forum acusticum 2026
8-12 September · Graz, Austria

TONAL VS. NOISE-LIKE SHARPNESS: AN INDIVIDUAL VOCABULARY PROFILING STUDY

Matt Torjussen[1], Roland Sottek[2], Zuzanna Podwinska[1], Antonio Torija Martinez[1]

[1]*Acoustics Innovation Institute, University of Salford, 43 Crescent, M5 4WT Salford, UK*

[2]*HEAD acoustics GmbH, Ebertstr. 30a, 52134 Herzogenrath, Germany*

Abstract

Sharpness is a well-established psychoacoustic attribute describing high-frequency spectral balance, but current objective measures treat it as a single dimension regardless of tonal content. This study was motivated by whether tonal and noise-like sounds give rise to perceptually distinct types of sharpness and whether metrics should accommodate this distinction. Fourteen loudness-matched stimuli (pure tones, harmonic tone complexes, filtered noise, and tone-in-noise mixtures) span a two-dimensional space of sharpness and tonality. Selected pairs are contrasted by their tonal or sharpness content and participants are asked to generate their own descriptive vocabulary through pairwise comparison, then rate every stimulus against their elicited constructs.

After a priori removal of attitudinal terms, the pooled constructs are analysed by hierarchical agglomerative clustering with bootstrap stability assessment, principal component analysis, and generalised Procrustes analysis. The vocabulary shows one reproducible division, tonal against noise-like, crossed by two graded dimensions that recover the four designed stimulus conditions. This study contributes to a wider project developing an improved model of psychoacoustic sharpness; the reduced lexicon extracted here will carry forward to a confirmatory experiment quantifying how tonality moderates perceived sharpness.

Keywords: *sharpness, tonality, individual vocabulary profiling, psychoacoustics*

1. Introduction

Sharpness is the auditory sensation associated with the balance of high-frequency energy in a sound: it distinguishes a hiss from a rumble and is a dominant dimension of timbre [1]. Its standardised metric, DIN 45692:2009 [2], computes sharpness in the unit "acum" as the first spectral moment of a stationary specific loudness pattern under a high-frequency weighting. The procedure descends directly from von Bismarck's experiments on synthetic, steady-state, loudness-equalised sounds [3], and carries those origins into real-world use.

Among the standard's documented limitations, one has not been directly investigated: tonal and noise-like signal components of equal spectral position are treated identically, though the two may be perceived differently. Von Bismarck found that, for upper limiting frequencies above 1 kHz, broadband noise is rated somewhat sharper than a harmonic complex tone with the same spectral envelope [3]. Sottek and Becker showed that tonal annoyance diverges from tonal loudness in the same frequency range [4]. Therefore, spectral fine structure, not envelope alone, appears to influence the perception of high-frequency content; however, this is not currently part of the DIN 45692:2009 sharpness models.

This paper presents the results of a listening experiment, the first in a wider programme developing perceptually validated hearing model-based metrics, that examines this question using an elicited-construct method. Section 2 sets out the method, Section 3 the results, and Sections 4 and 5 the discussion, conclusion and further work.

Corresponding author: *m.c.torjussen@edu.salford.ac.uk*

2. Method

2.1 Individual vocabulary profiling

How listeners are asked to respond determines what the data can subsequently show [5]. Conventional given-construct methods such as magnitude estimation, category rating, the semantic differential and paired comparison, all supply the response dimension and ask the listener to rate stimuli against it. This is efficient, but "you only get an answer to what you ask" [6]. The semantic differential, the method through which von Bismarck identified sharpness as the dominant timbral dimension [7], presupposes that the experimenter already knows which descriptors matter; a method that fixes the descriptors in advance cannot reveal an unanticipated distinction like the one under study and could be subject to training bias.

The alternative is to let listeners generate their own vocabulary. On Kelly's personal construct theory [8], each person construes experience through personal bipolar constructs, so no researcher-imposed adjective set represents every listener. Methods built on this insight form a family including Free Choice Profiling, Flash Profile, the Repertory Grid Technique, Individual Vocabulary Profiling (IVP) [9] and pairwise variants such as the Pivot Profile. IVP imposes no pre-defined descriptors, so any sharpness-like vocabulary that appears does so unprompted.

