forum acusticum 2026
8-12 September · Graz, Austria

# RAPID BACKGROUND SOUND GENERATOR FOR URBAN SOUNDSCAPES STUDIES IN STUTTGART

Hou Hin Au-Yeung[1,2], Daniel Wolfgang Kling[1,3], Josep Llorca-Bofí[1,4]

[1] *Urban Architectural Acoustics Group, Fraunhofer Institute of Building Physics, Germany*
[2] *Institute of Acoustics and Building Physics, University of Stuttgart, Germany*
[3] *Stuttgart Media University, Germany*
[4] *Department of Architectural Representation, Polytechnic University of Catalonia, Spain*



## Abstract

European cities are adapting soundscape-based urban planning to create better living spaces for residents [1]. Auralization acts as one of the possible tools that provides immersive auditory experiences for urban planners, architects and the community to participate in the design process without acoustical expert [2]. Existing auralization tools is computationally expensive [3]. This paper explores using surrogate background sound generators. By combining Text-to-Audio models (TTA) and Binaural Room Impulse Response-based models, the paper highlights the potential of surrogate audio renderers to support more perceptually grounded, design-oriented decision-making in urban environments.

**Keywords:** *auralization, urban planning, synthetic soundscape, soundscape composition*

## 1. Introduction

The word *Soundscape* stems from the early sonography works [4], [5], where the authors define soundscape as "the acoustic (somewhat) equivalent to the landscape" which contains heard events. Soundscape acted as an instrument for preservation of Acoustic Ecology and History. Modern acousticians have adapted the concept of soundscape and further elaborated the definition to "an acoustic environment as perceived or experienced and/or understood by a person or people, in context" [6] and "a perceptual construct based on physical places with the sound sources of the acoustic environment heard in real time" [7], which shifts toward human perception and built environment. Soundscape is increasingly used as a carrier and assessment criteria for urban planning.

***Corresponding author**: hou.hin.au-yeung@ibp.fraunhofer.de.*


There were researches on synthesizing soundscapes, with/without the use of generative AI. For instance, soundscape can be composed using machine-generated audios [8], generated based on textual spatial contexts [9], or based on geo-spatial and weather context [10], or relying on GPT models to synthesize audio parameters for downstream granular synths from contextual inputs [11]. These researches largely matches the framework given in [12]:

1. Specify an environmental context
2. Sound file retrieval
3. Synthesize / Extract representative audios from recordings
4. Mix and sequence

Despite the attempts in synthesize perceptually realistic soundscape, there are only a few works that model physical effects of sound propagation. Work in [13], for instance, attenuates audio samples with respect to the synthesized source distance and sound propagation effects. The effect of diffraction and reflection from the surrounding geometry is not explicitly considered. On the other side of spectrum, some attempted to generic synthetic sounds. State-of-the-art Text-To-Audio (TTA) models is capable of generating realistic sound effects, usually under 5 minute. More generic models such as [14], [15] is capable of generating multiple sequential, realistic sound effects. [16], [17] provides spatialization support, but lacks the interface of inputting the geometrical and environmental context. The common issue with these TTA model is, the inference time is heavily reliant on the GPU hardware, and would take more than 2-3 minutes in generating single clip,which hence failing the purpose of rapid soundscape generation.
In this paper, we aim to fill the gap between previous synthetic soundscape works, and state-of-the-art,

full-scale heavy-duty acoustic solvers. We propose a framework for synthesizing background sounds, which composes sounds based on geo-spatial contexts and physically-based sound propagation effects, via surrogate Binaural Room-Impulse-Response (BRIR) models. This methodology would allow urban planners and architects to rapidly generate relevant sound for efficient soundscape evaluation.

## 2. Methodology

A brief breakdown schematic diagram is shown at Fig. 1. The framework largely follows the layout from [12], namely *context specification, sound file retrieval, and mixing*. We intentionally skipped the *sound synthesis* step, as we are using the sound clips from the datasets as-is. Instead, we propose to add the *spatial mixing* stage that could increase the immersive-ness of generated background soundscape. The stages will be each elaborated in details in the following.

