forum acusticum 2026
8-12 September · Graz, Austria

EVALUATION OF PSYCHOACOUSTIC TONALITY MODELS FOR CABIN NOISE ACROSS VEHICLE POWERTRAINS

Zhenxian LI[1], Etienne Parizet[1]

[1] *LVA, UR677, INSA Lyon, France*

Abstract

Tonality strongly affects the perceived quality and annoyance of vehicle interior noise. This study evaluates two widely used tonality metrics—the Aures tonality metric and the Sottek psychoacoustic Hearing Model tonality metric (ECMA-418-2)—against subjective ratings collected in two complementary listening experiments. Stimuli were drawn from chassis-dynamometer recordings of four vehicles representing different propulsion technologies (two electric, one hybrid-electric, one diesel), captured at the driver and rear-passenger head positions under full and partial throttle. Study 1 used 112 unprocessed 3 s segments rated by 33 listeners. Study 2 used 48 order-tracking resynthesised stimuli, which preserved the spectral and tonal structure while suppressing temporal fluctuations and equalising loudness across stimuli; each of 19 listeners rated every stimulus three times in randomised order. Hierarchical clustering of Study 1 ratings revealed two listener subgroups with distinct evaluation strategies: one showed a moderate association with the Sottek metric (r = 0.59) but none with Aures, whereas the other showed only weak associations with either metric; ratings in both clusters also correlated with loudness. In Study 2, with loudness and temporal fluctuations controlled, both tonality metrics were strong predictors of subjective ratings (Sottek: r = 0.77; Aures: r = 0.72). Overall, the Sottek metric outperforms the Aures metric for complex, realistic stimuli, while both metrics perform comparably well under controlled conditions. Although tonality alone is a difficult perceptual attribute for non-expert listeners to evaluate, both metrics remain valuable as input features for higher-level perceptual models such as annoyance prediction.

Keywords: *tonality, vehicle interior noise, psychoacoustics, listening experiment, order tracking*

Corresponding author: *zhenxian.li@insa-lyon.fr.*

1. Introduction

The electrification of road vehicles has changed the acoustic environment inside the passenger compartment. Reducing or removing the broadband combustion noise of an internal combustion engine lowers the overall sound pressure level but also weakens the masking that previously concealed the narrow-band components radiated by the electric powertrain, the gearbox and the auxiliary systems [1]. Early measurements on standard production electric vehicles already documented that, despite the lower overall level, individual tonal components of the traction system stand out clearly above the residual broadband background [2]. Tonal components linked to the electromagnetic excitation of the stator, the pulse-width modulation of the inverter and gear-mesh whine therefore become salient features of electric and hybrid-electric cabin noise, and their perceived strength varies systematically with operating condition [3]. Such tonal content has long been associated with increased annoyance at equal sound pressure level [4], motivating the development of dedicated tonal-annoyance metrics for the electric-vehicle context [5]. Sound-quality assessment of electrified vehicles therefore needs *tonality* metrics that match what listeners actually hear.

Engineering practice distinguishes *tone-prominence* descriptors—such as the Tone-to-Noise Ratio and the Prominence Ratio standardised in ECMA-74 [6]—from full *tonality* models. The former compare a candidate tone to its local spectral background and are useful for ranking individual tones, but they do not integrate contributions across the spectrum or place the result on a perceptual scale, and thus agree only poorly with subjective ratings of complex signals [7]. Two model families address these limitations. The procedure of Aures [8], building on the pitch-salience analysis of Terhardt et al. [9], extracts audible tonal components from the spectrum and combines them through empirical weighting functions into a single tonality value. The *psychoacoustic Hearing Model* of Sottek [10, 11], now standardised in ECMA-418-2 [12], operates instead on a critical-band auditory representation and uses an autocorrelation-based periodicity analysis to estimate a specific tonal loudness modulated by an across-band loudness signal-to-noise ratio [13]. Although both model families are widely deployed in commercial sound-quality software, their relative performance across propulsion technologies has rarely been compared systematically [14, 15]. In naturalistic recordings, tonality is strongly correlated with loudness and temporal fluctuations, so the agreement between a metric and listeners' ratings can be inflated by these confounds; listeners themselves may also differ in whether they evaluate the tonal strength of the signal or the annoyance it elicits [16]. This study takes up both issues with two listening experiments on chassis-dynamometer recordings of vehicles spanning current propulsion technologies (electric, hybrid-electric and diesel). The first confronts the Aures tonality metric and the ECMA-418-2 Sottek metric with subjective ratings of unprocessed cabin-noise segments; the second uses an order-tracking analysis-by-synthesis procedure [17] that suppresses temporal fluctuations and equalises loudness across stimuli, isolating the contribution of tonality.

