forum acusticum 2026
8-12 September · Graz, Austria

EXPECTED REPRODUCTION LEVELS OF SOUNDSCAPES

Ching Jo Hsu[1], Nils Meyer-Kahlen[2,3], Markus von Berg[4], Tapio Lokki[2]

[1] *Dpt. of Arts and Media, Aalto University, Finland*
[2] *Acoustics Lab, Dpt. of Information and Communications Engineering, Aalto University, Finland*
[3] *Dyson School of Design Engineering, Imperial College London, United Kingdom*
[4] *Institute of Sound and Vibration Engineering, Hochschule Düsseldorf, Germany*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

Humans appear to possess internal expectations of reproduced sound pressure levels of auditory scenes, but they might not be very accurate. Previous studies have shown a systematic underestimation of reproduction levels for speech and traffic soundscapes. This study investigates the factors underlying such underestimation by examining level expectations across six real-world soundscapes reproduced using a multichannel loudspeaker array. Participants adjusted playback levels to match their expectations of the soundscapes' real-world counterparts, both with and without added recording noise. In addition, they rated perceptual attributes in accordance with ISO/TS 12913-3. Results show a strong dependence of level mismatch on the true sound pressure level of a soundscape: levels of louder scenes were consistently underestimated, while quieter scenes were slightly overestimated, indicating a compression effect in level expectations. The true level of the soundscape was a better predictor of estimation error than perceived pleasantness. Added recording noise had only a minor influence, primarily affecting quieter scenes, while familiarity with the soundscape showed no credible effect on accuracy. These findings suggest that level expectations may be anchored to specific sound sources and biased toward perceptually comfortable levels.

Keywords: *Soundscape, Loudness Perception, Psychoacoustics, Spatial sound reproduction*

1. Introduction

It is likely that humans possess a certain internal reference for the sound pressure level of a given acoustic scene. However, these level expectations might not be very accurate. In lab-based acoustical experiments, for example, it is common that reproduction levels are not set to the corresponding real acoustic situation but to what researchers deem comfortable. Such presentations usually do not strike participants as overly unnatural or wrong [1], [2], [3]. Recently, an experiment has shown that subjects tend to underestimate the level of a soundscape during reproduction by 9 dB, even if they had been exposed to the real-world scene minutes before the experiment [4]. Prior to this, similar results have been obtained when investigating expected levels for speech, which were underestimated by 3.6 dB, on average [5].

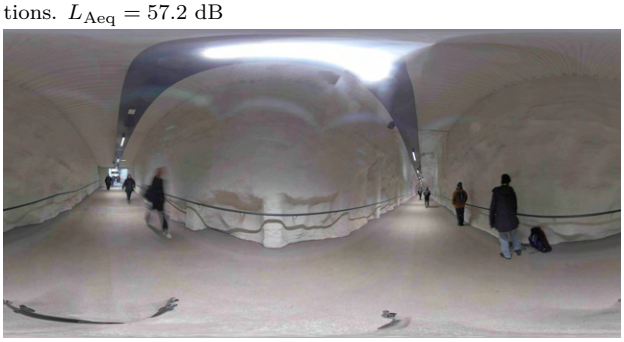
The reasons for this effect are unknown at this time, and several hypotheses can be formed. To explain their results, previous studies have suggested that biases might be caused by quieter background levels in laboratories [6], which would interfere with the perceptual scaling of an 'appropriate' reproduction level [4], [7]. In this work, we address five research questions, aiming to test several other possible explanations for these observations. We first want to assess whether there are differences in estimates of reproduction levels across the specific soundscapes reproduced (H1). Given that this is the case, we test three hypotheses to explain them. First, we test whether louder (H2.1) or more unpleasant (H2.2) soundscapes exhibit a stronger negative bias. This might be expected from larger differences to the quiet laboratory and could also point to a form of avoidance behavior, in which listeners simply avoid high levels [4]. We then examine whether added recording noise influences judgments, whereby reproductions from noisier equipment may be attenuated more to suppress perceptually unnatural scene-independent noise (H3). Finally, we assess whether absolute errors are associated with individual familiarity of a given scene—a robust internal representation of a familiar site's level might lead to more familiar scenes being judged more accurately (H4).