The experiment departed from standard IVP in one deliberate respect. Rather than open comparison of the full stimulus set, which risks constructs driven by incidental differences, sounds were presented as pre-selected pairs chosen so that the most salient contrast was the tonal/noise-like dimension under study. A directional response supplied the comparative anchor that free dyadic description otherwise lacks, using a more/less format similar to the Pivot Profile [10] while letting each pair, rather than one fixed reference, highlight the dimension of interest. Dyadic comparison was preferred to the triadic elicitation of the Repertory Grid Technique, which yields more differentiated constructs [11] but loses control over the acoustic contrast, since a third sound is more likely to introduce extraneous differences.

2.2 Stimuli

Fourteen sounds were synthesised to vary systematically along the tonal-to-noise-like and low-to-high sharpness (per DIN 45692:2009) dimensions under study, holding loudness constant. All stimuli were loudness-matched according to the **S**ottek **H**earing **M**odel Loudness at 4 sone$_{SHM}$ [12]. The set comprises filtered noise, Chebyshev band-pass noise with audible edge tones, pure tones, harmonic tone complexes (HTCs) and tone-in-noise mixtures.

Table 1 summarises the stimuli; Figure 1 locates them in the sharpness–tonality plane with the pairwise contrasts overlaid.

Table 1. The 14 stimuli: specification, tonality (unweighted Tu$_{SHM}$ and annoyance weighted Tu$_{SHM(aw)}$) and DIN 45692 sharpness S. All stimuli are loudness-matched at 4 sone$_{SHM}$. LP/BP/HP: low-/band-/high-pass filter, CBP: Chebyshev BP filter with steeper slope, HTC: harmonic tone complex.

Sound	Specification	Tu	Tu(aw)	S
S1	LP noise, f_c = 3 Bark	0.03	0.03	0.74
S2	BP noise, f_c = 3 Bark, 3 Bark wide	0.03	0.03	1.51
S3	BP noise, f_c = 17 Bark, 3 Bark wide	0.04	0.06	2.81
S4	HP noise, f_c = 21.5 Bark	0.03	0.03	4.71
S5	CBP noise, f_c = 3 Bark	0.77	0.77	0.42
S6	CBP noise, f_c = 17 Bark	0.62	1.49	2.84
S7	Pure tone at 3 Bark	2.54	2.54	0.41
S8	Pure tone at 17 Bark	3.04	7.21	2.85
S9	LP HTC, f_c = 3 Bark, f_0 = 80 Hz	0.27	0.27	0.66
S10	BP HTC, f_c = 3 Bark, envelope matched to S2	0.12	0.12	1.29
S11	BP HTC, f_c = 17 Bark, envelope matched to S3	1.59	3.77	2.93
S12	HP HTC, f_c = 21.5 Bark, f_0 = 2 kHz	1.33	3.51	5.10
S13	LP noise + tone at 3 Bark	1.56	1.56	0.58
S14	HP noise + tone at 17 Bark	2.24	5.31	4.00

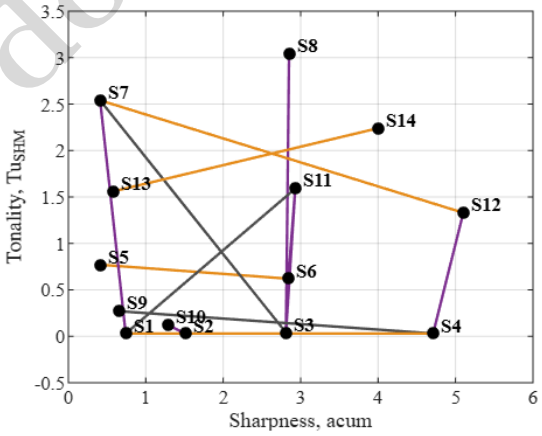

Figure 1. The 14 stimuli in the sharpness–tonality plane with the pairwise contrasts overlaid.

2.3 Pairs and presentation

Thirteen pre-selected pairs were organised into three contrast groups, as overlaid in Figure 1: vertical contrasts, in which sharpness was held approximately constant while tonality varied, including matched-envelope noise versus HTC comparisons; horizontal contrasts, in which tonality was held constant while sharpness varied; and diagonal contrasts, in which both varied, anchoring the extremes of the space. The pairs are listed in Table 2. Each pair was presented twice, in both orders (A–B and B–A), giving 26 presentations in total, order randomised per participant.

