Machine Learning for Acoustical In-Painting in AR: Enhancing Immersive Audio Realism

Nicolas Obin, Markus Noisternig, Benoit Alary (UMR 9912 - STMS)

In recent years, the emergence of virtual and augmented reality technologies (VR/AR) has produced many scientific and technical challenges in closely related fields. For sound reproduction, one key challenge is reproducing a virtual sound source in a specific acoustics environment, a process called acoustic auralization. More specifically, in AR applications, real world sounds need to coexist seamlessly with virtual ones in the same acoustics environment, and this highly critical auralization process is performed using a very limited set of information. Typically, a head-mounted device is equipped with cameras which can yield information about a visual environment, and microphones to capture information on the sound field. Currently an active area of research, the aim is to use an analysis-synthesis process [1], using the data collected through these devices, to resynthesize a virtual acoustic scene seamlessly placing any virtual sound sources in the same acoustical environment as real ones.

Machine learning (ML) is rapidly becoming the prevalent foundation for solving complex audio signal processing problems as well as tasks related to room acoustics. The inherent statistical properties embedded within audio signals makes them good candidates for ML optimization. In the context of room acoustics, recently proposed methods include the use of physically-informed neural networks for sound field reconstruction [2], blind acoustic parameter estimation [3, 4], and acoustical inverse problems to infer physical information, such as walls locations, from acoustical measurements [5]. However, training these ML models remains a significant challenge, as they require very substantial datasets of acoustical footprints, known as room impulse responses (RIRs), which are inherently difficult and time-consuming to acquire. As a result, research typically relies on acoustic simulation to produce large datasets [6, 7, 8]. However, due to simplifications inherent to these simulation algorithms, the data they produce tend to differ – from a signal processing perspective – from data measured in real rooms. Consequently, many acoustic-related ML models struggle to deliver accurate results in practical scenarios, when real data is used for inference [5].

This thesis project aims to overcome these limitations by developing state-of-the-art parametric reverberation methods [9, 10] to synthesize large datasets of realistic spatial room impulse responses (SRIRs) optimized for immersive audio applications [11, 12]. The synthesis approach will rely on acoustically relevant parameters and be evaluated for realism to enhance ML training. Beyond broader implications for acoustics-related ML, this method will significantly improve blind auralization in AR for real-world applications. The parametric reverberator will model key acoustic properties, including time-frequency and directional decay profiles [12], source-receiver positions, multi-exponential reverberation times, spatial coherence [13], spatially distributed early reflections [14], and temporal echo density evolution. Novel synthesis techniques will be tailored to reproduce the complex transition from (early) specular reflections to late reverberation, incorporating scattered reflections and background noise to simulate natural environments. Ultimately, this parametric approach will achieve superior spectral, temporal, and spatial alignment with real-world SRIRs compared to traditional reverberation and geometric simulation methods.

At the start of the project, we will use a dataset of generated SRIRs to build an acoustic parameter inference model. This initial ML model will validate our approach by training on synthetic data and inferring parameters – such as reverberation time [15] – from measured data. In the next phase, focused on AR applications, our synthesis method will generate a larger and more diverse set of SRIRs (Fig.1). These will be convolved with anechoic speech to train a diffusion model for reverberant speech. Using microphone signals as input, the model will perform acoustic in-painting on dry sounds, generating realistic reverberated signals. This process will enable a fully automated, end-to-end blind auralization framework that dynamically places virtual sound sources at unseen locations based on evolving acoustic conditions. Our approach, will also be assessed and improved for challenging practical conditions. For instance, beamforming analysis of microphone array input signal allow multiple simultaneous sound sources in a room. Furthermore, in a field highly dependent on the quality of its training data, the