### 2.1. Environmental context

A Point of Interest (POI) is a term used in cartography to represent a particular feature in a given area [18]. This is helpful in associating geo-spatial context with different sound objects. A Rhino plugin has been developed that allows user to select any area on a map. The POIs, for instance, post office and restaurants, are subsequently extracted from the downloaded OpenStreetMap (OSM) data.

### 2.2. Sound file retrieval

#### *2.2.1. Urban Sound Datasets*

While multiple previous works ([12], [19]) source sound samples from *Freesound* [20], we adapt the audio samples from *UrbanSound8K* [21] which is a labeled, 4-second datasets stemming from Freesound and *SONYC* [22] (which contains recordings of soundscapes in New York cities classified in 23 categories) and [23], composed of 5-second-long recordings organized into 50 semantical classes.

#### *2.2.2. Audio-Text Embedding in Vector Database*

While previous attempts make use of table-based databases that associates each sound file with its title, in this paper we experiment with storing Urban Sound data in ChromaDB, which is a vector database that stores the LAION-CLAP embedding [24]. This allows to retrieve files based on the similarities between latent representations of given query (prompt) and audio.

#### *2.2.3. String List Chopping Experiment (SLiCE)*

We use the SLiCE method [12] to regularize the query prompts from Section 2.1. The SLiCE method allows audio file searches recursively based on the search results. In the event that embedding distance between query and search result is too large, sublists are used to generate sub-queries, until the point where enough results are returned to include all word-features (or soundscape features).

### 2.3. Mixing

#### *2.3.1. BRIR Fine-Tuning for Outdoor Scenarios*

In this experiment we build our surrogate BRIR model based on works from [25], [26], due to its capability of handling material data (scattering and absorption coefficients) via latent graphical representation. To start with, 10 neighborhoods are selected within Stuttgart based on urban typologies. We then extract the geometries via Blender and Blosm, subsequently generate the datasets for fine-tuning via outdoor sound propagation simulations in [27]. The output datasets is hosted at [28].

#### *2.3.2. ONNX Convolution*

The model produced in Section 2.3.1 is subsequently converted to Open Neural Network Exchange (ONNX) format, that allows interoperability in different platforms. To cover our use case as indicated in Section 1 for urban planners in Rhino, we integrated the `Microsoft.ML.OnnxRuntime` with plugins developed in the `RhinoCommon` framework.

#### *2.3.3. Temporal Sequencing*

We use the open-source framework in [29] to knit the input contexts, selected audio background and foreground files after BRIR convolution together. Each foreground sound effect are treated as independent probabilistic events. We derive the probabilities based on the raw recordings from datasets on [29] and [30].

## 3. Discussions, Limitations & Future Works

In this paper we described a framework that aims to synthesize efficiently plausible audios, that can digest general geo-spatial contexts and text prompts. Via revisiting some methodologies from the acoustic ecology community, we are able to produce high-quality sound and address the issue of lack of semantic understanding in the state-of-the-art auralizations as well as TTA models.

However, there are still a few limitations regarding the usage of this sound generator. Firstly, the uncertainties borne from the environment seriously affects the plausibility and applicability, for instance, a building with different type of concrete could make a difference. Besides, there is a lack of benchmarks that can vet the plausibility of given synthesized audio.

For future works, it is intuitive to extend the synthesizer frameworks to further automatize the sound generation flow to reduce user friction. Lightweight generative audio models can be adapted to generate different sound effects tracks. The integration of this framework into architects' workflow and its effectiveness for designing better outdoor soundscapes should also be further investigated.