Table 2. The 13 sound pairs and their contrast groups.

Pair	Sounds	Contrast group
1	S3, S8	Vertical (tonality varies)
2	S12, S4	Vertical (tonality varies)
3	S11, S3	Vertical (tonality varies)
4	S10, S2	Vertical (tonality varies)
5	S5, S6	Horizontal (sharpness varies)
6	S11, S6	Vertical (tonality varies)
7	S3, S7	Diagonal (both vary)
8	S4, S9	Diagonal (both vary)
9	S1, S11	Diagonal (both vary)
10	S12, S7	Horizontal (sharpness varies)
11	S13, S14	Horizontal (sharpness varies)
12	S1, S7	Vertical (tonality varies)
13	S1, S4	Horizontal (sharpness varies)

2.4 Procedure and participants

The experiment was delivered through a custom MATLAB interface in two phases, each preceded by written instructions and a practice session. In Phase 1 (dyadic elicitation), each pair appeared with two five-second playback buttons, a more/less selector and a single-word text field; the task was to complete the sentence "Sound A is [more/less] _____ than Sound B" for the most salient difference the listener could hear. In Phase 2 (construct rating), each listener's elicited words were deduplicated and editable prior to commencing Phase 2, and the listener then rated every sound against each of their own constructs on a four-point scale (Not at all, To some extent, Very, Extremely), sounds presented in randomised order.

Participants were recruited through the professional networks of the Institute of Acoustics in the UK rather than from a student population, these were people expected to describe sounds routinely without being psychoacoustics experts, whose textbook definitions might otherwise fail to separate the features under investigation. Ethical approval was obtained from the University of Salford's ethics committee; listeners gave signed informed consent and all data were anonymised. Because sharpness perception depends on sensitivity to high-frequency energy, age-related high-frequency hearing loss is a particular concern, and a hearing screen preceded each session, presenting pure tones diotically at octave frequencies from 250 Hz to 8 kHz using a descending threshold procedure. The screen departs deliberately from the BSA recommended procedure [13] in presenting descending runs only and diotic rather than monaural delivery, reducing administration to three to five minutes; it is a screening tool, not a pure-tone audiogram. Sessions took place individually at participants' workplaces in quiet meeting pods with low ambient level, and stimuli were reproduced through open-backed Sennheiser HD 660S2 headphones driven by a calibrated HEAD acoustics PlayPro chain.

2.5 Analysis

Attitudinal constructs were removed a priori (Section 3.3) and the remainder analysed three ways, each answering a different question. Hierarchical Agglomerative Clustering (HAC) asks where the pooled vocabulary separates, with the chosen cut tested by bootstrap resampling of participants. Principal Component Analysis (PCA) asks how many graded dimensions the vocabulary spans. Generalised Procrustes Analysis (GPA) asks whether listeners share a configuration of the sounds once differences in rating-scale use are removed, and places each word in that shared space. All three are computed from the same Phase 2 ratings and are descriptive: they extract structure for a future confirmatory experiment to test. Analyses were performed in MATLAB using the Statistics and Machine Learning Toolbox.

3. Results

3.1 Panel and sensitivity

Thirty-six listeners were recruited, of whom 34 were analysed, two having been rejected for failing to complete the hearing screen. For descriptive, attribute-based methods the literature indicates panels of the order of ten to twenty assessors, with a published individual-vocabulary guideline of $N > 8$ for experienced and $N > 20$ for inexperienced assessors [9].

3.2 Elicited vocabulary and consistency

Participants generated a rich free vocabulary, yielding 412 constructs across the 34 analysed listeners. High-frequency broadband contrasts tended to elicit terms such as *hissy*, *shushing* and *rushing*; high-frequency tonal contrasts terms such as *piercing*, *squealy* and *screechy*. Each listener's reliability was then checked against their own ratings. Each Phase 1 trial declared one sound of a pair to be more *X* than the other, and a trial counted as agreement when the two sounds' Phase 2 ratings on *X* ran in the declared direction, ties counting as one half. Pooled across all trials and listeners, directional agreement was 94.0% against a 50% chance baseline, with every participant above chance, and the mean signed contrast was 1.78 points on the 1 to 4 scale: declared differences were reproduced in direction and in magnitude. Because each pair appeared in both orders, an order-induced reversal would have registered as disagreement. The construct *whoosh*, elicited from a single participant, was removed because it was given an identical Phase 2 rating for all stimuli.

