7 - 12 SEPTEMBER 2014, KRAKÓW
FORUM ACUSTICUM

A method to uncover salient auditory features of gear unit sounds

Wolfgang Ellermeier, Angelika Kern, Josef Schlittenlacher
Institut für Psychologie, Technische Universität Darmstadt, D - 64283 DARMSTADT, Germany.

Summary
Triadic comparisons have been proposed as an indirect method for identifying the auditory attributes relevant for a given domain [Wickelmaier & Ellermeier, Perception & Psychophysics 69, 287-297 (2007)], without actually requiring listeners to name them. While the technique has been applied to simple synthetic sounds and to the spatial auditory reproduction of music with some success, the present investigation studied the complex auditory features of sounds emitted by gear units. To that effect, a sample of 15 listeners performed triadic comparisons indicating (with a 'yes' or 'no') whether the first two sounds had something in common that distinguished them from the third. Given certain requirements (replicability, transitivity), a lattice structure may be derived to represent the implicit auditory features. The data of only half of the participants, however, were consistent enough to be represented by such a structure. Their individual representations showed considerable agreement, and the majority of the features obtained were related to well-known psychoacoustic attributes such as loudness or tonal content.

PACS no. 43.66.Pn, 43.66.Cb

1. Introduction

Whenever auditory researchers set out to study a new domain using listening tests (e.g., electrical-vehicle sounds or the noise emitted by wind turbines), the question arises, which auditory attributes listeners might perceive in a given set of sounds. Thus before being in a position to ask listeners to quantify their sensations on different dimensions, investigators will have to determine which auditory sensations are elicited by the sounds in question in the first place. Often, like in the examples given above, there is not much prior empirical evidence as to what these sensations might be.

In the field of audio engineering, so-called 'elicitation' techniques have been developed to obtain the relevant dimensions from a set of (typically trained) listeners or experts. One of these techniques has been termed 'descriptive analysis' and involves guiding a panel of experts through a series of controlled exposures to the sounds in question [1]. Another one is based on the 'repertory grid technique' and requires detecting and labelling the similarities and differences between subsets of the sounds [2]. These and similar techniques have in common that the listeners will eventually have to explicitly label their auditory experiences, which is problematic for two reasons: (1) They might not have the vocabulary to do so, especially when confronted with an entirely novel set of sounds. (2) Even if they produce verbal descriptors, their reliability and validity remains doubtful and will be difficult to evaluate.

Therefore, psychoacousticians have borrowed methods from the theory of knowledge spaces [3] and the analysis of semantic concepts [4] to determine which auditory features characterize a set of sounds without requiring their participants to produce verbal labels. Heller [5] has formulated conditions under which a feature representation may be construed and has proposed a methodology involving triadic comparisons to derive semantic or perceptual structures for a given set of words or objects.

The method (a more strictly formal characterization of which may be found in [6]) requires all possible 'triple comparisons' of the form $\{a, b\}Q\{c\}$ to be made for a set of objects (e.g. sounds), with the relation Q being determined by the following question:

> Do the first two sounds have something in common which makes them different from the third?

If the domain is structured into subsets of sounds sharing common features, these triple comparisons should be transitive, i.e. responded to in a consistent fashion. Transitivity requires that if

$$\{a,b\}Q\{c\} \quad \text{“No”} \tag{1}$$
$$\{b,c\}Q\{d\} \quad \text{“No”} \tag{2}$$
$$\Rightarrow \{a,b\}Q\{d\} \quad \text{“No”}, \tag{3}$$

meaning that two 'inclusive' responses (eq. 1 and 2; implying that a pair of sounds is *not* distinguished from a third sound) entail a third response (eq. 3) to be 'inclusive' as well. If the subject responded “Yes” to the third question, that constitutes an intransitivity. Consequently collecting all triple comparisons from a listener on a set of sounds permits evaluating the data set with regard to transitivity, and - only if it is fulfilled - deriving an auditory structure for that listener.

The present study was designed to explore, whether the previously outlined approach to identify auditory features might be useful for industrial applications as well, namely in the domain of gear noise. Gear units are present in many industrial and everyday products (e.g. lifts, conveyor belts), and the industry has recently become more and more concerned with their contribution to overall sound quality. Thus, a selection of recordings from industrial gear units was subjected to the triadic comparison method in order to try to identify the sound features governing that domain.