2. Methods

2.1. Overview of the Two-Study Design

Two listening experiments characterised the relationship between the perceived tonality of cabin noise and objective tonality metrics. Study 1 used a broad set of unprocessed stimuli excerpted directly from in-cabin recordings, retaining the full temporal and spectral complexity of real interior noise across four vehicles, two seating positions and two acceleration regimes. Study 2 used a smaller subset processed by a custom order-tracking analysis-by-synthesis pipeline [17]: the resulting stimuli are temporally stationary and equalised in loudness but preserve the spectral envelope and powertrain-related harmonic structure with high fidelity, thereby isolating the spectral determinants of tonality perception from the confounding effects of loudness and temporal fluctuations.

2.2. Stimuli

2.2.1. Study 1: Original Recordings

The Study 1 stimuli were extracted from a measurement campaign on a chassis dynamometer in a hemi-anechoic chamber (Applus IDIADA NVH facility). Four passenger vehicles spanning three powertrain families were instrumented: two electric vehicles (EV1 and EV2), one full hybrid-electric vehicle (HEV) and one internal-combustion-engine vehicle (ICEV). The ICEV is mechanically a 48 V mild hybrid, but its electric machine does not drive the wheels under the steady throttle conditions used here and it is therefore treated acoustically as an ICEV.

In-cabin sound pressure was acquired with condenser microphones at the inner-ear position of the driver (DIE) and of the rear passenger on the driver side (RPIE), each channel being calibrated before every test session. Two throttle scenarios were applied—wide-open throttle (WOT) and a mechanically held part-open throttle (POT, nominally 20–30 %)—over acceleration ramps spanning 5–110 km/h (5–60 km/h for EV2 under POT; 30–110 km/h for the ICEV, held in third gear). From each (vehicle × throttle × seat) recording, 3 s quasi-stationary segments were excerpted at target speeds spaced approximately every 10 km/h, and segments lying on gear shifts or otherwise non-stationary were discarded after audio-visual inspection of the waveform and spectrogram. Because each vehicle covers a different operating range, the final count per condition is not uniform (EV1: 36; EV2: 24; HEV: 24; ICEV: 28), yielding 112 stimuli in total. The use of chassis-dynamometer recordings follows established practice for cross-vehicle tonality benchmarking of electric-vehicle interior noise [14, 15].

2.2.2. Study 2: Order-Tracking Resynthesised Stimuli

For Study 2, a balanced subset of 48 stimuli (12 per vehicle) was drawn from the Study 1 pool, with selection constrained to be balanced across vehicles and operating combinations (POT–DIE, POT–RPIE, WOT–DIE, WOT–RPIE). The stimuli were processed through a three-stage order-tracking analysis-by-synthesis pipeline implemented in Python.

(i) Order extraction. An instantaneous fundamental frequency $f_0(t)$ is tracked frame by frame by maximising a Harmonic Sum Spectrum

$$S(f_0) = \sum_{h=1}^{H} |X(hf_0)|^2, \qquad (1)$$

with $H = 12$ harmonics. The signal is resampled into the angular domain and frame-wise magnitude averaging of an angular-domain STFT concentrates the energy of each harmonic into a sharp order peak.

(ii) Spectral reconstruction. The averaged order spectrum is merged with the time-domain noise-floor spectrum by element-wise maximum,

$$|\bar{X}_{\mathrm{comb}}(f)| = \max\left(|\bar{X}_{\mathrm{order}}(f)|,\, |\bar{X}_{\mathrm{noise}}(f)|\right), \qquad (2)$$

preserving broadband noise colour alongside sharpened order peaks. A stationary 3 s waveform is synthesised by inverse real FFT with random phases, RMS-matched to the original recording.

(iii) Loudness equalisation. The 48 waveforms were brought to a common Zwicker loudness using the MoSQITo implementation of ISO 532-1 [18], with each stimulus scaled to the mean N_5 across the set. The resulting stimuli are temporally stationary and equal in loudness, so that listeners can attend to spectral tonal salience without confounding cues from amplitude modulation or loudness differences.

