.435
Brightness	0	0	23	31.54	0.065
Clarity	2	3	16	19.24	0.570
Naturalness	1	1	18	11.72	0.947

## B. Scaling auditory attributes

The same logic of consistency checks, model evaluation, and scaling was applied to the more elementary auditory attributes. Table IV displays the violations of the stochastic transitivity 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 IV, 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.

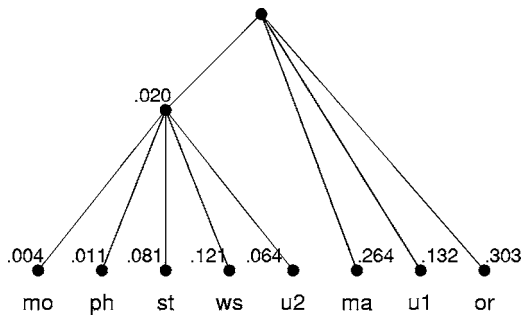


FIG. 3. Elimination-by-aspects (EBA) model structure and parameter estimates for *envelopment* (Steely Dan). Nodes represent aspects shared only by the connected reproduction modes [see Eq. (6)]. Scale values are obtained by adding up the parameters belonging to each reproduction mode.

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  $\alpha$  level of 5%. The original Steely Dan material, however, was different from the other three excerpts in that it not only contains 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 [Eq. (6)]. Among the EBA models with only one additional parameter, the best fitting one for *envelopment* is depicted in Fig. 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 that do *not* reproduce any discrete source at the side of or behind the listener.<sup>1</sup> This simple EBA model was found to fit the data [ $\chi^2(20)=26.55; p=0.148$ ] the improvement over the BTL model being significant [ $\chi^2(1)=12.85; p<0.001$ ]. The parameter estimates are also displayed in Fig. 3. In order to derive *envelopment* scale values from the model, the parameters belonging to each reproduction mode were added up. For example,  $u(\text{mo})=0.02+0.004=0.024$ . Similarly, an EBA model was found for the *width* attribute, which accounted for the data [ $\chi^2(20)=27.27; p=0.128$ ] and outperformed the BTL model [ $\chi^2(1)=8.74; p=0.003$ ]. 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, a considerable similarity of the scales was observed across program materials, which was even more pronounced within a 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 nearest 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

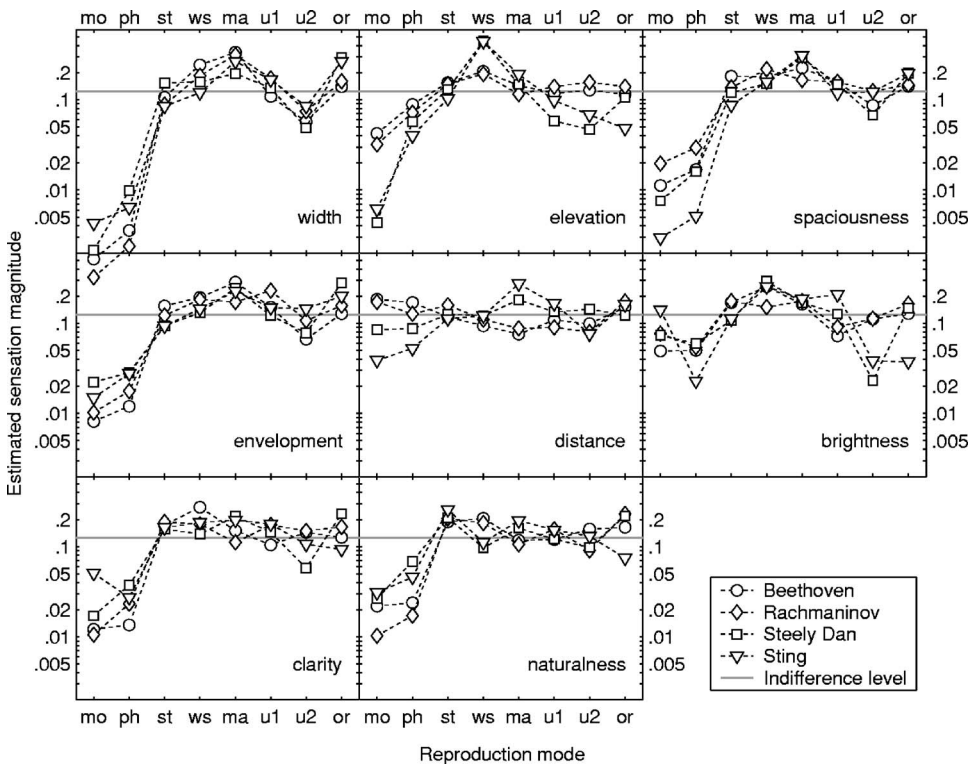


FIG. 4. Ratio scales of eight auditory attributes estimated using BTL and EBA models for four types of program material.