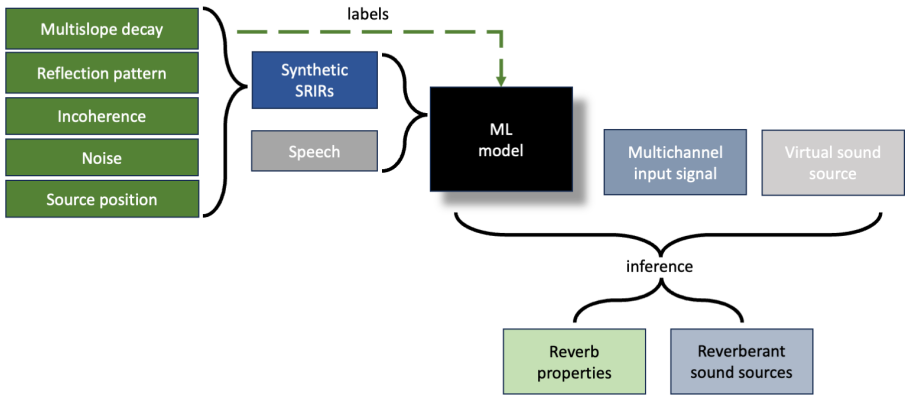

Figure 1: *Proposed acoustical audio in-painting and parameter inference model.*

developed synthesis framework will offer much broader applications, and have a deep impact in many acoustics-related ML tasks. Toward this objective, the resulting synthesis framework will be released as open-source.

The PhD candidate will work under the guidance of a multidisciplinary team, including thesis director Nicolas Obin (HDR) and co-advisors Markus Noisternig (PhD) and Benoit Alary (PhD), bringing expertise in machine learning, audio signal processing, acoustics, and spatial audio analysis/synthesis. The research will be conducted at IRCAM (STMS, UMR 9912), a leading institution dedicated to bridging scientific research, technological innovation, and artistic creation in sound and music. The ideal candidate should have a strong background in signal processing and a solid understanding of machine learning methods, enabling them to contribute effectively to this cutting-edge research.

References

[1] F. Lluís and N. Meyer-Kahlen, "Blind spatial impulse response generation from separate room- and scene-specific information," 2024.

[2] X. Karakonstantis, D. Caviedes-Nozal, A. Richard, and E. Fernandez-Grande, "Room impulse response reconstruction with physics-informed deep learning," *The Journal of the Acoustical Society of America*, vol. 155, pp. 1048–1059, 02 2024.

[3] P. Srivastava, *Realism in virtually supervised learning for acoustic room characterization and sound source localization.* Theses, Université de Lorraine, Nov. 2023.

[4] P. Srivastava, A. Deleforge, and E. Vincent, "Blind room parameter estimation using multiple multichannel speech recordings," in *WASPAA 2021- IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (New Paltz, NY, United States), Oct. 2021.

[5] T. Sprunck, A. Deleforge, Y. Privat, and C. Foy, "Fully Reversing the Shoebox Image Source Method: From Impulse Responses to Room Parameters," *working paper or preprint*, May 2024.

[6] E. A. Lehmann and A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Trans. Audio, Speech, Language Proces.*, vol. 18, no. 6, pp. 1429–1439, 2010.

[7] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *J. Acoust. Soc. Am.*, vol. 138, pp. 708–730, Aug. 2015.

[8] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl, "A round robin on room acoustical simulation and auralization," *J. Acoust. Soc. Am.*, vol. 145, pp. 2746–2760, Apr. 2019.

[9] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 1421–1448, Jul. 2012.

[10] V. Välimäki, B. Holm-Rasmussen, B. Alary, and H.-M. Lehtonen, "Late reverberation synthesis using filtered velvet noise," *Appl. Sci.*, vol. 7, May 2017.

[11] B. Alary, A. Politis, S. J. Schlecht, and V. Välimäki, "Directional feedback delay network," *J. Audio Eng. Soc.*, vol. 67, pp. 752–762, Oct. 2019.

[12] B. Alary, P. Massé, S. J. Schlecht, M. Noisternig, and V. Välimäki, "Perceptual analysis of directional late reverberation," *J. Acoust. Soc. Am.*, vol. 149, pp. 3189–3199, May 2021.

[13] P. Massé, T. Carpentier, O. Warusfel, and M. Noisternig, "Denoising Directional Room Impulse Responses with Spatially Anisotropic Late Reverberation Tails," *Applied Sciences*, vol. 10, p. 1033, Feb. 2020.

[14] P. Massé, A. Gallien, W. Kreuzer, and M. Noisternig, "Echo detection using the herglotz wavefunction in spatial room impulse responses measured with spherical microphone arrays," 2025. Manuscript submitted for publication.

[15] G. Götz, R. Falcón Pérez, S. J. Schlecht, and V. Pulkki, "Neural network for multi-exponential sound energy decay analysis," *The Journal of the Acoustical Society of America*, vol. 152, pp. 942–953, 08 2022.