## References

[1] Eurocities, "Eurocities report on noise in cities 2025," Eurocities, Tech. Rep., Jun. 2025, Accessed via Horizon Europe NCP Portal. [Online]. Available: `https://horizoneuropencpportal.`

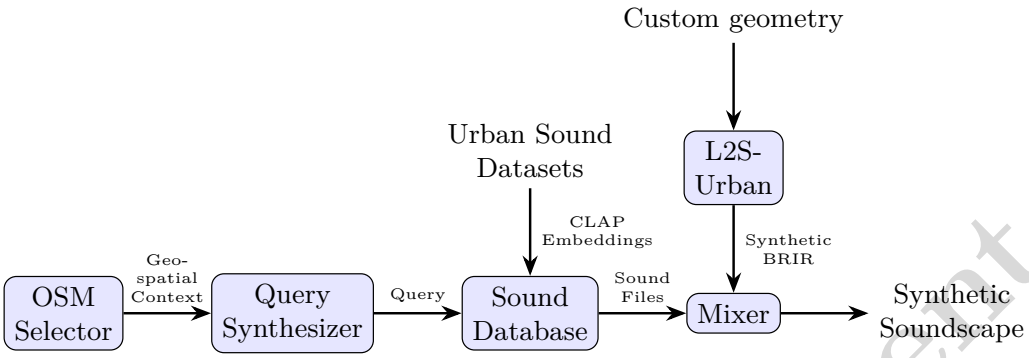


**Figure 1:** Functional Blocks with Inputs and Outputs



```
You are a sound design assistant for ambient soundscapes.
Based on the POI type and any provided mood, action, atmosphere, distance, or location context,
describe the typical continuous background sound floor of the place.
Return exactly 5 descriptors total, one for each category:
mood, setting, sound_type, intensity, character.
Choose concise, accurate, audible, and varied descriptors.
Prefer specific terms over generic ones.
Avoid repeating common fallback words unless clearly justified.
Only describe realistic continuous ambience, not isolated foreground events or one-shot SFX.
If uncertain, choose a broader but plausible descriptor.
If distance data is provided, prioritize nearer POIs.
Use location context to adapt to urban, suburban, rural, industrial, residential, coastal, or similar settings.
Output exactly this JSON format:
{"sounds": [
{"category": "mood", "value\": "..."},
{"category": "setting", "value": "..."},
{"category": "sound_type", "value": "..."},
{"category": "intensity", "value": "..."},
{"category": "character", "value": "..."}
]}.
```

**Figure 2:** Sample prompt for generating POIs

eu / sites / default / files / 2026 - 01 / eurocities-report-on-noise-in-cities-2025.pdf

[2] J. Llorca-Bofí, H. H. Au-Yeung, M. Lippold, and C. Reicher, "Bringing soundscape auralizations to the urban city planning," in *Proceedings of Forum Acusticum / Euronoise 2025 (11th Convention of the European Acoustics Association)*, Málaga, Spain, Jun. 2025, pp. 1155–1158. DOI: 10.61782/fa.2025.0181

[3] X. Yang, Z. Han, X. Lu, and Y. Zhang, "A rapid approach to urban traffic noise mapping with a generative adversarial network," *Applied Acoustics*, vol. 228, p. 110 268, 2025.

[4] R. M. Schafer, *The soundscape: Our sonic environment and the tuning of the world*. Simon and Schuster, 1993.

[5] B. Truax, "Soundscape, acoustic communication and environmental sound composition," *Contemporary music review*, vol. 15, no. 1-2, pp. 49–65, 1996.

[6] International Organization for Standardization, "Acoustics – soundscape – part 1: Definition and conceptual framework," International Organization for Standardization, Geneva, CH, Standard ISO 12913-1:2014, 2014. [Online]. Available: https://www.iso.org/standard/52161.html

[7] J. Kang and B. Schulte-Fortkamp, *Soundscape and the built environment*. CRC press Boca Raton, FL, USA: 2016, vol. 525.

[8] R. Ayoubi, S. B. Park, and L. Lescop, "Artificial intelligence in creating, representing or expressing an immersive soundscape," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Institute of Noise Control Engineering, vol. 270, 2024, pp. 7818–7828.