2.3. Participants, Apparatus and Procedure

Study 1 enrolled 33 listeners and Study 2 enrolled 19, all volunteers and predominantly students at the host university aged 20–29 years, apart from one listener aged about 60 in each study. All participants reported no history of hearing impairment, and in

Study 2 this was complemented by a clinical hearing examination dated within the previous year and showing no anomalies. Both experiments took place in a semi-anechoic chamber. Stimuli were played back through an RME Fireface UCX audio interface and Beyerdynamic DT 990 PRO open-back headphones under the control of PsychoPy [19]. The headphone reproduction level was calibrated against a laboratory artificial head so that the sound pressure level reproduced at the artificial-ear reference point matched the in-cabin pressure recorded at the corresponding position (DIE or RPIE) during the original measurements. At the start of each session, participants were familiarised with the attribute of *tonality* in two stages: a *demonstration* stimulus (a 1 kHz tone fading in and out of a constant broadband background, presented for listening only) and a *warm-up* block of more realistic stimuli—a stationary narrowband tone on a broadband background, with centre frequency increasing and tone-to-noise ratio decreasing across the series—used to practise the rating procedure. Both could be replayed freely.

Tonality was then rated on a continuous unipolar slider anchored at *Not at all* and *Extremely.* In Study 1, the stimuli were presented once each in fully randomised order. In Study 2, each stimulus was presented three times: the first repetition traversed the stimulus set once in random order, after which the second and third repetitions were drawn in fully randomised order. The within-subject mean across the three repetitions was used for subsequent analyses. The repeated presentations in Study 2 also enabled an assessment of within-subject test–retest reliability.

2.4. Data Analysis

2.4.1. Subject-wise Rating Correction

Inspection of the raw data revealed substantial between-listener differences in slider usage, with individual medians and scale widths varying widely across the 0–100 range. Because all subsequent geometric procedures are sensitive to this variability, every rating r_{sn} was replaced by its panel-centred counterpart

$$r^*_{sn} \;=\; r_{sn} - \bar{r}_s + \bar{r}, \tag{3}$$

where $\bar{r}_s$ is subject s's mean across all trials and $\bar{r}$ is the panel grand mean (49.83 in Study 1, 50.07 in Study 2).

2.4.2. Within-Subject Reliability (Study 2 Only)

Because Study 2 included three repeated presentations per stimulus, within-subject test–retest reliability was assessed using pairwise Pearson and Spearman correlations, mean absolute error, within-sound standard deviation, and intraclass correlation coefficients (ICC(3,k)).

2.4.3. Hierarchical Clustering of Listeners

The corrected ratings were arranged into a subject-by-stimulus matrix and submitted to hierarchical agglomerative clustering with Ward's minimum-variance linkage to identify listener subgroups with distinct perceptual profiles. In Study 1 ($N = 33$), the dendrogram was cut to yield a two-cluster solution. In Study 2 ($N = 19$), the small sample size precluded meaningful clustering, so all participants were analysed as a single group.

2.4.4. Correlation and Regression Analysis

Cluster-wise (Study 1) or group-wise (Study 2) mean corrected ratings were related to the objective metrics via Pearson and Spearman correlations and simple OLS regressions, with robust Huber-loss regression as a sensitivity check. In Study 1, multiple regression models including both tonality predictors simultaneously were also fitted to the individual-level data, with and without prior vehicle-wise centering of the dependent variable.

3. Results

3.1. Study 1 – Original Recordings

3.1.1. Listener clustering and inter-rater agreement

Ward hierarchical clustering of the 33×112 subject–stimulus matrix of corrected ratings yielded a two-cluster solution that partitioned the panel into a smaller group, Cluster 1 ($n_{C1} = 15$), and a slightly larger group, Cluster 2 ($n_{C2} = 18$). The dendrogram in Fig. 1(a) shows that the two clusters separate at a clearly higher linkage distance than the within-cluster sub-branches. A single-factor repeated-measures ANOVA on the full panel showed a significant effect of stimulus on corrected ratings, $F(111, 3552) = 7.31$, $p < 10^{-90}$, $\eta_p^2 = 0.19$, demonstrating that the 112-stimulus set evoked reliable subjective discrimination overall. The same analysis run separately within each cluster revealed a large difference in sensitivity: Cluster 1 yielded $F(111, 1554) = 2.99$, $p \approx 8 \times 10^{-21}$, $\eta_p^2 = 0.18$, with Kendall's coefficient of concordance $W = 0.18$; Cluster 2 yielded $F(111, 1887) = 10.65$, $p \approx 6 \times 10^{-132}$, $\eta_p^2 = 0.39$, $W = 0.38$. Cluster 2 thus displayed roughly twice the inter-rater concordance and effect size of Cluster 1, indicating that its members responded more consistently and with greater sensitivity to the 112-stimulus structure.