Initially, two sets of gear noise recordings defined by different technical specifications will be subjected to our methodology in order to determine, whether common auditory features may be consistently identified in that domain. Subsequently, an attempt shall be made to interpret these common features by relating them to (a) instrumentally measured sound metrics such as loudness and tonal content, and (b) adjective ratings on a large number of attributes pertinent to gear sounds made by a different set of listeners.

2. Method

2.1. Subjects

Fifteen listeners (10 female, 5 male) between 19 and 45 years of age ($M = 25.5$, $SD = 6.9$) all of whom were psychology students and received course credit for their participation took part in the experiments.

2.2. Apparatus and stimuli

2.2.1. *Stimuli*

The stimuli employed in the present experiments were selected from a large set of recordings of industrial gear units provided by SEW-EURODRIVE (Bruchsal, Germany). They had been made in a semi-anechoic chamber using a single microphone (Brüel & Kjær Type 4190) that was placed at 1 m distance from the most sound-emitting area of the gear unit.

For the present set of experiments, nine recordings were selected (see Table I) to reflect a wide range of sound attributes and operating conditions. For making triadic comparisons, they were divided into two partially overlapping sets, with set 1 primarily varying the type of gear mechanism and set 2 primarily varying rotational speed (between 750 and 3000 rpm). For better comparability, only recordings made while the gear unit was running in idle mode were used.

Table I. Gear unit recordings used in the experiments.

sound	*sound set*	*gear type*	*load*	*rpm*
1	1	C	0	1500
2	1, 2	A	0	1400
3	1, 2	B	130	1470
4	2	B	130	2400
5	2	C	0	750
6	1, 2	C	0	1500
7	2	C	0	3000
8	1	B	0	1500
9	1	B	150	1500

The recordings were shortened to a uniform duration of 5 s and played back at their original level yielding loudness values between 9 and 50 sones, calculated according to DIN 45631/A1 [7]. They were D/A converted with a sampling rate of 48 kHz and 16-bit resolution.

2.2.2. *Apparatus*

A personal computer (PC) was used for controlling the experiment and registering responses. The stimuli were D/A converted by an external sound card (RME Hammerfall DSP Multiface II). Subsequently they were amplified by a headphone amplifyer (Behringer HA8000 Powerplay Pro-8), and diotically delivered to diffuse-field equalized headphones (Beyerdynamics DT-990). The experiment was carried out in a double-walled sound-attenuated (Industrial Acoustics Company) chamber.

2.3. Procedure

All Participants were exposed to the six sounds in a given set and had a chance to listen to them repeatedly. Only after this familiarization phase did the experiment proper start. On each trial three gear sounds of 5 s duration were presented successively. During the entire trial, the instruction 'Haben Geräusch A und B etwas gemeinsam, was C nicht hat?' (German for: 'Do sounds A and B have something in common that C doesn't have?') was displayed, along with a 'Yes', a 'No' and a 'Repeat' button. Only after all sounds were presented could the subject press one of the two response buttons, or repeat the presentation of the triple. $[6 \times (6-1) \times (6-2)]/2 = 60$ triples were presented per sound set, requiring approximately 30 min.

For 10 of the 15 participants, the procedure was repeated in a second session on another day (these data will be refered to as 'block 2' in the data analysis), this time working on the two sound sets in the reverse order. At the end of the second session, only those trials

Table II. Number of response changes between repetitions for sound set 1.

participant	Block I-II		Block II-III	
	total	*percentage*	*total*	*percentage*
ANON02	22	36.67	6	27.27
ADNE03	13	21.67	7	53.85
AZIE04	16	26.67	10	62.50
ANNA06	31	51.67	12	38.71
ARRA08	17	28.33	12	70.59
ANLO09	16	26.67	3	18.75
ADNA10	12	20.00	9	75.00
AMTE11	21	35.00	7	33.33
ZYNA13	23	38.33	3	13.04
ENEL15	21	35.00	8	38.10
Mean	19.20	32.00	7.70	43.11
SD	5.61	9.36	3.20	21.48

on which a given participant disagreed between session 1 and 2 were repeated a third time ('block 3') to resolve inconsistencies.

