forum acusticum 2026
8-12 September · Graz, Austria

INDIVIDUAL ACOUSTIC PREFERENCE IN SELF-ASSEMBLED SLEEP SOUNDSCAPES: A POPULATION-SCALE FIELD STUDY

Stefan Dominic Schucker[1,2]

[1] *Empa, Swiss Federal Laboratories for Material Science and Technology, 8600 Dübendorf, Switzerland*
[2] *Sonora Audio Lab, 4500 Solothurn, Switzerland*

Abstract

Functional sleep soundscapes are a mainstream listening category, yet which acoustic properties keep users returning is unknown. We analyze twelve months of replay data from *Sonora: Sleep and Focus Sounds* (77,408 users), reconstructing each user's self-assembled sound mixes from event logs (app-initiated audio excluded) and computing every mix's spectrum from calibrated spectra of its constituent sounds. Classical psychoacoustic descriptors (ISO 532-1 loudness, sharpness, tonality) carry little predictive power: no association with return frequency survives control for usage intensity, and they separate day-one churn only weakly (AUC 0.59). In a mel-cepstral representation, however, the content of a user's first-day mixes alone (source categories plus spectral shape, no usage information) predicts whether that user ever returns (AUC 0.69). Within-person mixed models on 2,600 multi-session users show that the coupling between a session's acoustics and its listening time is strong but individually signed: between-user variation in this coupling is seven to nine times the population-mean effect, and is organized along two largely independent axes, level response and brightness response. Opposite individual signs cancel in any population average, which is why group-level analyses of sleep soundscapes find so little. There is no universally optimal sleep soundscape, and fixed-stimulus designs in sleep-noise research may underestimate effects for the same reason. A within-user randomized field study is planned.

Keywords: *psychoacoustics, sleep soundscapes, individual differences, mel-frequency cepstrum, field data*

***Corresponding author**: info@sonoraa.com.*

1. Introduction

Millions of consumers fall asleep to functional soundscapes: layered mixes of rain, wind, fire, and noise assembled inside mobile applications. Yet the acoustics of the soundscapes users actually return to are essentially uncharacterized. The standard psychoacoustic vocabulary (loudness, sharpness, tonality) was developed for short, often aversive product sounds [1]; whether it captures what governs prolonged, self-chosen ambient listening around sleep is an open question. Soundscape research has long held that the acoustic environment must be evaluated as perceived by a listener in context [2], and individual differences in reaction to sound are well documented [3], yet sleep-sound experiments typically administer one fixed stimulus to every participant [4].

This study asks whether the acoustic properties of self-assembled sleep mixes predict user retention, using twelve months of replay data from a deployed application. Three findings emerge. Standardized descriptors carry little retention signal at population level once usage intensity is controlled. In a mel-cepstral basis [5], however, the content of a user's first-day mixes predicts whether they ever return, and quieter, darker-tilted mixes predict higher return frequency among continuing users. Within-person models explain the discrepancy: the coupling between a session's acoustics and listening time is strong but individually signed, with a between-user spread seven to nine times its population mean, so that opposite preferences cancel in any population average.

2. Data and mix reconstruction

Source. Firebase Analytics events exported unsampled to BigQuery: 2025-07-29 to 2026-07-20, 77,408 pseudonymous users of *Sonoraa: Sleep and Focus Sounds*, 1.27M playback-start events over a catalog of ∼110 sounds. Start events carry the sound ID and the user-chosen volume (100% coverage); device timezone offsets localize events to the user's clock. A *session* is the analytics session (a new one begins after 30 min of

inactivity). Data are pseudonymous behavioral logs collected with user consent under the application's privacy policy; users who declined analytics are absent, so this population is smaller than the install base. No audio is recorded from users and no demographic attributes are collected; all sounds analyzed are catalog assets.

Mix-state reconstruction. Per user and session, a sweep-line over start/stop events maintains the active {sound → volume} set; each inter-event interval is a *mix-state* (volumes quantized to 0.1). Layering is the norm: only 32.5% of listening time is a single sound. A complete inventory of app-defined mixes (onboarding preview, auto-played defaults, 28 curated presets), reconstructed from the application's version history, allows app-initiated states to be excluded, so all features describe *user-assembled* listening. Durations are reconstructed from event timestamps, as client-reported values are unreliable; overnight playback partially outlives logging, so composition- and count-based features are favored throughout.