3.1.2. Tonality metric performance

Per-cluster mean corrected ratings were related to the two tonality metrics across the 112 stimuli (Fig. 1(b)–(e)). The two clusters showed a double dissociation. In Cluster 2, the Sottek metric was a moderate-to-strong predictor of mean rating ($r = 0.59$, $p < 10^{-11}$), whereas Aures tonality was uncorrelated with rating ($r = -0.03$, $p = 0.79$). In Cluster 1, the pattern was reversed but much weaker: Aures tonality showed a small significant positive association ($r = 0.30$, $p = 10^{-3}$) while Sottek tonality was uncorrelated ($r = -0.02$, $p = 0.86$). Multiple regression on the individual-level ratings confirmed these patterns: for Cluster 2 the Sottek metric was the unique significant predictor (standardised $\beta_{\text{Sottek}} = 0.36$, $t = 17.6$, $p < 10^{-60}$; $\beta_{\text{Aures}} = -0.003$, $p = 0.89$; $R^2 = 0.13$), and for Cluster 1 Aures was the unique significant predictor ($\beta_{\text{Aures}} = 0.13$, $t = 5.2$, $p < 10^{-6}$;

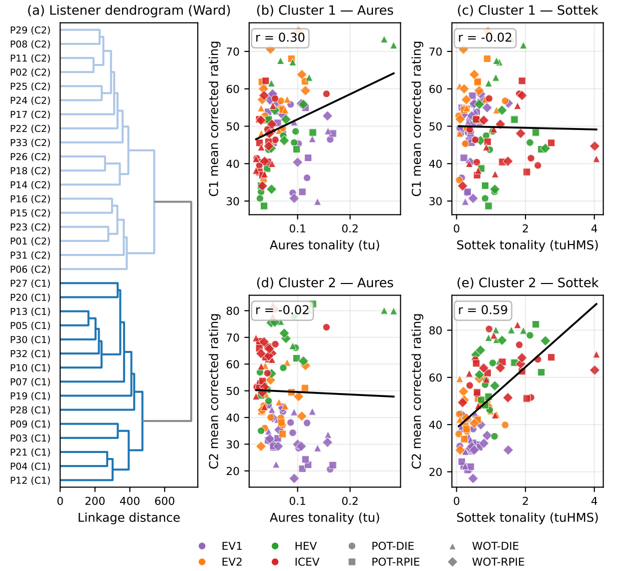

Figure 1: Study 1 overview. **(a)** Ward dendrogram of the 33 listeners; the two-cluster cut yields Cluster 1 ($n = 15$, orange branches) and Cluster 2 ($n = 18$, red branches). **(b–e)** Cluster-mean corrected rating vs. Aures (left) and Sottek (right) tonality; marker colour: vehicle, marker shape: throttle–microphone condition. Solid line: OLS fit; inset: Pearson r. Each metric track sone cluster only.

$\beta_{\text{Sottek}} = -0.003$, $p = 0.91$; $R^2 = 0.02$). Variance inflation factors were close to unity, so the two predictors were not collinear in the present stimulus set.

3.1.3. Broader psychoacoustic context

The Pearson correlations of each available psychoacoustic descriptor with the cluster-mean rating (Figure 4) show that ratings correlated with overall loudness in both clusters, moderately in Cluster 1 ($r = 0.33$) and strongly in Cluster 2 ($r = 0.68$). In Cluster 2 the dominant correlates were Sottek roughness ($r = 0.82$), fluctuation strength ($r = 0.74$), and loudness, all of which exceeded the correlation with Sottek tonality itself; in Cluster 1 the broader metric set produced uniformly weak correlations ($|r| \leq 0.43$). Cluster 2 therefore appears to have responded to a composite tonal-and-modulation salience that the Sottek metric captures only partially, whereas Cluster 1 responded inconsistently to all available descriptors.