3. Results

Even though participants judged two sets of sounds using the triadic-comparison method, this paper presents only results from sound set 1. The outcome for sound set 2 was quite similar and will be summarized in a subsequent report.

Initially, the data will be analyzed with respect to their reliability and consistency, with reliability refering to the degree in which identical trials produce identical responses (or relatively few response changes across blocks), and consistency being operationalized as a lack of transitivity violations as defined in Eq. 1-3.

3.1. Response changes between sessions

For the 10 listeners who participated in two sessions, the reliability of the triadic comparisons may be determined by inspecting for how many trials (of a total of 60 per set) they changed their responses between block 1 and 2 (see Table II). Furthermore, since - at the end of session two - only the inconsistently evaluated trials were presented a third time, it may be seen how often response changes occurred between block 2 and 3, or how often the subject reverted back to the decision originally made in block 1 (see the last two columns of Table II).

It is evident that responses in the triple comparisons are of only moderate reliability: When the identical trials were repeated in a different order, roughly one third (32%) of the responses changed (see Table II). When those inconsistent trials were presented a third time, an even greater proportion (43%) was reverted back to the original judgment. Note, that if the subject responded randomly to the triple question (by saying 'yes' or 'no'), 50% change were to be expected. A sign test on the individual percentages revealed that response changes between block I and II were significantly different from chance ($p = .02$), while response changes between blocks II and III were not: Here, 4 of 10 participants had changed more than 50% of their responses (see the last column of Table II).

3.2. Transitivity of triadic comparisons

The transitivity of the judgments may be assessed (a) after subjects have completed a block of 60 trials (block I, 15 participants), (b) after subjects have repeated the jugments in a second session (block II, 10 participants), and (c) after the 'inconsistent' trials have been presented a third time (block III, 10 participants). For the latter analysis, the judgments made on the third occasion (and interpreted to settle the issue) were incorporated into the response matrix of block I.

Violations of transitivity reveal inconsistencies in the subjects' responses and were determined based on triples of triadic comparisons containing a 'No' response in the premises, as defined in Eq. 1-3. They were identified using a computer program cited in [9] and subsequently checked by inspection of the respective trials. Table III summarizes these analyses by listing the absolute number of transitivity violations and estimates of their proportion of all relevant triples. Subjects produced a median number of 32 transitivity violations in their first block, corresponding to 6% of all theoretically possible ($60 \times 3 \times 3$) tests or 11% of those tests having two 'No' responses in the premises. That proportion remained approximately the same in the second block ($Mdn = 31$) and dropped somewhat (to a median value of $Mdn = 19$) when the data set combining block I with the corrections made in block III is considered. A Wilcoxon signed-rank test showed that the number of transitivity violations did not differ significantly when comparing blocks I and II ($V = 34; p = .54$) but was significantly reduced (i.e. for 8 of 10 participants, see Table III) when comparing the 'final' outcome including block III with block II, $V = 49; p = .03$.

3.3. Lattice representations

In order for a lattice representation to be made, the data have to be consistent, i.e. devoid of any transitivity violations. Since it may be assumed, however, that some of the transitivity violations produced by our subjects are due to inattention, careless errors, and the like, it was investigated, how many responses of a given subject would have to be changed to produce a consistent data set.

The last column of Table III shows for each participant how many responses had to be altered until an entirely transitive data set was obtained. That number ranges from 1 to 13 with a median value of

Table III. Transitivity violations in block 1, block 2 and the combined data. The last column lists the response reversals required to make a subject's data set transitive.

participant	Block I		Block II		I & III combined		responses
	total	*percent*	*total*	*percent*	*total*	*percent*	changed
ANON02	63	0.12	52	0.10	46	0.09	11
ADNE03	15	0.03	22	0.04	19	0.04	4
AZIE04	18	0.03	22	0.04	16	0.03	7
ANNA06	54	0.10	55	0.10	19	0.04	6
ARRA08	21	0.04	50	0.09	9	0.02	4
ANLO09	22	0.04	26	0.05	2	0.00	1
ADNA10	32	0.06	25	0.05	22	0.04	6
AMTE11	71	0.13	64	0.12	61	0.11	9
ZYNA13	47	0.09	36	0.07	27	0.05	7
ENEL15	27	0.05	8	0.01	16	0.03	3
DNRC01	35	0.06					9
ARIA05	19	0.04					5
LOLK07	36	0.07					9
ESPG12	74	0.14					13
ATIK14	24	0.04					7
Median	32	0.06	31	0.06	19	0.04	7

($Mdn = 7$) response alterations required. It was decided that changing no more than 10% (i.e. 6 of 60 triadic comparisons) is acceptable to represent a subject's data by an auditory structure, thus leaving 7 subjects' data sets for a lattice representation.

