Audio Engineering Society
Convention Express Paper 90

Presented at the 154th Convention
2023 May 13-15, Espoo, Helsinki, Finland

This Express Paper was selected on the basis of a submitted synopsis that has been peer reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This express paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

The effect of geometry information in blind estimation of room acoustic parameters

Nikolaos Vryzas[1], Lazaros Vrysis[1], Marina Eirini Stamatiadou[1], Rigas Kotsakis[2], and Charalampos Dimoulas[1]

[1] *Aristotle University of Thessaloniki, 54124, Greece*

[2] *Department of Information and Electronic Engineering, International Hellenic University of Greece (IHU), Sindos Campus, GR-57400 Thessaloniki, Greece*

Correspondence should be addressed to Nikolaos Vryzas (nvryzas@auth.gr)

ABSTRACT

This paper investigates the effect of prior knowledge of a room's geometry on the accuracy of predicting its acoustic properties. The motivation comes from the need for the creation of a framework that reduces the complexity of modeling real spaces. The project aims to model cultural heritage spaces for film-making applications, like location scouting. Preliminary investigation is performed on an existing dataset, where reverberation and early decay time are estimated in several octave bands through audio feature-driven regression. The room volume is treated as a feature, and performance is calculated with and without including it in training. Results indicate a slight improvement in performance in most bands. This leads to the conclusion that additional geometrical information should be provided during dataset creation for the significant improvement of robustness.

1 Introduction

Room acoustics play a significant role in the audio attributes of an event. It is a common practice in audio and film production to carefully select or treat a room to achieve a target result. Auralizing the effect of a specific space refers to the task of creating a simulated version of a recording as if it had been recorded in that space. The recording can take place in anechoic studio conditions and then is edited to resemble another environment [1].

This process is essential for the creation of simulated spaces, which is an indispensable part of immersive experiences, video games, and XR environments. It can also be proved very useful in location scouting (repérage) where there is the need to forecast the adequacy of a place, according to the desired aesthetic criteria. In this direction, the SCENE (Searchable multi-dimensional Data Lakes supporting Cognitive Film Production & Distribution for the Promotion of the EuropeaN Cultural HeritagE) project[1] aims to address the cycles of pre- and post- production of the film industry by modeling places of significance for the European cultural heritage, both visually and acoustically. The present research takes place within the context of the SCENE project.

In most cases, the described task consists in extracting the Room Impulse Response (RIR) of a place [2]. Convolving the RIR with the signal of a recording that was held in anechoic conditions results in a simulation of how it would sound. This can be achieved by measuring the RIR with appropriate equipment. This method adds practical complexity and might not be a suitable choice for the creation of

[1] http://thesceneproject.eu/

large datasets, which is the scope of the current project. Another approach is to simulate the acoustic conditions taking into account the geometrical characteristics, surface materials, etc, which adds computational complexity to the problem. There are however modern solutions for the GPU acceleration of such processes [3].

In the past few years, with the advance of deep learning techniques, blind audio matching has got more attention [4], [5]. In this approach, having a dataset of recordings along with the outcome of their convolution with RIRs of spaces with known acoustic conditions, a deep model is trained to transform a signal or calculate room acoustics attributes [6], [7]. In some applications, the change of acoustic conditions through time can be taken into consideration [8].

In [1], motivation from the empirical fact that experienced audio engineers can make rough estimations of room acoustics by seeing a room leads to the evaluation of audiovisual input in the blind estimation of acoustic parameters. Evidence from the literature indicates that the knowledge of certain acoustic properties can practically be enough for audio matching and signal generation. Acoustic matching aims at transforming and audio recording as it had been recorded in another space [9]. In auralization tasks, it is feasible to make assumptions based on psychoacoustic perception [10]. This means that is not necessary to simulate completely the audio signal in an intended space, but rather the listener's perception of it. Besides human listeners, in some cases, such techniques can be used for data augmentation and generation, to improve the robustness and generalization of machine learning classifiers [11].

The contribution of this work is that it investigates a solution for a practical problem in realistic conditions, highlighting the significance of room geometry in performance improvement. The procedure is meant to be integrated into a framework for reducing the complexity of the modeling of real spaces, allowing the crowdsourcing of room measurements. The overall goal of this research aims at the creation of a real-world dataset, which is also an important expected outcome of the project, and the evaluation of acoustic matching performance in real conditions through subjective listening tests. Different information fusion techniques will be applied for the most efficient inclusion of geometrical information within the model. At the current stage, emphasis is given to the investigation of available resources and material, preliminary evaluation of the project's research hypothesis, and the formulation of a solid methodology for audio measurements and dataset creation.