striking. From the high correlations it is evident that the attributes did not vary independently in the stimuli under study, or that listeners were not able to distinguish between all of them.

### C. Relation between specific sensations and overall preference

A simple way of relating the auditory attributes to overall preference is multiple regression, where the predicting variables are the attribute scales and the predicted variable is the preference scale. While this approach might at first glance provide a satisfactory goodness of fit, it suffers from two shortcomings, which make it inapplicable in many situations such as the present investigation. First, 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. The second problem is the high correlation between some of the attributes; collinearity of the independent variables in multiple regression often yields an unstable and therefore unreliable model.

In previous studies, several methods have been used to relate the overall quality to more specific perceptual dimensions, while simultaneously addressing the problem of collinear attributes (Nakayama *et al.*, 1971; Zacharov and Koivuniemi, 2001; Mattila, 2002; Rumsey *et al.*, 2005). Such methods are based on multidimensional scaling (MDS), principal component analysis (PCA), or related techniques. PCA reduces the attributes to a few independent factors (or *principal components*) that are orthogonal by construction, as are MDS dimensions. This makes them suitable as predictors in a regression model, circumventing the problem of collinearity mentioned above.

In the present study, multiple regression based on principle components was used to predict preference. In order to increase the generalizability of the model, the data were combined within a 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 Fig. 4). *Naturalness* was excluded from the analysis because it was considered more global than the other (more specific) attributes and not sufficiently separate from preference, the correlation between *naturalness* and preference ranging from 0.94 (Steely Dan) to 0.98 (Rachmaninov).

PCA with varimax rotation 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 components 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 V. Although the relationship between the attributes and the two factors is more clear cut 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*

TABLE V. Attribute loadings on the factors ( $F_1$  and  $F_2$ ) obtained from principal component analysis, and variance explained by these factors after varimax rotation. Loadings higher than 0.6 are indicated in boldface.

Attribute	Classical		Pop	
	F1	F2	F1	F2
Width	0.50	<b>0.75</b>	<b>0.94</b>	0.17
Spaciousness	<b>0.68</b>	<b>0.68</b>	<b>0.93</b>	0.26
Envelopment	0.56	<b>0.77</b>	<b>0.94</b>	0.17
Distance	-0.16	<b>-0.88</b>	<b>0.84</b>	0.13
Clarity	<b>0.90</b>	0.35	<b>0.78</b>	0.47
Brightness	<b>0.91</b>	0.24	0.24	<b>0.92</b>
Elevation	<b>0.83</b>	0.41	0.15	<b>0.93</b>
Var. explained (%)	48	39	58	30

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 Fig. 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 5 and 6 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.

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 the second measurement (after attribute scaling). The resulting regression equations are

$$\hat{P} = 0.138 + 0.075F_1 + 0.017F_2 - 0.014F_1^2 \quad (\text{classical}), \quad (7)$$

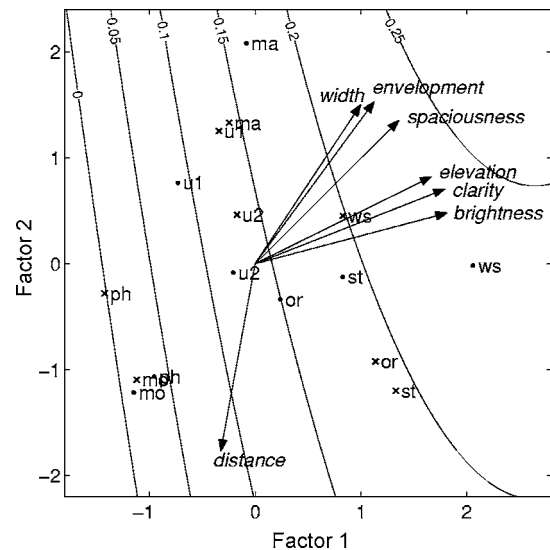


FIG. 5. Graphical representation of the factor space obtained from principal component analysis of the attribute scales, and predicted preference [Eq. (7)] for the classical music material. Factor loadings of the attributes are shown as arrows, and the scores of the reproduction modes along the two factors are represented as dots (Beethoven) or crosses (Rachmaninov). The preference estimated from the two factors is represented by contour lines.