To address these hypotheses, we test level expectations for six soundscapes reproduced using Ambisonics in a surrounding loudspeaker array. We let participants

***Corresponding author**: nils.meyer-kahlen@aalto.fi.*

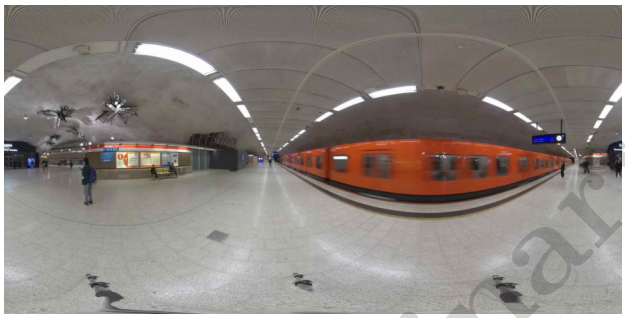
(a) Kaisaniemi Underpass Tunnel – under the central railway station, with bicycles, pedestrians, and cars going in both directions. $L_{\mathrm{Aeq}} = 57.2$ dB

(b) Oodi Library – spacious third floor of the central library in Helsinki; the microphone was facing a cafe, with an event area on the left and a children's area on the right. $L_{\mathrm{Aeq}} = 59.4$ dB



(c) Opintoputki Metro Entrance – a rocky tunnel leading to the entrance of University of Helsinki metro station, with a street performance inside the tunnel. $L_{\mathrm{Aeq}} = 67.4$ dB

(d) Otaniemi Shore – on the campus of Aalto University beside the road and nearby seaside during sunset, with birds chirping from the surrounding forest. $L_{\mathrm{Aeq}} = 71.6$ dB



(e) Kamppi Metro Platform – one meter away from the side of the train door during rush hour. $L_{\mathrm{Aeq}} = 77.7$ dB

(f) Central Railway Station – reverberant hall outside the metro entrance and exit to the railway station, with large movement of people and synthetic bird sounds. $L_{\mathrm{Aeq}} = 81.1$ dB

Figure 1: Measurement locations used in the public sounds experiment, A-weighted equivalent sound pressure levels during reproduction.

adjust the reproduction level to what they believe is the real-world level, both with and without adding noise. In addition, we let them rate the eight attributes defined in ISO/TS 12913-3:2019 [8] to obtain a pleasantness score. To gauge their familiarity with the scene, we also ask them to describe where the soundscape may have been recorded.

2. Methods

2.1. Recordings

The recording setup consisted of one Zylia ZM-1 3rd-order Ambisonics microphone and one omnidirectional Behringer ECM8000 measurement microphone, connected to a Macbook Pro via Zoom H4n as an audio interface, all mounted on the same microphone stand. Both microphones were recorded simultaneously in Reaper. Before each recording, a 15-second 1-kHz calibration tone at 94 dB from a Brüel & Kjær Type 4231 sound level calibrator was recorded by the measurement microphone. Windshields were used in outdoor environments. Images of the scenes were taken by an Insta360 Pro 2 Spherical 360° camera[1].

All recordings were made at a sampling rate of 48 kHz with a 24-bit-per-sample resolution. Recording notes included date, gains of the two microphones, and a short description of the scene. Microphone gains were reduced for windy environments and for the railway station recordings to prevent clipping.

Six scenes of indoor and outdoor public spaces in Helsinki and Espoo, Finland, were selected for the lis-

[1]The pictures were not used in the present experiment.

tening test to balance familiarity and diversity of the scenes. One minute of each recording was selected for the experiment. The 360° pictures and short descriptions of the selected scenes can be seen in Fig. 1, along with the equivalent continuous A-weighted sound pressure levels (L_{Aeq}). The true levels ranged from 57.2 dB to 81.1 dB.