3.3 Construct reduction

Reduction before analysis is typical for this family of methods [6]. It was applied here on linguistic grounds alone, without reference to how the words behaved in the ratings; constructs expressing evaluation or affect rather than a property of the sound, such as *pleasant*, *annoying* and *soothing*, were set aside. Forty-six of the 411 constructs were removed, leaving 365, and because the exclusion precedes every analysis they take no part in what follows.

Variants of one root were pooled onto a canonical form (*buzzing* onto *buzzy*, *stabbing* onto *piercing*) only when counting words for reporting, never prior to fitting, and

purely to produce a candidate attribute set for a subsequent experiment.

3.4 Construct clustering

Because each listener rates on their own vocabulary, ratings cannot be averaged word-by-word across the panel. Each construct was instead represented by its profile of Phase 2 ratings across the fourteen sounds, and similar profiles grouped by hierarchical agglomerative clustering, the conventional treatment of an elicited vocabulary [6], using city-block distance and complete linkage.

The number of clusters was not imposed and the largest gap in the agglomeration schedule falls at two. An independent analysis of the same ratings using correlation distance, average linkage and a silhouette criterion also returns two. The division, shown in Figure 2, is broadly tonal against noise-like. The first cluster (130 constructs) collects the tonal and high vocabulary, *tonal*, *piercing*, *clear*, *high* and *sharp*; the second (235) the noise-like and low, *hissy*, *natural*, *buzzy*, *broadband*, *deep* and *rumbly*. The boundary does not fall along tonality alone: low-pitched tonal words such as *droney* sit in the second cluster, so the division separates tonal-and-high description from the remainder.

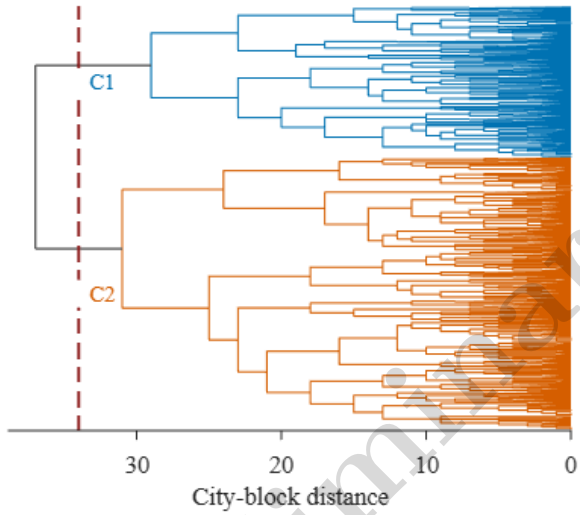

Figure 2. Hierarchical clustering of the 365 retained constructs, each one listener's use of one word across the 14 sounds (city-block distance, complete linkage). Leaf labels are omitted and branches are coloured by cluster. The dashed line is the cut at d = 34.0, the largest gap between successive merge heights, giving a tonal pole of 130 constructs and a noise-like pole of 235.

Hierarchical clustering imposes a hierarchy whether or not one exists, so the two-cluster cut was tested by bootstrap resampling of participants rather than constructs, since one listener's constructs share that person's scale use. Over 1000 replicates the mean adjusted Rand index against the full-sample cut was 0.642, and per-cluster Jaccard recovery [14] was 0.756 for the tonal cluster and 0.851 for the noise-like. The division is reproducible, the noise-like pole the more coherent and the tonal pole, which carries *piercing*, *sharp* and *shrill*, the weaker.

3.5 Principal component analysis

How many graded dimensions the vocabulary spans is a separate question. A single correlation PCA was run on all 365 standardised construct profiles, with the number of components assessed by permutation parallel analysis, a component being kept only where it explained more of the variation than shuffling the ratings produced. Two dominant components accounted for 56.8% of construct variance; two further components cleared the chance level only narrowly, taking the total to 76.1%. Two dimensions are secure, a further two possible but weak.