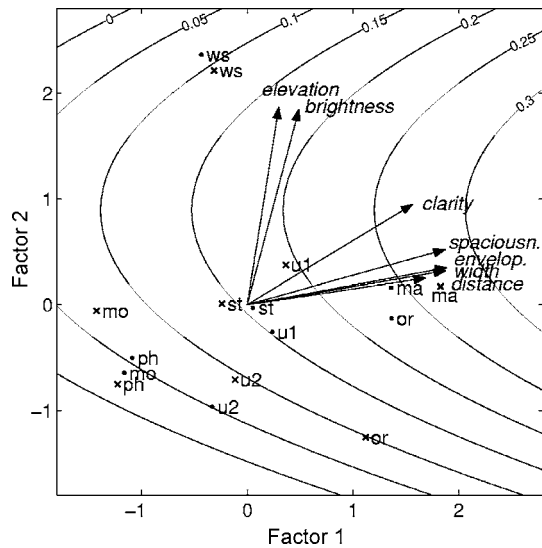


FIG. 6. Graphical representation of the factor space obtained from principal component analysis of the attribute scales, and predicted preference [Eq. (8)] for the pop music material. Factor loadings of the attributes are shown as arrows, and the scores of the reproduction modes along the two factors are represented as dots (Steely Dan) or crosses (Sting). The preference estimated from the two factors is represented by contour lines.

$$\hat{P} = 0.155 + 0.057F_1 + 0.058F_2 - 0.032F_2^2 \quad (\text{pop}), \quad (8)$$

all three terms in each equation being significant. In both genres, the quadratic term refers to the factor correlating with *brightness* and *elevation*, suggesting an ideal point on this dimension. The response surfaces resulting from Eqs. (7) and (8) are shown in Figs. 5 and 6 for classical and pop music, respectively. Each contour line connects points of equal preference, as predicted by the regression model. In Fig. 5, for example, the predicted preference increases when moving from the left to the upper right part of the panel. Generally, the two models were found to predict the preference quite well (Figs. 7 and 8), 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.

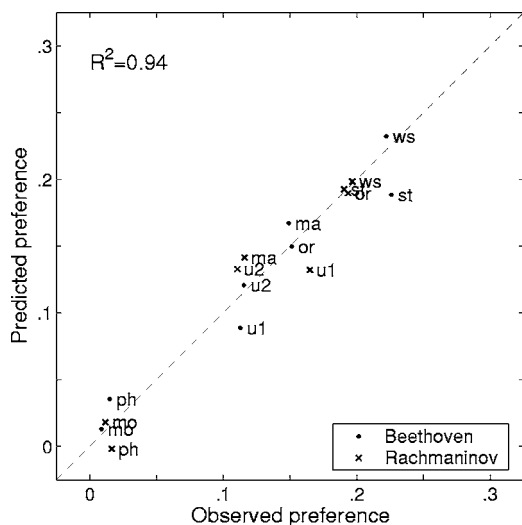


FIG. 7. Predicted [Eq. (7)] versus observed preference for the classical music material (Beethoven and Rachmaninov).

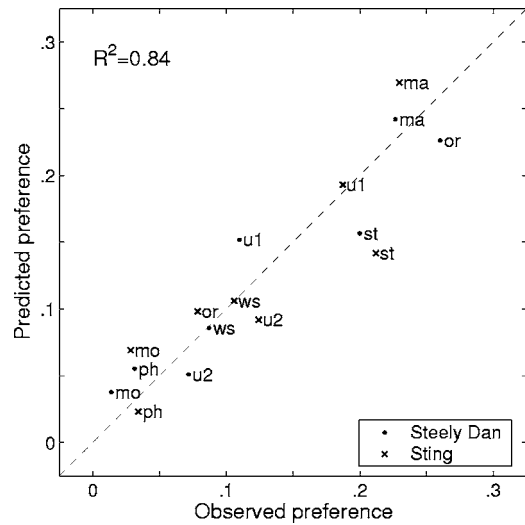


FIG. 8. Predicted [Eq. (8)] versus observed preference for the pop music material (Steely Dan and Sting).