2.2. Experimental Setup

Laboratory calibration and the final experiment were conducted in the multichannel anechoic chamber "Wilska" at the Aalto Acoustics Lab, equipped with 47 Genelec 8331AP loudspeakers surrounding the listener, of which 37 were used for playback. The subset of 37 speakers was selected so that it roughly approximates a uniform directional distribution on the sphere. An advantage of reproducing soundscapes using a multichannel loudspeaker array in an anechoic chamber compared to headphone reproduction is that participants listen with their own ears, avoiding issues of non-individual binaural reproduction or head-tracking latency.

The six one-minute recordings were first encoded to Ambisonics using SPARTA array2sh, then decoded to a 37-channel loudspeaker layout using SPARTA ambiDEC [9] hosted in Reaper, and exported for playback in Pure Data. When comparing long-term power spectral density estimates between recording and reproduction, a certain loss of high frequencies was observed. It was compensated using a high-shelf filter with a Q-factor of 0.7, a gain of 7.6 dB, and a cut-off frequency of 2700 Hz.

For the noisy version, 19-channel pink noise at −18 dB was added to simulate Zylia microphone self-noise of the 19 capsules prior to Ambisonics decoding. The level was chosen such that it was barely audible to the authors, as we did not want participants to perceive it as an obvious noise artifact.

Prior to the listening test, the selected one-minute real-world recordings were reproduced in Wilska and re-recorded using the same omnidirectional measurement microphone used along with the Zylia in the field. Also, a calibration tone was recorded using the same B&K calibrator. With that, it became possible to compare the level of reproduction at the omnidirectional microphone with the level during the recording. A gain factor was determined based on the A-weighted difference and applied to the decoded Ambisonics signal. Like this, when a gain of 0 dB was set in the experiment, the level should correspond to the real-world level. The calculations for the required level adjustment were conducted in MATLAB.

2.3. Experimental Procedure

Each participant completed two blocks, with a short break in between. The participants were identified by a number; odd numbers started with the no-noise condition, and even numbers started with added pink noise. Participants were instructed to adjust the playback level of each recording to match their expectation of the sound level in the real-world environment where

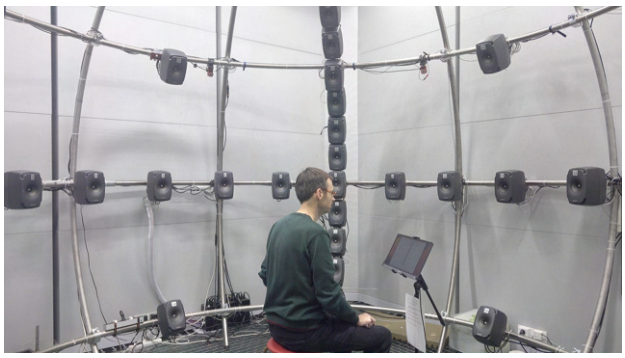


Figure 2: During the experiment inside the multichannel anechoic chamber "Wilska" at Aalto Acoustics Lab.

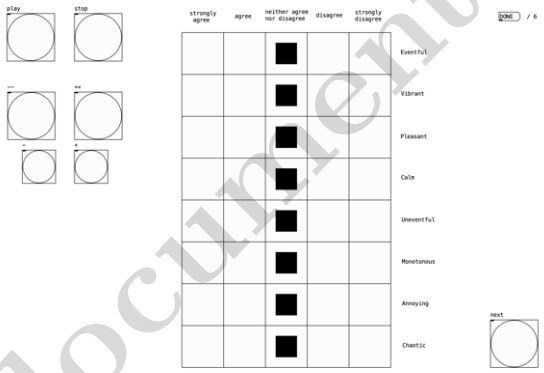


Figure 3: Graphical User Interface in Pure Data.

the recording was made. Then, the participants rated the eight attributes (eventful, vibrant, pleasant, calm, uneventful, monotonous, annoying, chaotic) defined in ISO/TS 12913-3:2019 [8] on a six-point scale from "strongly agree" to "strongly disagree", and provided a guess of the recording location. They wrote their guess of the location as free text on a piece of paper. The location answer was later used as an indicator of scene familiarity. In both blocks, participants followed the same procedure, except that the location was written down only in the first block. After both blocks, participants filled out a short post-experiment survey.