If a representation is possible, it can be illustrated in the form of a lattice graph that is sometimes called a 'Hasse diagram'. In such a graph, the bottom nodes represent the entities to be classified (here: the six sound recordings), and higher nodes represent features shared by the nodes to which connecting lines are drawn. These higher nodes can also be characterized as the set of elements defined by the lower nodes. For example, $\{1, 2, 9\}$ might be a superordinate node depicted as a filled circle in a lattice graph (as for subject 'ANNA06' in Figure 1), implying that sounds 1, 2, and 9 have an auditory feature in common.

Figure 1 shows Hasse diagrams for the 7 participants for whom transitive judgments were obtained. These lattices vary considerably in complexity, having four (subjects ANLO09, ENEL15) to nine superordinate nodes (subject ANNA06). Nevertheless, some common features emerge: All subjects distinguish sound 2 from the rest of the recordings $\{1, 3, 6, 8, 9\}$, and four of seven participants treat $\{1, 9\}$ as sharing a feature. Based on the features shared by more than half of the participants, yet another tree representing these common features was drawn (see the bottom right panel of Figure 1).

3.4. Instrumentally measured sound metrics

In order to interpret the auditory structures obtained, two strategies were employed: (a) relating them to instrumentally measured metrics computed from the signals directly, and (b) using ratings of the sounds performed by a different set of listeners on a number of verbal descriptors of auditory qualities.

Standard instrumental metrics were obtained from the (monophonic) recordings using sound-quality analysis software (ArtemiS 12). From the large number of metrics initially computed, loudness according to DIN 45631/A1 [7], sharpness [8], roughness (Artemis algorithm), tonality and fluctuation strength – all based on their 'statistical' versions, i.e. using those values exceeded only 5% of the time – turned out to account best for the qualitative differences observed in the gear sound lattices. In order to characterize the feature(s) shared by a given set of sounds (or node), the instrumental metrics of all elements of that set were computed, and it was determined, whether the sounds in question clustered on a given metric when compared to the others, i.e. whether their values were all smaller or all larger than those of the remaining sounds, or whether they fell into the mid-range without interjacent elements from the other set. The results of this analysis are illustrated in Figure 2.

When the tree structure comprising the most frequently encountered common nodes of all 7 participants is inspected (left panel of Fig. 2), it appears that the sound being at least twice as loud as the others, sound $\{2\}$, is judged as distinct from the remainder $\{1, 3, 6, 8, 9\}$ which scores lower in loudness, roughness, and fluctuation strength. The next node below, $\{1, 3, 6, 9\}$, places sound $\{8\}$, the softest and least tonal, on a separate branch. A further subset consisting of sounds $\{1, 9\}$ comprises sounds that are comparatively low in sharpness and fluctuation strength.

3.5. Independently rated sound attributes

Since in a parallel study [10], 57 sounds – including the present set 1 and set 2 – were judged according to 16 subjective auditory attributes by a sample of 19 naive listeners, the outcome of the present experiment may