3.2. Study 2 – Order-Tracking Resynthesised Stimuli

3.2.1. Within-subject test–retest reliability

Because each of the 48 Study-2 stimuli was presented three times, within-subject consistency could be quantified. Mean pairwise Pearson correlations between repeats ranged from 0.20 to 0.81 across the 19 participants, with ICC$(3, k)$ between 0.44 and 0.92 (median ICC $= 0.79$); all participants exceeded the conventional ICC $= 0.4$ minimum-acceptable threshold. The three least reliable listeners (ICC < 0.6) were retained for the group-level analyses because their inclusion

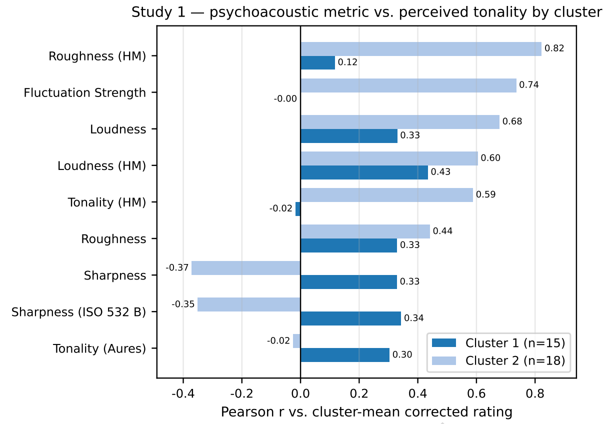

Figure 2: Study 1 – Pearson correlation between the cluster-mean corrected rating and each available psychoacoustic descriptor across the 112 stimuli, sorted by Cluster-2 magnitude (HM: Sottek hearing-model implementation, ECMA-418-2). Cluster 2 is dominated by modulation and loudness descriptors; each tonality metric tracks one cluster only.

did not materially affect the cross-stimulus correlations reported below; sensitivity analyses restricted to ICC ≥ 0.7 subjects yielded near-identical metric rankings (Pearson r within ± 0.03 of the full-sample values).

3.2.2. Stimulus effect

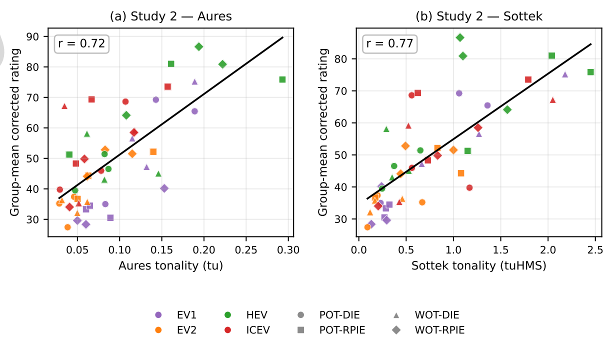

Figure 3: Study 2: group-mean corrected rating vs. **(a)** Aures and **(b)** Sottek (ECMA-418-2) tonality across the 48 resynthesised, loudness-equalised stimuli; markers as in Fig. 1. With loudness equalised, both metrics become strong predictors.

A repeated-measures ANOVA showed a significant effect of stimulus identity on corrected ratings across all 19 participants, $F(47, 846) = 23.19$, $p < 10^{-100}$, $\eta_p^2 = 0.56$. Even with loudness equalised and temporal fluctuations suppressed, the 48 stimuli elicited reliable and substantial differences in perceived tonality.

3.2.3. Tonality metric performance

With overall loudness equalised across stimuli and slow temporal fluctuations suppressed by the resynthesis pipeline, both tonality metrics emerged as strong predictors of the group-mean corrected rating (Fig. 3). The Sottek metric reached $r = 0.77$ ($p < 10^{-9}$) and the Aures metric $r = 0.72$ ($p < 10^{-7}$); the corresponding Spearman correlations ($\rho = 0.80$ and $\rho = 0.66$) were of the same order, so the association does not hinge on the linear form of the fit. The two metrics were close in predictive validity, with a small advantage for the Sottek metric.