The participants remained seated throughout the test in the center of the multichannel anechoic chamber "Wilska", with a touchscreen in front of them. A camera and microphone were placed inside the room for communication between the participant and the experimenter. The experiment, with its graphical user interface (GUI), was implemented in Pure Data, see Fig. 3. The GUI provided controls for playback ("play", "stop", "next"), rating scales for the soundscape attributes, and buttons to adjust playback levels in steps of 1 dB and finer adjustments of 0.5 dB. Each scene was one minute long and could be replayed. Starting playback levels of each sample were randomized within a range of ±3 dB, and participants were informed that level differences of up to 6 dB could occur between recordings. The order of the six sound

scenes was randomized within each block. The average duration of the listening test was 30 minutes, including the training session and a short break between the blocks.

2.4. Participants

In total, 24 participants (71 percent male, 25 percent female, 4 percent diverse) volunteered for the listening experiment. The age ranged from 25 to 54 years, with a median age of 32.8 years (SD = 8.9 years). Participants included professors, PhD candidates, staff members, and students of the Acoustics Lab and the Department of Art and Media at Aalto University, as well as non-university members. Eleven of the participants had formal training in acoustics, 16 had some and extensive experience in sound design, sound art, or related artistic work. Ten participants had professional musical training. Four participants reported having no training in fields related to acoustics, music, or sound. One participant reported a mild hearing impairment with a few dB loss in the left ear, and two noted potential hearing loss.

2.5. Bayesian Data Analysis

First, the responses to the soundscape attributes were processed following the standardized procedure [8], so that a score for *eventfulness* and *pleasantness* was obtained by forming weighted combinations of the eight attribute ratings.

The free text location answers were encoded into a three-level scale (1–3) by the first author, where 1 corresponds to no answer or a completely incorrect location, 2 stands for a partly correct, but vague location (e.g. “a cafe”, “restaurant”, “outdoors”), and 3 denotes answers that include the exact name of the place (e.g. “University of Helsinki metro entrance pathway”, “Kaisaniemi Underpass”, “Central Railway Station hallway”).

To address our research questions, Bayesian linear mixed-effects models were fitted using the *brms* package in *R* [10]. For models used to address H1–H3, the outcome variable was the signed error between the real level and the level to which participants set the reproduction. All models included the participant as a random effect. As an initial step to check H1, a baseline model was specified with soundscape as a categorical fixed effect. Then, different predictors were tested to assess the remaining questions. First, the true level was used as the only predictor to assess H2.1. This model was compared with a model that used the mean pleasantness instead (H2.2). Next, the simulated recording noise was added as a factor to the categorical model to assess H3. As we assume that familiarity influences absolute error rather than signed error (H4), a separate model was fit using the categorical 3-level familiarity scores as a predictor of the absolute error.

The results are reported in terms of mean posterior coefficient estimates and their credibility intervals (CrIs), defined through the 95% highest posterior density interval. Comparisons between models were conducted using leave-one-out (LOO) cross-validation [11] and computation of the expected log pointwise predictive density (ELPD) difference. A large ELPD difference indicates that one model better predicts the data than another model.

In all models, a weakly informative normal prior ($\mu = 0$, $\sigma = 5$) was specified for the fixed effect coefficients. Default Student-t priors were used for the intercept and variance parameters. Posterior inference was based on Markov Chain Monte Carlo sampling using four chains with 2000 samples each, with the first 1000 samples discarded as warm-up. $\hat{R} \leq 1.01$ was verified for all models, indicating that the models converged and mixing was achieved.