2 Creating XR content data lakes for location scouting and cultural heritage preservation

2.1 An exploration of datasets for room acoustic parameter estimation

There are several datasets in modern literature for the blind estimation of acoustic parameters that can be proved useful. The ACE challenge dataset provides the RIRs of several rooms, along with known acoustic properties [12]. The Arni [13] dataset includes different RIRs from the same room with different configurations in the placement of acoustic panels. The dEchorate [14] dataset includes annotations of early echo timings and 3D positions of microphones, real sources, and image sources under different wall configurations in a cuboid room. MeshRIR [15] consists of two sub-datasets: one with IRs in a 3D cuboidal region from a single source, and the other with multi-source IRs in a 2D square region. In BUT Speech@FIT Reverb Database [16] various RIRs are measured, for different microphone and sound source positionings, in 8 different rooms. In OpenAir dataset, ambisonics recordings are provided in over 46 places [17]. Along with the audio recordings, photos are available, allowing a joint audiovisual estimation of acoustic parameters [1]. A more comprehensive list of available resources is presented in [18], along with links and documentation.

2.2 The SCENE approach.

The SCENE project aims at providing an end-to-end solution for filmmaking, addressing the stages of preproduction, production, and postproduction, promoting collaboration, and reusability of resources. Monuments of cultural heritage are considered a very important asset for the European film industry, as they are highly favored filming locations.

In this direction, 3D modeling aims at supporting the procedures of location scouting, through the creation of XR data lakes. Important spaces will be modeled so that directors, producers, etc. can perform the previsualization of the spaces before filming. An XR environment will allow the virtual placing of actors on set. An artificial lighting module will allow the simulation of different conditions for shooting. Along with the previsualization functionality, an

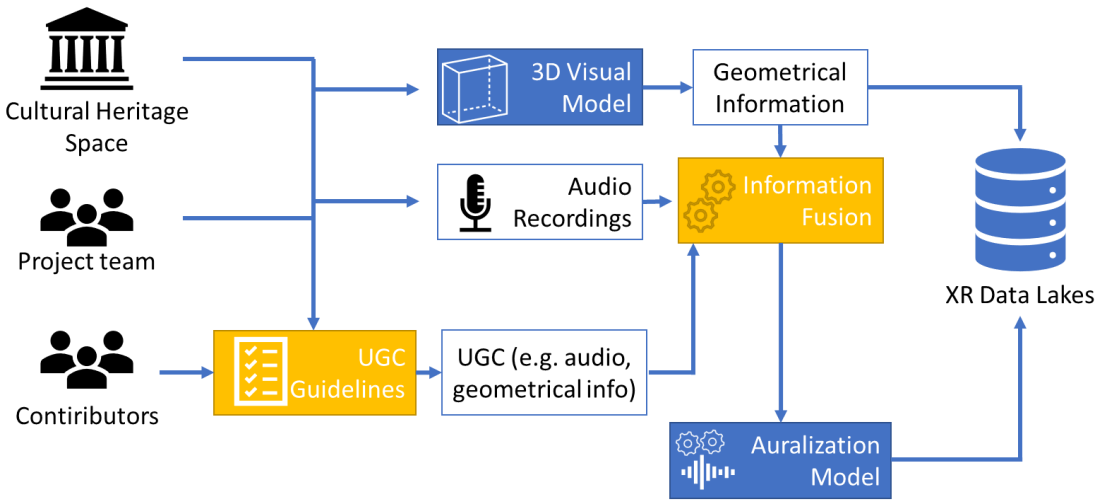


Figure 1. The SCENE dataset creation ecosystem. XR data lakes are created to store 3D visual and auralization models of cultural heritage spaces. Independent users can follow the defined guidelines for content contribution to store recordings, along with geometrical information, images, etc. to the data lakes.

auralization module will also be in place. The creators should be given the capability to have a prediction of the acoustic behavior of this space. To achieve this, an acoustic matching model should be in place for every modeled space, that accepts audio recorded in anechoic conditions and provides as output a prediction of how it would sound on set. It is evident from 2.2 that existing datasets for addressing similar tasks are quite heterogeneous.

