Scaling sound quality using models for paired-comparison and ranking data

Florian Wickelmaier[1], Nora Umbach[1], Konstantin Sering[1], Sylvain Choisel[2]
[1] *Department of Psychology, University of Tübingen, Germany, Email: florian.wickelmaier@uni-tuebingen.de*
[2] *Philips Consumer Lifestyle, Leuven, Belgium*

Introduction

Several psychometric methods exist to measure listeners' preferences among different systems, such as loudspeakers, upmix algorithms and audio codecs. The de-facto standard in sound quality evaluation are direct scaling procedures, as described in recommendations of the International Telecommunications Union (ITU-R BS.1116-1, ITU-R BS.1534-1, ITU-T P800). In direct scaling, participants are required to assign a number or a verbal label to the sounds that are supposed to reflect the magnitude of the sensation evoked by these sounds. Direct scaling methods are widely used, but they rely on many implicit and untested assumptions and are known to give rise to multiple biases [7].

An alternative approach to the quantification of auditory attributes is indirect scaling. Within indirect scaling, global and specific auditory attributes, like audio quality, preference, or spaciousness are defined as latent psychological variables. Their magnitudes cannot be observed directly but need to be inferred from observable behavior. Paired-comparison scaling is an example, where the observable pairwise judgments of listeners are analyzed using psychological models in order to estimate these latent quantities [1]. A possibly less time consuming response collection method than a complete paired comparison are ranking procedures: Listeners have access to all stimuli and rank them, for example, according to their preference. A numerical scale can be derived by subsequent modeling of the ranking data. Some listeners, however, when performing such a ranking, may spend a long time reconsidering and altering their rankings. Therefore, a new ranking procedure (ranking by elimination) is proposed, where the task is to identify the least preferred stimuli, which will subsequently and irrevocably be eliminated from the list. The elimination continues until only one stimulus is left. This is thought to increase the speed of the listening test and to ease the subject's task.

It was the goal of this study, to compare three response collection methods, paired comparisons, a traditional ranking procedure and ranking by elimination with respect to the scaling of the sounds, and to the time required to complete the respective task. We hypothesized that ranking by elimination would be faster than conventional ranking while being as accurate as the pairwise comparisons. More details of the study are given in [6].

Methods

Subjects and stimuli. The sample consisted of 52 participants (41 female). They were between 18 and 44 years old ($M = 23.7$). None of the subjects reported any hearing problems. The experiment was conducted in a sound-attenuating booth. The sounds were played back by a personal computer and were delivered by headphones.

Three musical excerpts were selected from audio CDs, recorded into WAV format at 44.1 kHz sampling frequency and 16-bit resolution, and carefully cut to include a musical phrase of about 5 s duration. None of the excerpts included vocals. The material was selected to be representative of changes in audio quality that are induced by compression algorithms: a castanets excerpt being especially prone to coding artifacts, the other two excerpts constituting examples of bandwidths and content typical of pop and classical music. These unprocessed files were encoded using MP3 and OGG VORBIS codecs. For each type of program material, three MP3 and three OGG encoded versions were generated employing (nominal) bit rates of 128, 96, and 64 kbps.

Procedure. Participants completed three blocks in which they judged the three program materials using one response collection method. In Block 1, subjects did all 21 pairwise comparisons for each program material. Their task was on each trial to choose the sound that had the higher audio quality. In Block 2, half of the listeners performed a regular ranking task for the three musical excerpts, the other half of the subjects completed a ranking by repeatedly eliminating the sound having the lowest quality. On each trial in the regular ranking, participants were presented with all seven compression settings represented as buttons on the screen and asked to rank them with respect to audio quality. Subjects were free to replay each sound as often as they wanted. They were also free to alter their rankings as often as they desired, until they confirmed their final ranking. No time restrictions were given. The response interface for the ranking by elimination was similar to the regular ranking procedure, but this time only the highest rank (lowest quality) could be clicked which subsequently disappeared from the screen. The elimination continued until only the compression setting with the highest audio quality was left.

Analysis of the choices and rankings. Pairwise judgments were aggregated across subjects in order to estimate pairwise choice probabilities, P_{xy}, of choosing sound x over sound y with respect to its audio quality. In order to test whether the judgments fulfill the structural requirements necessary for a scale to be derived, the number of violations of the weak (WST), moderate (MST), and strong (SST) stochastic transitivities were counted [1]. While WST is a necessary condition for an ordinal scale, choice models that imply higher scale levels require MST

or SST to hold. When SST was satisfied, the Bradley-Terry-Luce (BTL) model [2] was fit, which predicts P_{xy} as a function of parameters associated with each sound, $P_{xy} = u(x)/(u(x) + u(y))$, where $u(\cdot)$ is a ratio scale of the perceived audio quality.