3. Results

Figure 4 shows the results of the soundscape analysis, showing eventfulness and pleasantness scores for each participant and soundscape. The means provide a first overview of how the soundscapes were perceived. The recording from Otaniemi Shore was clearly perceived as the most pleasant, followed by the Bike Tunnel and the Metro Entrance. In terms of eventfulness, the railway station scored the highest, followed by the library and the metro platform. Overall, these results are in line with our expectations.

Figure 5 shows the distributions of adjusted levels for each scene. It is immediately apparent that there are big differences in the set gain between scenes. When comparing a categorical model using scene as a predictor with a null model that only uses the participant as a random effect and performing LOO, the expected log predictive density difference is ELPD = 150, providing very strong evidence that the model including the soundscape as a predictor better explains the data. This clearly confirms H1, showing that differences in estimation error exist between scenes. Specifically, the strongest underestimation was found for the Railway Station scene (median of −14.3 dB in the case of no added noise), followed by the Metro Platform and the Otaniemi Shore (medians of −5 dB and −4.5 dB, respectively). In the case of the Bike Tunnel, Oodi Library, and the Metro Entrance, the level was even overestimated on average (with medians of +3.0 dB, +2.3 dB, +2.0 dB).

The soundscapes in Fig. 5 were ordered by ascending sound pressure level. Hence, it is easy to see that the levels of louder soundscapes were underestimated to a greater extent. Consequently, the model using the soundscape level as a predictor accounts for this effect very well. The estimate of the level coefficient was −0.59 with a CrI of $[-0.66, -0.52]$, meaning that for every decibel of increase in absolute sound pressure level of the scene, the level set by the participants decreases by 0.59 dB. The mean of the estimated intercept was 37.22 (CrI: $[32.40, 42.06]$). Hence, for the example of Otaniemi Shore, with an absolute level of 71.6 dB, the predicted underestimation is $37.22 - 0.59 \cdot 71.6 = -5.02$ dB, which is very close to the observed median value of −5.0 dB. The CrI of the

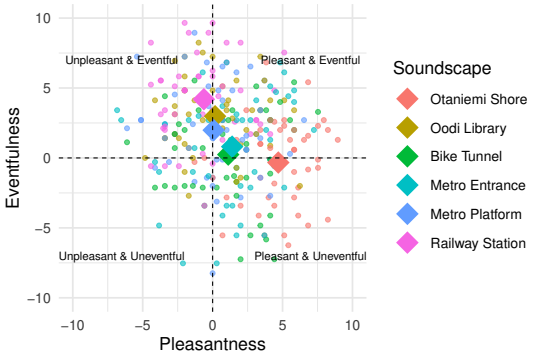


Figure 4: Soundscape profiling according to ISO/TS 12913-3:2019 for the six tested scenes.

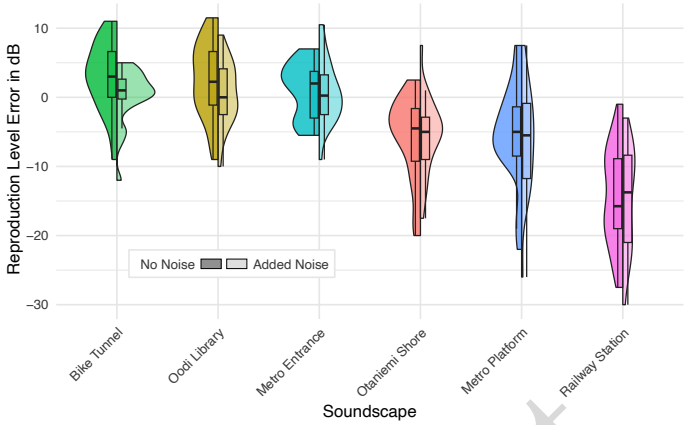


Figure 5: Difference between true level and level set by the participants. A set gain of 0 dB corresponds to the true level.

slope parameter clearly excludes zero, providing strong evidence for the presence of this effect, confirming H2.1.