Within the project duration, some pilot cases will be carried out for the validation of this concept. However, in terms of sustainability and extendibility of the project's impact, a methodology has to be set for the evolution of the formulated data lakes in time. Blind acoustic matching has emerged in the past years as a feasible approach for auralization, as mentioned in the introduction. This allows the creation of datasets without precise acoustic measurements. For the acoustic modeling of cultural heritage spaces around Europe, collaboration is the only sustainable approach. Following the paradigm of [19], crowdsourcing can be integrated, allowing users to provide audio data captured with non-professional devices. However, this requires the definition of precise guidelines for content creation. It is evident from 2.2 that existing datasets for addressing similar tasks are quite heterogeneous. This makes it difficult to use such data efficiently in a cross-corpora task. The ecosystem for the creation of such data lakes by the project team, and in collaboration with external contributors who follow the defined guidelines for User Generated Content Creation (UGC) is depicted in Figure 1.

The theory of room acoustics defines the significance of room geometry and material in reverberance and the overall acoustic behavior of a closed space [20]. In the case of blind acoustic matching or parameter estimation, data-driven techniques are used to fine-tune the model without any prior knowledge. The research hypothesis behind the SCENE approach is that additional information concerning the room geometry can be utilized as a useful feature or act as a correction factor in blind modeling. In the case of the 3D modeling of spaces, which is relevant for the pilot validation of the project, all room geometry will be known and could be extracted from the model. In the case of audio measurement crowdsourcing, the type of information that can improve robustness is under investigation. Through the overview of current datasets and materials, it is evident that the room volume is a type of information that can be accessed more easily. In the direction of defining the guidelines for dataset contribution, the volume is evaluated as the minimum prior knowledge of room geometry.

3 Investigation of the room size effect on acoustic parameter estimation

Geometrical information includes precise dimensions, room volume, wall/floor/ceiling surface, etc. For the experimentation, the OpenAir [17] dataset was selected. The dataset includes several rooms,

including churches, public buildings, places related to cultural heritage, etc. In this aspect, it is quite relevant to the SCENE context. For every room, the Reverberation Time RT60 – T30, and Early Decay Time are provided, as well as some speech intelligibility metrics, namely Definition D50, Clarity C50, and Clarity C80. The same set of audio examples is used in every room, including a voice and a drum sample. The recordings are available in B-Format. For the current experiment, the W component of the B-Format is used, providing the output for the omnidirectional polarity. A short description, some photos, and the volume of each space in m^3 are provided as well.

For the fusion of audio input and volume information, an early information fusion approach is selected. In previous work, in [6], convolutional neural networks were used for the estimation of acoustic parameters. This is feasible when the room impulse responses are available and it is possible to artificially generate a big dataset, by providing audio input of unconstrained length. When the room geometry is fused within the available information, there is a constraint in dataset size, which is defined by the number of measured rooms. Providing additional instances from the same room would cause overfitting, as there is a one-to-one matching of room size and room acoustic parameters. For this reason, a feature-based approach was preferred in the current context, which is more suitable for smaller datasets. The feature vector was based on the MIR toolbox set and previous experimentation with it [22], [23]. The provided room volume in m^3 was integrated into the original audio feature vector.

4 Results and discussion

For the evaluation, only the selected spaces from the Openair dataset with available volume information were selected, leading to a subset of 29 rooms. For every room, the human voice example was used, as it is intuitively considered more relevant to the context of acting in film production. The feature vector that was described in Section 3 was extracted for each audio file. For modeling evaluation, the RapidMiner Studio environment was used [23].

The calculation of RT60 and EDT for every octave band was treated as a distinct regression task. The Random Forest algorithm was used with 100 random trees and a least square criterion. The mean squared error was used as the performance metric. The models were trained twice, once without including the volume information as a feature, and the second one including it, to compare performance. This led to a total of 40 separate trained models, 2 for each defined task.

The results are presented in Table 1, for each parameter and octave band respectively. It is evident that the room volume provides additional useful information, improving performance in most bands. Its contribution seems to be less important in the two lower octaves. This frequency range is however less crucial for speech recordings. In any case, performance is not greatly improved. The current results provide an intuition on the effect of volume as a feature vector. Additional experimentation is needed to examine its effect in more complex architectures that allow unsupervised feature extraction, like convolutional neural networks, as it was evaluated in previous related work, in [6], [7]. This will be feasible when a more extended dataset in terms of room size information will be available. However, initial experimentation and exploratory data analysis indicate that the volume information itself, while it offers convenience in terms of calculating is rather insufficient. Additional geometrical parameters should be included in the dataset creation guidelines.

5 Conclusions and future work

This paper introduces the problem of blind calculation of acoustic room parameters with prior knowledge concerning room geometry. In the current,

Table 1. Results of the effect of room volume information in various octave bands. The mean square error is used as a metric.