Neither leading component aligns with a single design factor. PC1 runs from noise-like-and-low vocabulary (*broadband*, *rumbly*) to tonal-and-high (*tonal*, *shrill*, *piercing*), PC2 from hiss-like (*hissy*, *sibilant*) to tonal-and-low (*deep*, *hummy*, *droney*). The components are the diagonals of the tonality × sharpness design, and the space is stretched along one of them: the run from tonal-and-high to noise-like-and-low carries twice the variation of the other, so tonality and sharpness are not heard independently. The two marginal components lie outside the design, collecting roughness and temporal fluctuation, attributes not fully engineered out of the stimuli, to which Section 4 returns.

3.6 Consensus of sounds and words

Clustering and PCA pool constructs directly, so a listener who used more of the rating scale weighs more heavily. Free-choice profiling with generalised Procrustes analysis [15] removes this: each participant's rating matrix was matched to a running consensus by rotation, reflection and isotropic scaling, preserving how each listener arranged the sounds. The consensus accounted for 77.0% of the fitted variation, its two leading dimensions carrying 74.8%, and the scaling factors varied by more than a factor of two across listeners, the very difference in range use the method exists to remove.

Figure 3 shows the sounds as the panel jointly placed them: the four designed conditions sit in four separate regions of the plane. The first dimension tracks tonality and the second sharpness, each with little contamination from the other, so the plane recovered from the ratings alone already corresponds to the design; placements below are read on the dimensions as extracted, without rotation.

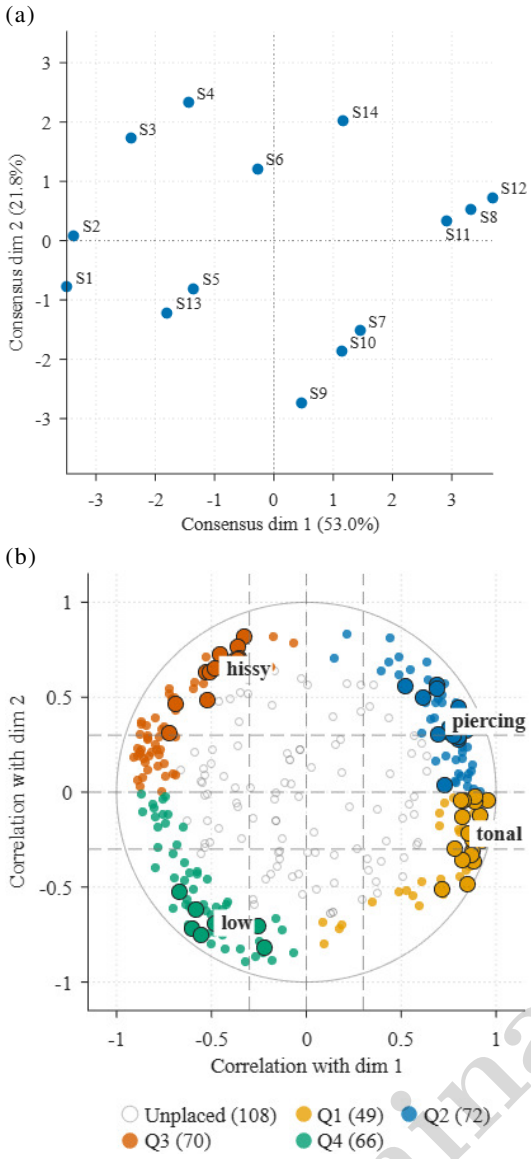

Figure 3. The free-choice profiling consensus. (a) The fourteen sounds in the leading plane. (b) Every construct as its correlation with the two dimensions; open grey markers are the 108 whose position in the plane does not survive correction, and one word per region is emphasised and labelled at the centroid of its instances. Q1 tonal-low, Q2 tonal-high, Q3 noise-high, Q4 noise-low.