The ranking data were aggregated across participants in oder to estimate the probability of each ranking, $P(R_m)$. In those cases where the BTL model provided an adequate fit, the Mallows-Bradley-Terry (MBT) model [3] was applied. The MBT model predicts that

$$P(R_m) = \frac{\prod_{i=1}^{t} u_i^{t-r_{im}}}{\sum_{m=1}^{t!} \prod_{i=1}^{t} u_i^{t-r_{im}}}, \qquad (1)$$

where t is the number of sounds. This is, in essence, the product of all pairwise probabilities consistent with a given ranking. Since only few of these possible rankings can actually be observed with $n = 52$ subjects, a goodness-of-fit test is not possible. It is assumed that whenever the BTL model fits the corresponding paired-comparison data, the MBT model would fit the rankings, and vice versa. As for the BTL model, the u parameters of the MBT model are a ratio scale. Software used for fitting and testing these models is described in [5, 1, 4].

Results and discussion

Table 1 shows the number of violations of the weak, moderate, and strong stochastic transitivities. There were only few violations of WST, which suggests that an ordinal scale can be derived from the judgments for each of the excerpts. For castanets, MST and SST violations are relatively few, for the remaining two excerpts they seem to be more substantial.

Figure 1 shows the scale values derived for the castanets excerpt. The scale values are the parameter estimates of the Bradley-Terry-Luce model applied to the paired-comparison data and the parameter estimates of the Mallows-Bradley-Terry model applied to the ranking data. Scale values are normalized such that $u(\text{Orig}) := 1$. For castanets, the perceived audio quality spans a considerable range: Since $\hat{u}(\text{M}_{64}) = 0.05$, the uncompressed version has an audio quality about 20 times higher than the MP3 at 64 kbps for the paired comparisons.

Comparing the time participants needed to complete an evaluation of the compression settings using a given method, paired comparisons took longest on average ($M = 242$ s) for all types of program material. The variances were smallest for the paired comparisons, and they were very large for the ranking methods. There is a tendency that the ranking by elimination procedure is slightly faster than the regular ranking; this is significant for castanets, $M_r = 135$ s, $M_e = 117$, $t(51) = 2.26$, $p = .028$, but not for the pop, $M_r = 171$, $M_e = 156$, and the classical excerpt, $M_r = 169$, $M_e = 161$.

Table 1: Number violations of the stochastic transitivities and goodness-of-fit tests of the BTL model for each excerpt.

	WST	MST	SST	$G^2(15)$	p
Castanets	0	1	8	13.18	.588
Pop	4	13	20	21.01	.137
Classic	2	12	29	27.03	.029

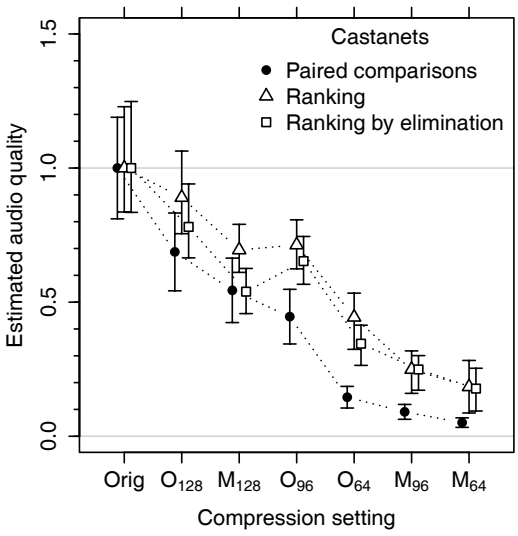

Figure 1: Parameter estimates of the BTL model with approximate 95% confidence intervals (pairwise comparisons) and of the MBT model (ranking and ranking by elimination).

It has been shown that all three methods for sound quality evaluation derive similar result patterns concerning perceived quality of the stimuli, but differ with respect to the time needed to gain these results. The herein introduced ranking-by-elimination procedure seems to be faster than a common ranking procedure when easy stimulus material is used. Both ranking procedures are faster than paired comparisons. This speed gain, however, comes at the price of large individual time differences.

References

[1] Choisel, S. & Wickelmaier, F. (2007). Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. *J. Acoust. Soc. Am.*, *121*, 388–400.

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

[3] Mallows, C. L. (1957). Non-null ranking models: I. *Biometrika*, *44*, 114–130.

[4] Wickelmaier, F. (2011). Elimination-by-aspects (EBA) models. R package manual, version 1.7-0.

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

[6] Wickelmaier, F., Umbach, N., Sering, K., & Choisel, S. (2009). Comparing three methods for sound quality evaluation with respect to speed and accuracy. *126th Conv. Audio Eng. Soc., Munich, Germany, May 7–10*. Preprint 7783.

[7] Zieliński, S. K., Rumsey, F., & Bech, S. (2008). On some biases encountered in modern audio quality listening tests – a review. *J. Audio Eng. Soc.*, *56*, 427–451.