## IV. DISCUSSION

### A. Scaling auditory attributes using probabilistic choice models

The quantification of attributes that play a role in the context of multichannel reproduced sound is a nontrivial problem because of the complex nature of the stimuli that typically gives 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 evaluate statistically 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 prob-

lem for deriving a meaningful sensation scale. To simply assume unidimensionality, however, is hard to justify.

It is worth noting that the aspects represented by EBA parameters might or might not have a direct correspondence to physical characteristics of the stimuli. Rather, the aspects may be viewed as perceptual effects relevant in the decision process. Furthermore, listeners might or might not be fully aware of the aspects when choosing among the sounds, that is to say that—depending on the context of stimuli—aspects may relate to more elementary sensory or higher-level cognitive mechanisms. A sensation scale derived from probabilistic choice models, therefore, reflects to more or less a degree both sensory and judgmental (cognitive) processes (see also Baird, 1997).

It is an encouraging result of the present study that the listeners' overall preference was measurable at a high scale level, and that it was stable over a period of six months. The experience that the listeners had gained during their participation in the experiments 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 nonexpert listeners have a clear and stable concept of the versions of reproduced music to which they would prefer to listen.

The highly restrictive BTL model that 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 that 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."<sup>2</sup> 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 Fig. 4 that the effect of the reproduction mode on *width*, *envelopment*, and *spaciousness* is preserved across programs.

For other attributes (e.g., *elevation* and *distance*), certain patterns can be observed that distinguish the classical from

the pop music selections. This is also true for preference (Fig. 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  $\pm 30^\circ$ . Increasing the angle to  $\pm 45^\circ$  in the present study resulted in a perceived elevation of the sound sources (cf. Fig. 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* (Fig. 4).

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, the 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 may be explained by the subjects' familiarity with two-channel stereo reproduction, or, in the case of classical music, by the only subtle changes introduced by downmixing. Zieliński *et al.* (2003) and Zieliński *et al.* (2005) reported that the perceived quality of material containing dry sources both in the front and in the surround channels (*foreground-foreground*) is more impaired by downmixing than is material containing predominantly reverberation in the surround channels (*foreground-background*). The result of the present study that the stereo downmix was less preferred than the original *only* for the foreground-foreground material (Steely Dan) 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).

## 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 might play a role when assessing the overall quality of reproduced sound.

The four recordings were grouped into two musical genres, resulting in two models: one for classical music [Eq. (7)] and one for pop music [Eq. (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 V and in Eqs. (7) and (8) are encouraging, as they suggest that similar sensations might have played a similar role in the preference judgments across program materials.

In several studies (e.g., Toole, 1985; Letowski, 1989; Zacharov and Koivuniemi, 2001), the overall sound quality is conceived as consisting of timbral and spatial quality. Recently, Rumsey *et al.* (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*, *distance* and *spaciousness* loaded on one of the components, while *brightness* and *elevation* (attributed to spectral changes) loaded on the other one. These results support the notion that both timbral and spatial auditory attributes are important predictors of overall listener preference.

It is not possible from the present data to determine whether the collinearity of certain auditory attributes results from a common underlying sensation, or whether distinct sensations are involved, but covary, in the context of the selected stimuli. Therefore, the relation between single attributes and overall preference must be interpreted with care. 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 overall preference and the underlying (more specific) auditory attributes in the context of multichannel sound.

## V. CONCLUSIONS

In summary, the following conclusions can be drawn:

(1) By applying probabilistic choice models to binary paired-comparison judgments, it is possible to scale auditory attributes in complex sounds, while revealing inconsistencies related to multidimensionality.

(2) Consistent judgments (with respect to transitivity) were obtained from nonexpert listeners on overall preference as well as on more specific attributes.

(3) The preference judgments were highly reliable both across repetitions and when retesting after six months, indicating that listeners have a clear and stable concept of what they prefer to listen to.

(4) Perceptual similarities were observed between materials; those were more pronounced within musical genres (classic and pop) than across.