Cohorts. Each analysis uses the subset meeting its observability, outcome, and minimum-listening requirements, so cohort sizes are smaller than the 77k total. Two constraints apply throughout: every acoustic feature is computed from *user-assembled* mix-states only, and the population is sleep-anchored (73% of onboarding intent responses select sleep; 60% of users are bedtime-dominant, with ≥60% of their early sound starts between 21:00 and 04:00 local time; bedtime use is stratified or controlled in every model). Three cohorts are used. (i) *Return-frequency cohort*: users observable ≥30 days who reached day 3 (n=15,007); outcome: sessions in days 4–30; every model controls a first-72 h engagement baseline (log seconds, sessions, sound starts) in a negative-binomial GLM. Acoustic features use the first 24 h only, nested inside that baseline window; this is deliberately conservative, since part of the acoustic variance is absorbed by the control. (ii) *Day-one churn cohort*: one-shot users (lifespan <2 days, never seen again; n=4,106 with ≥60 s day-0–1 sessions) versus returners (≥1 active day in days 4–30, same session requirement; n=8,419). Here features span days 0–1 rather than 24 h, because one-shot users are defined by a lifespan of up to two days and a 24 h window would discard part of their only listening. (iii) *Within-person cohort*: users with ≥3 user-composed sessions of ≥60 s in their first 30 days (n=2,600; 13,225 sessions), for mixed-effects models of session-level listening.

3. Acoustic representation

For uncorrelated sources, power spectra add [6]. Each sound's power spectral density $S_i(f)$ is estimated once (Welch, full file, $\Delta f = 1.46$ Hz); any mix-state's spectrum is

$$S_{\mathrm{mix}}(f) = \sum_i v_i^2 \, S_i(f), \tag{1}$$

with v_i the user's volume for sound i, calibrated so the duration-weighted median mix sits at 68 dB SPL. Loudness computed from power-summed parts agrees with the rendered mix to 0.02%; frequency- and time-domain loudness computations agree to ≤0.2%.

From S_{mix} we derive two descriptions (Fig. 1). *Classical*: ISO 532-1 loudness [7], DIN 45692 sharpness [8], and ECMA tone-to-noise ratio via MOSQITO [9]. *Mel-cepstral*: a 40-band mel filterbank (20 Hz–16 kHz) applied to S_{mix} (mel energies are linear in power, so per-sound mel vectors compose exactly under Eq. (1)), followed by log and DCT, yielding coefficients c_1 through c_8 plus a mel-level term [5]. Here c_1 measures spectral tilt, with *higher* values indicating a darker mix (empirically, $r = -0.94$ with sharpness and +0.35 with low-frequency share). Where a session rather than a mix-state is the unit of analysis, its mel spectrum is the duration-weighted mean over the mix-states it contains.

4. Results

4.1. Classical metrics carry little predictive power

Across loudness, sharpness, tonality, SPL, and six spectral-shape descriptors, no metric survives Benjamini–Hochberg correction for predicting return frequency once the engagement baseline is controlled, either in the pooled cohort or within any stratum tested (bedtime users, sleep-onset vs. all-night use, preset vs. free composition). Uncontrolled associations (e.g. loudness, $\rho \approx 0.10$) reverse sign once app-initiated states are excluded and usage intensity is controlled. Two properties of the data limit these metrics further. Tone-to-noise ratio is degenerate at population scale: broadband layering masks tonal peaks for 98.5% of users, leaving no variance to correlate. Group-average spectra are equally uninformative: with thousands of users drawing from one finite sound library, group means converge to the library's usage-weighted average regardless of group composition. All analyses therefore use per-user distributions, controlled regressions, and out-of-fold prediction.

4.2. The mel-cepstral basis predicts

Day-one churn. Table 1 shows out-of-fold logistic AUC for classifying one-shot users against returners from the *content of their day-0–1 mixes alone*, with no usage information (user-level cross-validation). Source-*category* shares (rain, noise, fire, …) reach 0.611, pooled session MFCC statistics 0.656 (per user, the mean, standard deviation, minimum, and maximum across their sessions of the mel level and c_1–c_8; 36 features), and their combination 0.690: the two feature sets are complementary, since semantically similar sources can differ in spectrum and vice versa. These features are actionable: an application can influence what a first mix contains. The contribution is not attributable to usage intensity: adding intensity covariates (session count and durations; AUC 0.646 alone) raises the combined model only to 0.705, so mix content contributes +0.059 beyond intensity. Gradient boosting and shallow neural networks do not