To test H2.2, the scene-wise mean pleasantness score was used as a predictor instead of the total level. It yielded a coefficient of 0.54 with a CrI of $[0.05, 1.04]$. The ELPD difference is 108, clearly indicating that the true sound pressure level is a better predictor of set level than perceived pleasantness.

The results obtained when additional simulated microphone noise was added to the recordings are shown on the right side of the violins in Fig. 5. It appears that the differences from the noise-free condition are small. When fitting a model using soundscape as one predictor and noise as a second one, the contrast between the coefficients for no noise and added noise was 0.98 (CrI: $[0.03, 3.02]$). Yet, there is an interaction with the specific soundscape: In line with what can be seen in Fig. 5, the levels of the quieter scenes are overestimated less when the additional simulated recording noise was added. The contrast between noise and no noise for the Bike Tunnel is 2.31 (CrI: $[-0.08, 4.59]$), and Oodi Library is 1.68 (CrI: $[-0.7042, 3.98]$). For the Metro Entrance there is a contrast of only 0.15, and the CrI very clearly includes zero (CrI: $[-2.18, 2.49]$), making the estimate unreliable. The same is true for the remaining soundscapes.

Finally, we tested H4, i.e., whether familiarity with a given scene leads to lower absolute errors. The model using familiarity as a predictor to explain absolute set level error did not yield any real explanatory power. The strongest difference in familiarity (level 1 vs. 3) resulted in a contrast of -0.68 (CrI: $[-2.4, 0.99]$).

4. Discussion

4.1. Mentioned Strategies

When asked after the tests, many participants reported using specific sound sources as reference points for adjusting playback levels, such as the departing metro or the perceived distance of the platform to the microphone. However, sound sources in public environments are often dynamic rather than fixed. Some environments, such as metro stations, can have a variation in sound level depending on time of day and passenger density. Consequently, results are tied to events within the specific recordings rather than the places as such.

Two participants mentioned that the low frequencies were louder than real-life experiences, either due to the microphone or because they are subconsciously ignored in the real-world. Additionally, the lack of vibration is a limitation when evaluating experienced sound levels, particularly in the metro scene. One participant further noted that everyday use of noise-canceling headphones may affect loudness expectations by reducing exposure to real-world reference sound. A relationship between level expectation judgment and the use of noise-canceling headphones could be a topic for future research.

4.2. Perceived Differences Between First and Second Block

When asked after the experiment, 79.2% of participants reported feeling different about the scenes between the two presentations. Several participants reported that in the first block, mental energy was directed towards identifying the recording location, whereas during the second block, attention was focused on playback level adjustment. However, based on the numerical values, the difference in mean error across all scenes and noise conditions between the first and second runs was only 0.21 dB (0.1 dB when looking at the absolute error).

4.3. Causes of Estimation Error

Summarizing the results, we found a strong dependence of reproduction level expectation on the true level of the scene. However, even though the difference in true A-weighted equivalent continuous level was approximately 24 dB, the mean differences in set level were only -16 dB. A similar trend was found in [5], where loud speech samples were underestimated more severely than quieter ones. The mean estimates of

quieter scenes were even above the true level, which had not been observed in previous studies. Thus, there seems to be a compression effect rather than a complete lack of reference or consistent underestimation. Underestimation of loud scenes may be due to the fact that expectations are generally imprecise. Given a wide range of acceptable values, participants may have preferred more comfortable listening levels. Alternatively, participants might use the current background noise as a reference for scaling [4]. Hence, the level of louder scenes that differ most from the quiet laboratory is underestimated more than that of quieter scenes, whose true level is closer to the current laboratory surroundings. Future studies should try to disentangle simple avoidance of loud sounds from relative scaling by including conditions with background noise in the reproduction space.

Moreover, in the present experiment, participants were only provided with auditory information, which reduces potential influences from visual context. Seeing the scene would make understanding what happens in it much easier. However, at least in our current results, scene understanding, as indicated by recognition of the place, did not seem to influence outcomes.