Placing individual words is harder: each construct's position rests on just fourteen ratings, and with 730 correlations tested at once, chance alone produces large values. Placing a word in one corner of the design, axis by axis, is beyond an experiment of this size and after correction for multiple testing [16] just one construct in 365 survives. A fairer question is whether a word has a reliable place in the plane at all, that is whether the plane accounts for more of its ratings than chance would, and on that criterion 257 of the 365 constructs pass, with 72 tonal-high, 70 noise-high, 66 noise-low and 49 tonal-low. The strongest placements are the words elicited most often: hissy for noise-high, piercing and high tonal-high, tonal for tonal-low, and low and deep for noise-low (Figure 4).

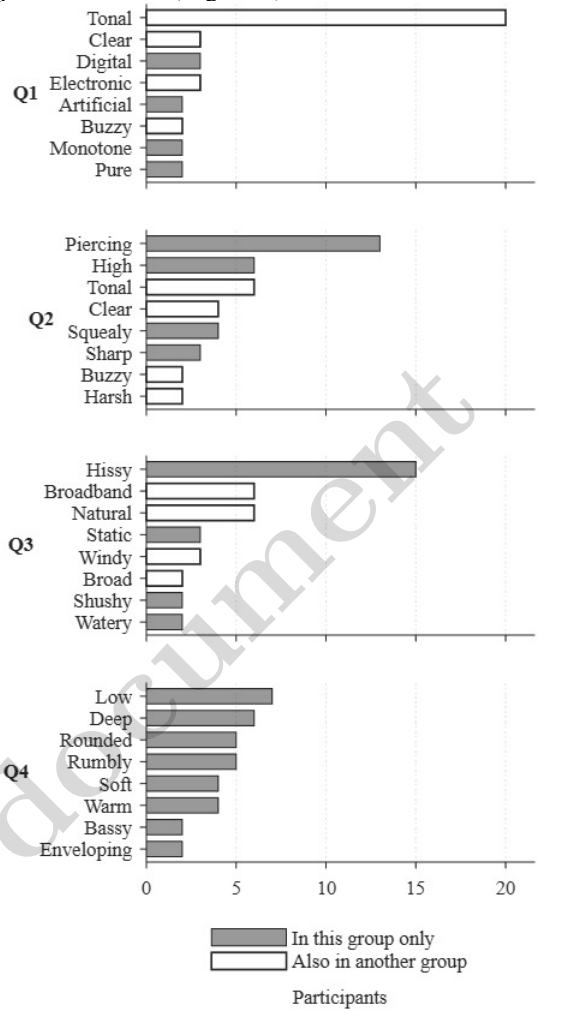

Figure 4. The placed vocabulary pooled across listeners, one panel per region of the plane: Q1 tonal-low, Q2 tonal-high, Q3 noise-high, Q4 noise-low, each showing its eight most frequent words.

4. Discussion

The three analyses answer different questions: one well-supported division, two graded dimensions stretched so that the clustering boundary lies across the dominant diagonal, and a shared configuration in which the four designed conditions occupy four directions. Clustering and the consensus map describe the same split, since a single straight line across the consensus plane puts more than 90% of constructs on the side their cluster predicts, and lies close to the tonality direction, tilted towards sharpness.

Clustering on rating profiles presumes the constructs entering it mean something stable, which a refinement step with the listeners present [6] would normally achieve. This experiment went straight from elicitation to rating, and each word was elicited as a contrast which Phase 2 removed. Judged only on the pair that produced them the constructs behave well, with 2.3% of trials reversing the declared direction; judged across the full set, a quarter of ratings peak on a sound other

than the one that elicited them. Clustering cannot tell a construct that genuinely groups with its neighbours from one whose meaning has drifted, and drift cannot be removed retrospectively.

Fourteen stimuli cannot separate a specifically tonal sharpness effect from a general tonality effect. What the experiment establishes is that the distinction is available to listeners, and that their vocabulary for it has been extracted.

Two observations point back at the stimuli. The positive pole of the dominant component collects attention vocabulary, *intrusive* and *penetrating* against *background*, so salience may run along the same axis as tonal-and-high against noise-like-and-low. That could mean sharp tonal sounds genuinely demand attention, or only that these stimuli differed in ways the loudness matching did not remove. Roughness and fluctuation, which the design did not set out to vary, are audible in three stimuli (S5 fluctuating, S9 and S10 rough).