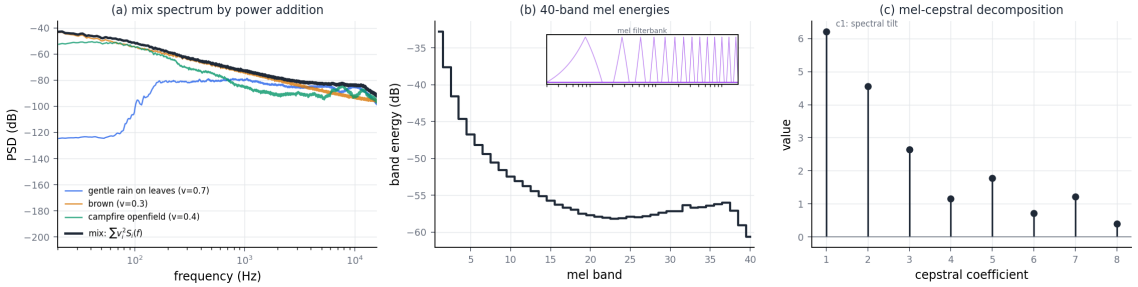

Figure 1: Acoustic pipeline on real catalog sounds. (a) A three-sound mix (rain 0.7, brown noise 0.3, campfire 0.4): per-sound PSDs and the power-added mix spectrum, Eq. (1). (b) The mix's 40-band mel energies (inset: filterbank). (c) Mel-cepstral decomposition; c_1 encodes spectral tilt, higher coefficients encode finer spectral curvature.

improve on the logistic model. The contrast has a consistent direction: one-shot users' mixes are darker and bassier, with per-user high-frequency share (>1.4 kHz) 0.7 dB lower in medians ($p = 3\times10^{-37}$). An AUC of 0.690 does not support individual-level classification, as the class distributions overlap substantially; for a predictor built from a single day of mix content and no behavior, however, it is a clear population-level signal, sufficient to rank new users by churn risk.

On the subcohort where the classical metrics are computable from first-24 h listening (n=11,696), loudness, sharpness, tonality, and SPL together reach AUC 0.587, against 0.655 for the MFCC features on those same users. The standardized descriptors thus carry genuine signal: as low-dimensional functionals of the same mix spectrum, they inherit part of its information. Adding them to the MFCC model, however, changes nothing (0.655 to 0.655), so what they capture is a subset of the mel-cepstral representation.

Table 1: Day-one churn prediction from mix content: one-shot users (n=4,106) vs. returners (n=8,419). Logistic regression, 5-fold user-level CV AUC. Categories = source-type shares (rain, noise, fire, ...); MFCC = pooled per-session mel-cepstral statistics. No usage features.

Mix-content features (day 0–1)	AUC
source categories	0.611
MFCC	0.656
categories + MFCC	**0.690**

Return frequency. The same representation also predicts return frequency among continuing users: in engagement-controlled negative-binomial GLMs (n=15,007), mel level (-0.051, $p_{\mathrm{BH}} = 2\times10^{-4}$), tilt c_1 ($+0.038$), and curvature terms c_4, c_5, c_8 (≈ -0.04) survive Benjamini–Hochberg correction: quieter, darker-tilted, spectrally smoother day-one mixes precede more frequent return. These effects are one third the size of the strongest behavioral predictor, early *settling*: the first principal component of five within-session concentration measures (dominant mix-state share, state entropy, and within-session variability of loudness, centroid, and mix composition), which indexes how quickly a user stops rearranging and stays with one mix ($+0.156$ per SD, $p = 3\times10^{-26}$, robust to country fixed effects and replicating in a temporal holdout).

Tilt therefore enters the two analyses with opposite signs: darker day-one mixes are more common among users who never return, yet predict *more* frequent return among those who do. The two cohorts ask different questions, one about selection into the continuing population and one about behavior within it, and both associations are small. A monotone reading (darker is better, or worse) is not supported; a plausible reconciliation is that very dark, bass-dominated first mixes deter some new users, while among users who stay, darker mixes accompany more regular use. This sign tension is one of the items in the pre-registered confirmation described below.

4.3. Preference is individual and two-dimensional

Figure 2 illustrates the scale of individual variation: day-one mix spectra range from deep-bass noise beds to bright rain-dominated mixes.