[9] M. Heydari, M. Souden, B. Conejo, and J. Atkins, "Immersediffusion: A generative spatial audio latent diffusion model," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.

[10] F. Rodrigues, "Synthetic ornithology: Machine learning, simulations and hyper-real soundscapes," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2025, pp. 84–92.

[11] G. Samson, "Perspectives on generative sound design: A generative soundscapes showcase," in *Arts*, MDPI, vol. 14, 2025, p. 67.

[12] M. A. Thorogood, "Audio metaphor: A framework for automatic soundscape composition," Ph.D. dissertation, Simon Fraser University, 2018. [Online]. Available: https://summit.sfu.ca/item/18165

[13] E. Grinfeder, C. Lorenzi, Y. Teytaut, S. Haupert, and J. Sueur, "Introducing evascape, a model-based soundscape assembler: Impact of background sounds on biodiversity monitoring with ecoacoustic indices," *Ecological Indicators*, vol. 178, p. 113 882, 2025.

[14] H. Liu et al., "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.

[15] C.-Y. Hung et al., "Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization," *arXiv preprint arXiv:2412.21037*, 2024.

[16] S. Della Torre, M. Pezzoli, F. Antonacci, and S. Gannot, "Diffusionrir: Room impulse response interpolation using diffusion models," *arXiv preprint arXiv:2504.20625*, 2025.

[17] C. Templin, Y. Zhu, and H. Wang, "Generating moving 3d soundscapes with latent diffusion models," in *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2026, pp. 2856–2860.

[18] OpenStreetMap Wiki, *Points of interest*, https://wiki.openstreetmap.org/wiki/Points_of_interest, Accessed: 2026-07-02, 2026.

[19] N. Finney and J. Janer, "Soundscape generation for virtual environments using community-provided audio databases," in *W3C Workshop: Augmented Reality on the Web*, Barcelona Spain, 2010.

[20] Freesound Team, *Freesound.org database*, https://freesound.org/, Accessed: 2026-06-25, 2026.

[21] J. Salamon, C. Jacoby, and J. P. Bello, "Urbansound8k," *Urban Sound Datasets*, 2014.

[22] M. Cartwright et al., *Sonyc urban sound tagging (sonyc-ust): A multilabel dataset from an urban acoustic sensor network.* Version 2.3.0, Zenodo, Sep. 2020. DOI: 10.5281/zenodo.3966543 [Online]. Available: https://doi.org/10.5281/zenodo.3966543

[23] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*, Brisbane, Australia: ACM Press, Oct. 13, 2015, pp. 1015–1018, ISBN: 978-1-4503-3459-4. DOI: 10.1145/2733373.2806390 [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390

[24] Y. Wu*, K. Chen*, T. Zhang*, Y. Hui*, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

[25] A. Ratnarajah, Z. Tang, R. Aralikatti, and D. Manocha, "Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 924–933.

[26] A. Ratnarajah and D. Manocha, "Listen2scene: Interactive material-aware binaural sound propagation for reconstructed 3d scenes," in *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, IEEE, 2024, pp. 254–264.

[27] P. Palenda and M. Vorlaender, "Auralizing arbitrary urban environments including diffraction," *Acta Acustica*, Apr. 2026. DOI: 10.1051/aacus/2026033

[28] Fraunhofer-IBP-UAA, *Listen2scene-stuttgart-outdoor (revision 0e82054)*, 2026. DOI: 10.57967/hf/9381 [Online]. Available: https://huggingface.co/datasets/FhG-IBP-UAA/listen2scene-stuttgart-outdoor

[29] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2017, pp. 344–348.

[30] M. Fuentes et al., "Soundata: Reproducible use of audio datasets," *Journal of Open Source Software*, vol. 9, no. 98, p. 6634, Jun. 2024. DOI: 10.21105/joss.06634 [Online]. Available: https://joss.theoj.org/papers/10.21105/joss.06634