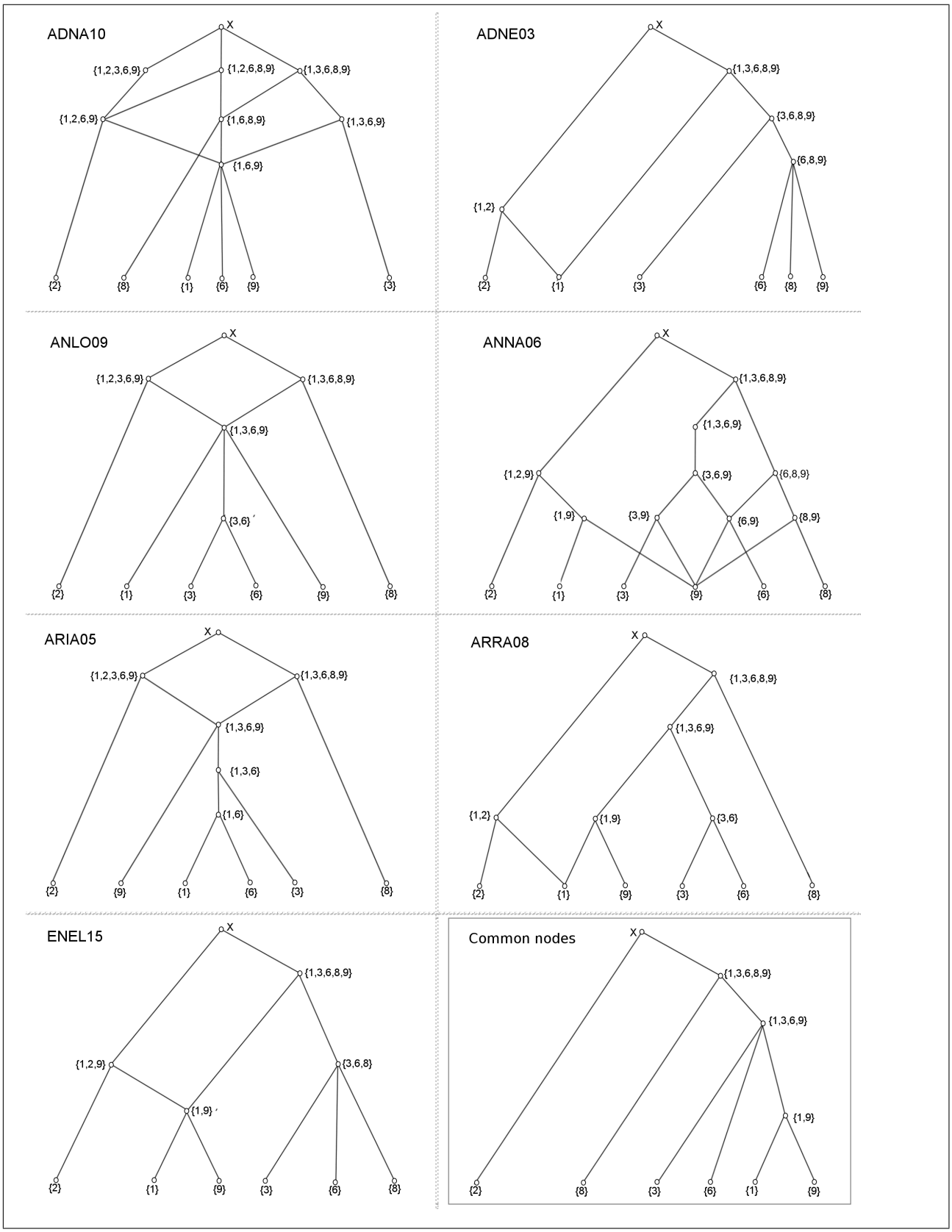

Figure 1. Hasse diagrams for the individual listeners and for common nodes identified by the majority (bottom right).

be tentatively interpreted in terms of these attribute ratings as well. The bottom panel of Fig. 2 shows the common nodes for sound set 1, interpreted in terms of the subjective ratings of the same sounds made by