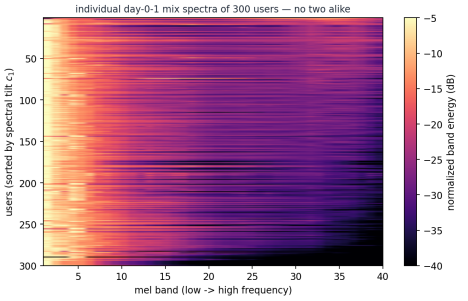

Figure 2: Mel spectra of 300 randomly chosen users' day-0–1 mixes (one row per user, sorted by spectral tilt c_1).

For the within-person cohort, linear mixed models relate log session listening time to the session's acoustic deviation from that user's own mean, with a random slope per user. Listening time is a tracked rather than an absolute quantity here (Section 2), but it enters only as a within-person contrast: each user is com-

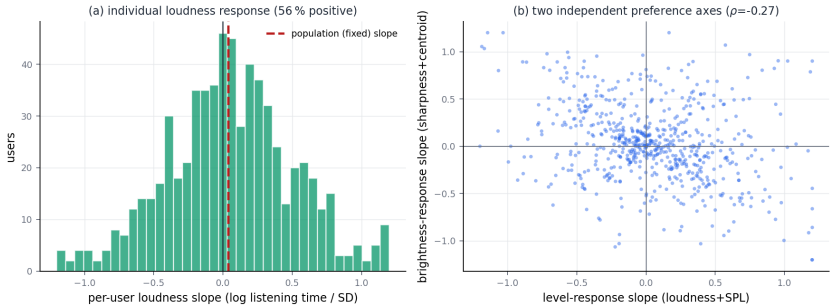

Figure 3: (a) Distribution of per-user loudness-response slopes (≥6 sessions): individually strong and of both signs, with a small population mean. (b) Level-response vs. brightness-response slope per user: the two preference axes are only weakly coupled, populating all four quadrants (17–31% each).

pared to their own sessions, and per-user truncation behavior cancels. Tracked sessions are short in any case (median 4.0 min; 99% below 1 h), and restricting the models to sessions under 2 h leaves every estimate in Table 2 unchanged to the third decimal.

On every dimension tested (loudness, SPL, sharpness, centroid, spectral flatness, low-frequency share, mel level, c_1, c_2), the population-mean slope is small while the between-user slope standard deviation is 0.20–0.27, seven to nine times larger (Table 2). The mean effects are not zero: most are significant at the session counts available, but they describe a weak average tendency that individual users depart from in both directions. For c_2 and flatness the mean effect is indistinguishable from zero while the between-user spread is undiminished. Roughly half of users listen longer when their mix is louder (or brighter) than their own norm; the other half, shorter. Adding a bedtime indicator to every model changes the slope standard deviations by at most 0.0004, so the effect is not a daytime-versus-night usage artifact.

Table 2: Random-slopes mixed models (≥3 sessions; ≈2,600 users, 13,225 sessions): log session listening time vs. within-person acoustic deviation (per SD). Fixed effect (per-SD, with standard error), SD = between-user standard deviation of the same slope, p = boundary-corrected likelihood-ratio test of slope variance.

Metric	fixed (SE)	SD	p
Mix SPL	+0.031 (0.012)	0.271	2×10^{-20}
Mel level	+0.034 (0.012)	0.256	8×10^{-17}
Loudness	+0.038 (0.012)	0.255	1×10^{-16}
MFCC c_2	+0.009 (0.012)	0.238	4×10^{-13}
Centroid	+0.029 (0.011)	0.230	3×10^{-12}
LF share	−0.026 (0.011)	0.224	1×10^{-10}
MFCC c_1	−0.021 (0.011)	0.208	2×10^{-10}
Flatness	−0.008 (0.011)	0.199	2×10^{-8}
Sharpness	+0.023 (0.011)	0.198	6×10^{-9}

The per-user slopes have low-dimensional structure (Fig. 3): loudness- and SPL-slopes correlate at 0.75 (a *level-response* axis), sharpness-, centroid-, and flatness-slopes at 0.63–0.75 (a *brightness-response* axis), while the two axes are only weakly coupled (loudness- vs. sharpness-slope −0.08; composite axes $\rho = -0.27$). Individual preference over sleep soundscapes is thus approximately two-dimensional: how a person responds to level and, largely independently, how they respond to brightness, with users populating all four preference quadrants. Users also converge over their first month toward a personal spectral anchor (consecutive-session spectral distance declines with session index; median per-user rank correlation −0.10), consistent with a search-then-settle process. What a user assembles predicts retention (Section 4.2); how the user responds to it, measured by slope strength or settling speed, does not predict retention further once engagement is controlled. Whether a mix that matches a user's preference causes them to return remains an experimental question.