(5) For the centered listening position investigated in the present study, stereo downmix was found to be among the most preferred formats, while mono was generally the least preferred.

(6) Preference could be predicted using two principal components derived from the attribute scales: one related to the spatial characteristics of the sounds, the other related to their spectral characteristics.

## ACKNOWLEDGMENTS

This research was carried out as part of the “Centerkontrakt on Sound Quality” that 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. Furthermore, we are grateful to the four anonymous reviewers for their comments.

## APPENDIX: ATTRIBUTE DEFINITIONS

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

**Which of these two sounds is wider?** Imagine the area occupied by the sound sources (e.g., the 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 (toward the ceiling). Indicate which of the two sounds you perceive as being higher in space.

**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 nonenveloping 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 farther away. If one of the sounds appears to be behind you, then choose the one that is farther 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 that 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.

<sup>1</sup>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.

- <sup>2</sup>This is consistent with Rumsey's (2002) proposal that a "macroattribute" (such as *envelopment*) consists of several "microattributes" (such as *individual-source envelopment* and *ensemble envelopment*).
- Baird, J. C. (1997). *Sensation and Judgment* (Lawrence Erlbaum, Mahwah, NJ).
- 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. (2006). "Identification of quality attributes of spatial audio by repertory grid technique," *J. Audio Eng. Soc.* **54**, 365–379.
- Blauert, J. (1997). *Spatial Hearing* (MIT Press, Cambridge, MA).
- Bloom, P. J. (1977). "Creating source elevation illusions by spectral manipulation," *J. Audio Eng. Soc.* **25**, 560–565.
- Böckenholt, U. (2001). "Hierarchical modeling of paired comparison data," *Psychol. Methods* **6**, 49–66.
- Bradley, R. A. and Terry, M. E. (1952). "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika* **39**, 324–345.
- Carroll, J. D., and De Soete, G. (1991). "Toward a new paradigm for the study of multiattribute choice behavior," *Am. Psychol.* **46**, 342–351.
- Choiel, S. and Wickelmaier, F. (2006a). "Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound," *J. Audio Eng. Soc.* **54**, 815–826.
- Choiel, S. and Wickelmaier, F. (2006b). "Relating auditory attributes of multichannel sound to preference and to physical parameters," *120th Convention of the Audio Engineering Society*, Paris, France, 20–23 May, preprint 6684.
- 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.
- David, H. A. (1988). *The Method of Paired Comparisons* (Oxford University Press, New York).
- 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.
- 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.
- Letowski, T. (1989). "Sound quality assessment: Concepts and criteria," *87th Convention of the Audio Engineering Society*, New York, USA, 18–21 October, preprint 2825.
- 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).
- 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, 5–8 October, preprint 5704.
- May, K. O. (1954). "Intransitivity, utility, and the aggregation of preference patterns," *Econometrica* **22**, 1–13.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* (Chapman and Hall, London).
- Nakayama, T., Miura, T., Kosaka, O., Okamoto, M., and Shiga, T. (1971). "Subjective assessment of multichannel reproduction," *J. Audio Eng. Soc.* **19**, 744–751.
- Rossi, P. E., Gilula, Z., and Allenby, G. M. (2001). "Overcoming scale usage heterogeneity: A Bayesian hierarchical approach," *J. Am. Stat. Assoc.* **96**, 20–31.
- 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 judgments 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.
- Toole, F. E. (1985). "Subjective measurements of loudspeaker sound quality and listener performance," *J. Audio Eng. Soc.* **33**, 2–32.
- 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 Choiel, S. (2005). "Selecting participants for listening tests of multichannel reproduced sound," *118th Convention of the Audio Engineering Society*, Barcelona, Spain, 28–31 May, preprint 6483.
- Wickelmaier, F. and Schmid, C. (2004). "A Matlab function to estimate choice model parameters from paired-comparison data," *Behav. Res. Methods Instrum. Comput.* **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.
- Zieliński, S. K., Rumsey, F., Kassier, R., and 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," *J. Audio Eng. Soc.* **53**, 174–192.
- Zimmer, K., Ellermeier, W., and Schmid, C. (2004). "Using probabilistic choice models to investigate auditory unpleasantness," *Acust. Acta Acust.* **90**, 1019–1028.