5. Conclusion and further work

Listeners spontaneously produced sharpness-like vocabulary, and it divides in two: words for tonal sounds and words for noise-like ones, both at elicitation and in the later ratings. The division is not a quirk of this panel, since resampling the listeners does not change the result. The vocabulary varies along tonality and sharpness, which put the four designed conditions in four separate regions of the shared map. The two high-sharpness regions draw on largely separate word sets: piercing, sharp, shrill and squealy for the tonal stimuli, hissy, broadband and static for the noise-like. Each has one dominant term, piercing for 13 of 34 listeners and hissy for 15, but of the two the noise-like region is the more concentrated, with eight words shared by at least two listeners against twelve for the tonal.

A second experiment is required: fourteen sounds were enough to extract the vocabulary but too few to prove that any word belongs to one corner of the design. The words carried forward will be those with a reliable position in the shared map that sit on the side their cluster predicts. That list becomes a fixed response set on which a panel rates a new stimulus set, quantifying how far tonality moderates perceived sharpness and informing an improved formulation.

6. Acknowledgments

The authors thank the volunteer listeners for their time, the host organisations that provided quiet listening spaces for the test sessions, and HEAD acoustics for providing listening test equipment. Funding provided by HEAD acoustics GmbH and EPSRC as part of the EPSRC Centre for Doctoral Training in Sustainable Sound Futures (EP/ Y034708/ 1).

References

[1] H. Fastl and E. Zwicker: Psychoacoustics: Facts and Models. 3rd edn, Berlin: Springer, 2007.

[2] DIN 45692:2009-08, Messtechnische Simulation der Hörempfindung Schärfe (Measurement technique for the simulation of the auditory sensation of sharpness). Berlin: DIN, 2009.

[3] G. von Bismarck, "Sharpness as an attribute of the timbre of steady sounds," Acustica, vol. 30, no. 3, pp. 159–172, 1974.

[4] R. Sottek and J. Becker, "Tonal annoyance vs. tonal loudness and tonality," in Proc. of Internoise 2019, (Madrid, Spain), 2019.

[5] N. Zacharov (ed.): Sensory Evaluation of Sound. 1st edn, Boca Raton: CRC Press, 2018.

[6] J. Berg and F. Rumsey, "Identification of quality attributes of spatial audio by repertory grid technique," Journal of the Audio Engineering Society, vol. 54, no. 5, pp. 365–379, 2006.

[7] G. von Bismarck, "Timbre of steady sounds: A factorial investigation of its verbal attributes," Acustica, vol. 30, pp. 146–159, 1974.

[8] G. A. Kelly: The Psychology of Personal Constructs. Volume One: Theory and Personality. London: Routledge, 1992.

[9] G. Lorho, "Individual vocabulary profiling of spatial enhancement systems for stereo headphone reproduction," in Proc. 119th AES Convention, (New York, USA), 2005.

[10] B. Thuillier, D. Valentin, R. Marchal, and C. Dacremont, "Pivot profile: A new descriptive method based on free description," Food Quality and Preference, vol. 42, pp. 66–77, 2015.

[11] P. Caputi and P. Reddy, "A comparison of triadic and dyadic methods of personal construct elicitation," Journal of Constructivist Psychology, vol. 12, no. 3, pp. 253–264, 1999.

[12] ECMA International, ECMA-418-2:2025 — Psychoacoustic metrics for ITT equipment — Part 2: Models based on the Sottek Hearing Model. Geneva: ECMA International, 2025.

[13] British Society of Audiology: Recommended Procedure: Pure-tone air-conduction and bone-conduction threshold audiometry with and without masking. BSA, 2018.

[14] C. Hennig, "Cluster-wise assessment of cluster stability," Computational Statistics & Data Analysis, vol. 52, no. 1, pp. 258–271, 2007.

[15] A. A. Williams and S. P. Langron, "The use of free-choice profiling for the evaluation of commercial ports," Journal of the Science of Food and Agriculture, vol. 35, no. 5, pp. 558–568, 1984.

[16] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," Journal of the Royal Statistical Society: Series B, vol. 57, no. 1, pp. 289–300, 1995.