3.2.4. Broader psychoacoustic context

Across all eight available psychoacoustic descriptors (Fig. 4), the two tonality metrics were the strongest predictors of perceived tonality in the controlled-stimulus set. Sottek roughness followed at $r = 0.52$, hearing-model loudness at $r = 0.39$, and fluctuation strength at $r = 0.28$. The two sharpness descriptors carried small negative correlations ($r \approx -0.46$ and $r \approx -0.47$), while ISO 532-1 overall loudness was essentially uncorrelated with the ratings ($r = -0.07$): the loudness equalisation had removed the level cue that dominated Cluster 2 ratings in Study 1.

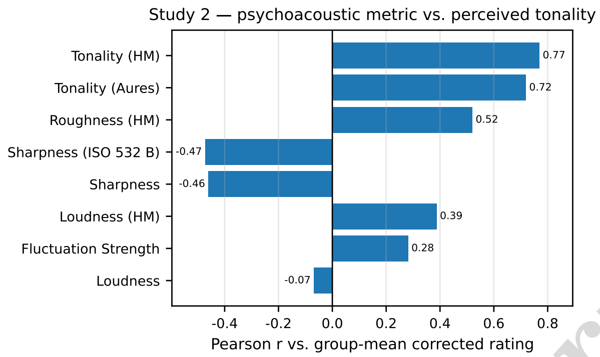

Figure 4: Study 2 – Pearson correlation between the group-mean corrected rating and each of the eight available psychoacoustic descriptors across the 48 stimuli, sorted by absolute strength. The two tonality metrics rank highest; ISO 532-1 loudness is essentially uncorrelated ($r \approx 0$), as intended by the loudness equalisation step.

3.3. Cross-Study Comparison

The comparison between the two studies carries the main result. In Study 1, where stimuli preserved the full temporal and spectral complexity of the original recordings, no single tonality metric tracked the subjective response of every listener: Sottek tonality predicted ratings only for the Cluster 2 subgroup ($r = 0.59$), and Aures tonality only weakly for the Cluster 1 subgroup ($r = 0.30$). Within Cluster 2, the strongest predictors were in fact Sottek roughness and fluctuation strength rather than the tonality metrics, suggesting that the perceived "tonality" of unprocessed cabin noise is intrinsically entangled with temporal-modulation and loudness cues that the listener cannot fully separate. In Study 2, where order-tracking resynthesis suppressed amplitude modulation and the loudness equalisation removed global level differences, both tonality metrics were strong and nearly equivalent predictors (Sottek $r = 0.77$; Aures $r = 0.72$). In short, the Sottek metric outperforms the Aures metric on complex, realistic stimuli, and the two perform comparably under controlled conditions.

4. Discussion and Conclusions

Two listening experiments on chassis-dynamometer recordings of four vehicles (two electric, one hybrid-electric, one diesel) were used to evaluate the Aures tonality metric and the Sottek psychoacoustic Hearing Model tonality metric (ECMA-418-2) against subjective ratings.

On the unprocessed recordings of Study 1, no single metric described the whole listener panel. Sottek tonality predicted the ratings of one subgroup ($r = 0.59$) and Aures those of the other, weakly ($r = 0.30$); within the more consistent subgroup, Sottek roughness ($r = 0.82$) and fluctuation strength ($r = 0.74$) were stronger correlates than either tonality metric. For non-expert listeners, the perceived "tonality" of real cabin noise is thus bound up with loudness and temporal-modulation cues. In Study 2, once order-tracking resynthesis suppressed the temporal fluctuations and the MoSQITo-based ISO 532-1 equalisation removed the loudness cue, both metrics ranked as the top two predictors among the eight available descriptors and performed almost equally (Sottek $r = 0.77$, Aures $r = 0.72$), while loudness itself dropped to $r \approx 0$.

Overall, the Sottek metric keeps a small but consistent advantage on unprocessed stimuli, in line with its more comprehensive auditory representation, and the two metrics converge under controlled conditions. Both remain useful as inputs to higher-level perceptual models such as annoyance prediction.

Several limitations should be kept in mind. The listeners were non-experts, mostly students in their twenties, and tonality proved a difficult attribute for them to isolate, as the Study 1 clustering shows. Study 2 rested on 19 participants, too few for subgroup analysis. All stimuli were 3 s quasi-stationary segments recorded on a chassis dynamometer; whether the same metric ranking holds on the road, with longer and time-varying sounds, remains to be tested.