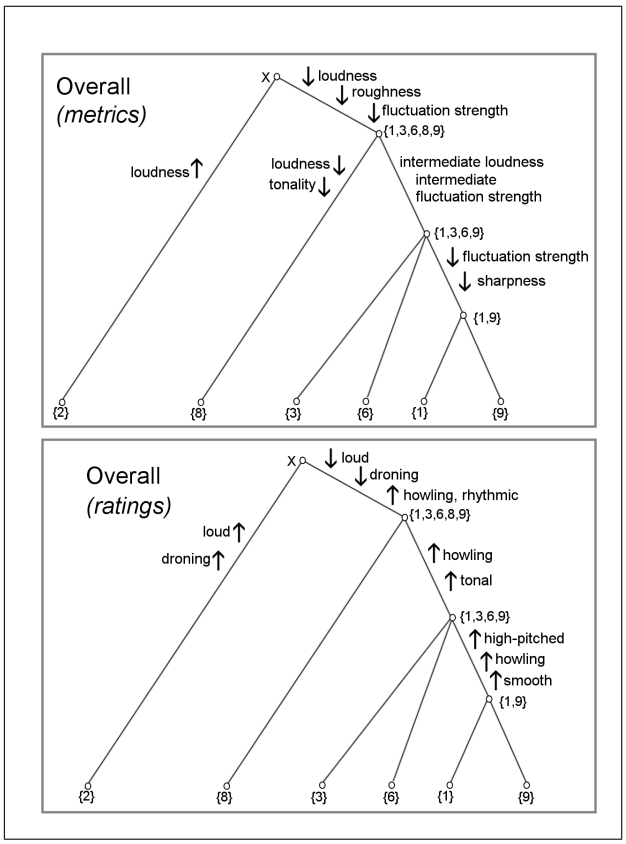

Figure 2. Hasse diagrams for the common nodes identified by the majority of listeners interpreted via instrumental metrics (top) and sound ratings (bottom). Upward-pointing arrows indicate the node right below to have larger values on the attribute in question than the remaining sounds; downward-pointing arrows indicate smaller values.

this different group of listeners. Consistent with the instrumental measurements, sound {2} distinguishes itself from the remainder by being *loud* and *droning* {2}. In addition to having less of these attributes, the node $\{1, 3, 6, 8, 9\}$ is perceived as sounding more *howling* and *rhythmic*. Furthermore, the subset $\{1, 3, 6, 9\}$ is rated more *tonal* than the rest. The subnode $\{1, 9\}$ is distinguished from all other sounds in set 1 by being rated more *high-pitched*, *smooth*, and *howling*.

4. Discussion

The present study shows that it is possible to use a non-verbal querying technique employing triadic comparisons to identify common auditory features in a set of machinery sounds. These features – interpreted with the help of additional instrumental measurements and subjective ratings of the same sounds – are largely interpretable in terms of known sound descriptors such as loudness or fluctuation strength. In some cases, however, it might be worth exploring new, domain-specific auditory attributes, such as the 'droning' sound sensation implied by the overall outcome of the present study.

The theoretical appeal of the procedure used, its objectivity and methodological rigour stand in sharp contrast, however, to its limited empirical success: In the present study, individual feature judgements were of low reliability, and less than half of the participants made triadic comparisons consistent enough to be modeled by a lattice structure. Further studies will have to show, whether these deficiencies are due to (a) participants discovering new sound features as they become familiar with the domain, (b) having trouble with maintaining a consistent set of features, or (c) misinterpreting the task as one of judging similarities rather than identifying the absence or presence of perceptual features. Of particular relevance to settling some of these issues might be a comparison between gear-noise experts and novices in that domain.

Acknowledgement

Portions of the data were presented at The 29th Meeting of the International Society for Psychophysics (Fechner Day), Freiburg, Germany, October, 2013.

References

[1] S. Bech, N. Zacharov: Perceptual audio evaluation: theory, method and application. Wiley, Hoboken, 2006.

[2] J. Berg, F. Rumsey: Identification of quality attributes of spatial audio by repertory grid technique. J. Audio Eng. Soc. **54** (2006) 365–379.

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

[4] B. Ganter, R. Wille: Formal concept analysis. Springer, Berlin, 1999.

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

[6] F. Wickelmaier, W. Ellermeier: Deriving auditory features from triadic comparisons. Percept. Psychophys. **69** (2007) 287–297.

[7] DIN 45631/A1: Berechnung des Lautstärkepegels und der Lautheit aus dem Geräuschspektrum - Verfahren nach E. Zwicker - Änderung 1: Berechnung der Lautheit zeitvarianter Geräusche (2010).

[8] W. Aures: Berechnungsverfahren für den sensorischen Wohlklang beliebiger Schallsignale. Acustica **59** (1985) 130–141.

[9] S. Choisel, F. Wickelmaier: Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound. J. Audio Eng. Soc. **54** (2006) 815–826.

[10] J. Schlittenlacher, W. Ellermeier: Psychoacoustic evaluation of gear noise using category ratings of multiple attributes. Proc. of the 42nd Conference on Noise Control Engineering (Inter-Noise 2013), Innsbruck, Austria.