5. Discussion

Acoustic response to sleep soundscapes is real but individually signed and two-dimensional, so population averages cancel it; no universal acoustic prescription should be expected to exist. What remains visible at population level is a day-one churn contrast (users who never return had assembled darker, bassier mixes) and a behavioral signature of fit-finding (early settling).

The result bears on sleep research beyond apps. Experimental studies of noise as a sleep aid typically administer a fixed, generic stimulus, most often broadband noise at a set level, and report mixed group-level effects [4], [10], [11]. Our data suggest a mechanism: if acoustic response is individually signed, a one-size-fits-all stimulus averages opposing responses toward zero. Individually titrated stimuli and within-subject designs may be prerequisites for measuring the true effect of sound on sleep.

Methodologically, the study provides a validated route to population-scale mix acoustics (power addition; exact composition in the mel basis) and identifies three requirements for acoustic analysis of replay data: app-initiated audio must be separated from user-initiated audio; usage-intensity confounding must be controlled, as it can reverse the sign of acoustic associations; and group-average spectra must be avoided, as they carry

no group information when users draw from a shared sound library.
Limitations: all spectra describe the signal delivered to the device, not the sound at the ear. Device volume, transducer (phone speaker, earbuds, bedside speaker), and room are unobserved, and the calibration constant is chosen so that the duration-weighted median mix sits at 68 dB SPL, a plausible bedside level but an assumption; absolute loudness and sharpness values are therefore level-dependent quantities on an arbitrary reference, and only relative comparisons between mixes are used. Further: observational design; listening time measures engagement, not sleep quality; within-person evidence covers multi-session users (one-shot users cannot contribute repeated observations by definition); mel-cepstral and churn findings are exploratory, with a pre-registered confirmation on later cohorts scheduled. The decisive test, a within-user field experiment randomizing suggested mixes near vs. far from the inferred personal preference, is planned.

6. Conclusion

Classical psychoacoustic metrics carry little retention signal, but acoustics are not irrelevant: mel-cepstral features of a user's first-day mixes predict day-one survival and, more weakly, return frequency, and within-person models establish that acoustic preference is individual, bidirectional, and two-dimensional (level × brightness). This structure also bears on sleep-sound studies, which typically administer one fixed stimulus to all participants: when responses are individually signed, such designs yield near-zero group effects even where individual effects are strong. For sleep-audio design, these results argue for helping each user find and settle on their own mix rather than optimizing a single soundscape for all listeners.

References

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

[2] *ISO 12913-1:2014 acoustics — soundscape — part 1: Definition and conceptual framework*, Geneva, Switzerland: International Organization for Standardization, 2014.

[3] N. D. Weinstein, "Individual differences in reactions to noise: A longitudinal study in a college dormitory," *Journal of Applied Psychology*, vol. 63, no. 4, pp. 458–466, 1978.

[4] S. M. Riedy, M. G. Smith, S. Rocha, and M. Basner, "Noise as a sleep aid: A systematic review," *Sleep Medicine Reviews*, vol. 55, p. 101 385, 2021.

[5] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[6] J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures*, 4th. Hoboken, NJ, USA: Wiley, 2010.

[7] *ISO 532-1:2017 acoustics — methods for calculating loudness — part 1: Zwicker method*, Geneva, Switzerland: International Organization for Standardization, 2017.

[8] *DIN 45692:2009 measurement technique for the simulation of the auditory sensation of sharpness*, Berlin, Germany: Deutsches Institut für Normung, 2009.

[9] R. San Millán-Castillo, E. Latorre-Iglesias, M. Glesser, S. Wanty, O. Jacqmot, and J. Zuponcic, "MOSQITO: An open-source and free toolbox for sound quality metrics in the industry and education," in *Proc. INTER-NOISE 2021*, Washington, DC, USA, 2021.

[10] L. Messineo, L. Taranto-Montemurro, S. A. Sands, M. D. Oliveira Marques, A. Azabarzin, and D. A. Wellman, "Broadband sound administration improves sleep onset latency in healthy subjects in a model of transient insomnia," *Frontiers in Neurology*, vol. 8, p. 718, 2017.

[11] M. R. Ebben, P. Yan, and A. C. Krieger, "The effects of white noise on sleep and duration in individuals living in a high noise environment in New York City," *Sleep Medicine*, vol. 83, pp. 256–259, 2021.