Adding simulated recording noise led to a trend of setting lower reproduction levels for the quieter soundscapes, which were less overestimated when noise was added. Given that and the large errors for the loud scenes, it is unlikely that equipment noise is a main reason for underestimation in general. Note that the noise added here was subtle. One participant remarked that their second block had added noise, even though, for this participant, the noise was added in the first block. Another participant noticed noise in general, but no difference between the two blocks. The effect of recording noise could be studied using more extreme noise levels. However, a between-subject design might be required to do so. A future study should also account for participants' experience level by testing a larger, non-expert population.

5. Conclusion

When asking people to adjust the playback level of soundscapes to what they believe is the correct real-world level, the results show underestimation of loud scenes. Quiet scenes were slightly overestimated in level. The true A-weighted equivalent continuous sound pressure level was a better predictor of estimation errors than the mean perceived pleasantness of each scene.

Furthermore, the results show a trend for the presence of added noise affecting participants' playback level adjustments when the sound pressure level of the scene is quiet. When background noise was present, participants selected lower playback levels compared to the no noise condition. Familiarity with the scene, on the other hand, did not show a clear relationship in level expectations.

References

[1] A. S. Sudarsono, Y. W. Lam, and W. J. Davies, "The effect of sound level on perception of reproduced soundscapes," *Applied Acoustics*, vol. 110, pp. 53–60, 2016. DOI: 10.1016/j.apacoust.2016.03.011.

[2] Y. Lu and S.-K. Lau, "Examining the ecological validity of vr experiments in soundscape and landscape research," *Computers in Human Behavior*, vol. 162, 2025. DOI: 10.1016/j.chb.2024.108462.

[3] T. Yang and J. Kang, "Perception difference for approaching and receding sound sources of a listener in motion in architectural sequential spaces," *The Journal of the Acoustical Society of America*, vol. 151, no. 2, pp. 685–698, Feb. 2022. DOI: 10.1121/10.0009231.

[4] M. Von Berg, S. Versümer, J. Bitta, and J. Steffens, "Reduced reproduction levels of outdoor soundscapes are deemed appropriate – even after real-world exposure," en, *Acta Acust.*, vol. 9, p. 78, 2025. DOI: 10.1051/aacus/2025062.

[5] N. Meyer-Kahlen, "Expected Levels of Reproduced Speech," in *AES Conference on Augmented and Virtual Reality*, Redmond, WA, USA, Aug. 2024.

[6] W. J. Davies, N. S. Bruce, and J. E. Murphy, "Soundscape reproduction and synthesis," *Acta Acustica*, vol. 100, no. 2, pp. 285–292, 2014. DOI: 10.3813/AAA.918708.

[7] L. D. Braida, J. S. Lim, J. E. Berliner, N. I. Durlach, W. M. Rabinowitz, and S. R. Purks, "Intensity perception. xiii. perceptual anchor model of context-coding," *The Journal of the Acoustical Society of America*, vol. 76, no. 3, pp. 722–731, Sep. 1984. DOI: 10.1121/1.391258.

[8] "Acoustics — soundscape — part 3: Data analysis," International Organization for Standardization, Geneva, Switzerland, Technical Specification ISO/TS 12913-3:2019, 2019.

[9] L. McCormack and A. Politis, "SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods," in *AES International Conference on Immersive and Interactive Audio*, York, UK, Mar. 2019.

[10] P.-C. Bürkner, "**brms** : An *R* Package for Bayesian Multilevel Models Using *Stan*," en, *Journal of Statistical Software*, vol. 80, no. 1, 2017. DOI: 10.18637/jss.v080.i01.

[11] A. Vehtari, A. Gelman, and J. Gabry, "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC," en, *Statistics and Computing*, vol. 27, no. 5, pp. 1413–1432, Sep. 2017. DOI: 10.1007/s11222-016-9696-4.