5. Acknowledgments

This work was supported by the European Commission through the Horizon Europe Programme (GAP-Noise project, with Grant Agreement No. 101073014)

References

[1] M. Munder and C.-C. Carbon, "A literature review [2000–2022] on vehicle acoustics: Investigations on perceptual parameters of interior soundscapes in electrified vehicles," *Frontiers in Mechanical Engineering*, vol. 8, p. 974 464, 2022.

[2] D. Lennstrom, T. Lindbom, and A. Nykanen, "Prominence of tones in electric vehicle interior noise," in *Proceedings of Inter-Noise 2013*, Innsbruck, Austria, 2013, pp. 508–515.

[3] F. Doleschal, H. Rottengruber, and J. L. Verhey, "Influence parameters on the perceived magnitude of tonal content of electric vehicle interior sounds," *Applied Acoustics*, vol. 181, p. 108 155, 2021.

[4] R. P. Hellman, "Perceived magnitude of two-tone-noise complexes: Loudness, annoyance, and noisiness," *Journal of the Acoustical Society of America*, vol. 77, no. 4, pp. 1497–1504, 1985.

[5] G. Pietila, W. Seldon, T. Roggenkamp, and T. Bohn, "Tonal annoyance metric development for automotive electric vehicles," SAE International, SAE Technical Paper 2019-01-1467, 2019.

[6] Ecma International, *ECMA-74: Measurement of airborne noise emitted by information technology and telecommunications equipment*, Ecma International, Geneva.

[7] H. Hansen, J. L. Verhey, and R. Weber, "The magnitude of tonal content. a review," *Acta Acustica united with Acustica*, vol. 97, no. 3, pp. 355–363, 2011.

[8] W. Aures, "Berechnungsverfahren fuer den sensorischen Wohlklang beliebiger Schallsignale," *Acustica*, vol. 59, no. 2, pp. 130–141, 1985.

[9] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *Journal of the Acoustical Society of America*, vol. 71, no. 3, pp. 679–688, 1982.

[10] R. Sottek, F. Kamp, and A. Fiebig, "A new hearing model approach to tonality," in *Proceedings of Inter-Noise 2013*, Innsbruck, Austria, 2013.

[11] R. Sottek, "Progress in calculating tonality of technical sounds," in *Proceedings of Inter-Noise 2014*, Melbourne, Australia, 2014, pp. 3319–3327.

[12] Ecma International, *ECMA-418-2: Psychoacoustic metrics for ITT equipment — part 2: Models based on human perception*, Ecma International, Geneva, Jun. 2025.

[13] M. J. B. Lotinga, M. Torjussen, and G. Felix Greco, "Verified implementations of the Sottek psychoacoustic Hearing Model standardised sound quality metrics (ECMA-418-2 loudness, tonality and roughness)," in *Proceedings of Forum Acusticum / Euronoise 2025*, Málaga, Spain, 2025.

[14] F. Doleschal and J. L. Verhey, "Tonality benchmark analysis for electric vehicle interior noise," in *Proceedings of Inter-Noise 2020*, ser. INCE Conference Proceedings, vol. 261, Seoul, Korea, 2020, pp. 5478–5489.

[15] D. Schecker, E. Deisling, and P. Uerlings, "Estimation of interior sound quality in BEVs regarding measurements on the NVH-Chassis-Dynamometer: Focusing on tonal components," SAE International, SAE Technical Paper 2020-01-1542, 2020.

[16] R. Sottek, J. Becker, and T. Lobato, "Tonal annoyance vs. tonal loudness and tonality," in *Proceedings of Inter-Noise 2019*, Madrid, Spain, 2019.

[17] T. Dupré, S. Denjean, M. Aramaki, and R. Kronland-Martinet, "Analysis by synthesis of engine sounds for the design of dynamic auditory feedback of electric vehicles," *Acta Acustica*, vol. 7, p. 36, 2023.

[18] R. San Millán-Castillo, E. Latorre-Iglesias, M. Glesser, S. Wanty, D. Jiménez-Caminero, and J. M. Álvarez-Jimeno, "MOSQITO: An open-source and free toolbox for sound quality metrics in the industry and education," in *Proceedings of Inter-Noise 2021*, ser. INCE Conference Proceedings, vol. 263, Washington, D.C., USA: Institute of Noise Control Engineering, 2021, pp. 1164–1175.

[19] J. Peirce et al., "Psychopy2: Experiments in behavior made easy," *Behavior research methods*, vol. 51, no. 1, pp. 195–203, 2019.