Reverberation Time RT-60 T30 (s)										
Octave	**31.25Hz**	**62.5Hz**	**125Hz**	**250Hz**	**500Hz**	**1kHz**	**2kHz**	**4kHz**	**8kHz**	**16kHz**
w/o volume	10.882	9.315	7.870	3.043	2.718	2.278	1.537	1.076	0.622	2.434
with volume	11.010	9.382	7.801	2.944	2.695	2.246	1.534	1.033	0.630	2.441
Early Decay Time (s)										
w/o volume	3.219	3.437	3.037	2.641	2.189	1.850	1.270	0.763	0.446	0.289
with volume	3.199	3.246	2.804	2.613	2.140	1.792	1.253	0.762	0.436	0.288

preliminary investigation, a well-known dataset is used, that provides volume information. Experimentation showed that this information alone improves the results, but not significantly. More geometrical parameters should be evaluated, along with unsupervised feature extraction, and deeper machine-learning architectures that offer advanced computational capacity. Late information fusion can be considered, where geometrical information acts as a correction factor after model results.
Existing datasets were investigated for the purposes of the current project. An important direction of this research is to define the guidelines for the creation of an extended dataset of measurements in cultural-heritage spaces. In the context of crowdsourcing information, the simplest solution should be found in terms of methodology and equipment, providing the best possible results, in combination with 3D modeling and geometrical information.

6 Acknowledgment

This research is part of the SCENE project and has received funding from the European Union's Horizon Research and Innovation programme under grant agreement No 101095303.

References

[1] Kon, H., & Koike, H. (2018, May). Deep neural networks for cross-modal estimations of acoustic reverberation characteristics from two-dimensional images. In Audio Engineering Society Convention 144. Audio Engineering Society.

[2] Yu, W., & Kleijn, W. B. (2020). Room acoustical parameter estimation from room impulse responses using deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 436-447.

[3] Diaz-Guerra, D., Miguel, A., & Beltran, J. R. (2021). gpuRIR: A python library for room impulse response simulation with GPU acceleration. Multimedia Tools and Applications, 80, 5653-5671.

[4] Chai, M., Li, T., Zhu, M., Wang, T., & Zhang, W. (2019, July). Blind estimation of reverberation time using binaural complex ideal ratio mask. In 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (pp. 378-383). IEEE.

[5] Deng, S., Mack, W., & Habets, E. A. (2020, October). Online Blind Reverberation Time Estimation Using CRNNs. In INTERSPEECH (pp. 5061-5065).

[6] Thoidis, I., Vryzas, N., Vrysis, L., Kotsakis, R., Kalliris, G., & Dimoulas, C. (2022, May). Disentangled estimation of reverberation parameters using temporal convolutional networks. In Audio Engineering Society Convention 152. Audio Engineering Society.

[7] Vryzas, N., Katsaounidou, A., Vrysis, L., Kotsakis, R., & Dimoulas, C. (2022). A Prototype Web Application to Support Human-Centered Audiovisual Content Authentication and Crowdsourcing. Future Internet, 14(3), 75.

[8] Götz, P., Tuna, C., Walther, A., & Habets, E. A. (2022, May). Blind reverberation time estimation in dynamic acoustic conditions. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 581-585). IEEE.

[9] Su, J., Jin, Z., & Finkelstein, A. (2020, May). Acoustic matching by embedding impulse responses. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 426-430). IEEE.

[10] Hacihabiboglu, H., De Sena, E., Cvetkovic, Z., Johnston, J., & Smith III, J. O. (2017). Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics. IEEE Signal Processing Magazine, 34(3), 36-54.

[11] Kim, C., Misra, A., Chin, K., Hughes, T., Narayanan, A., Sainath, T., & Bacchiani, M. (2017). Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home.

[12] Eaton, J., Gaubitch, N. D., Moore, A. H., & Naylor, P. A. (2016). Estimation of room acoustic parameters: The ACE challenge. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(10), 1681-1693.

[13] Prawda, K., Schlecht, S. J., & Välimäki, V. (2022). Calibrating the Sabine and Eyring formulas. The Journal of the Acoustical Society of America, 152(2), 1158-1169.

[14] Carlo, D. D., Tandeitnik, P., Foy, C., Bertin, N., Deleforge, A., & Gannot, S. (2021). dEchorate: a calibrated room impulse response dataset for echo-aware signal processing. EURASIP Journal on Audio, Speech, and Music Processing, 2021, 1-15.

[15] Koyama, S., Nishida, T., Kimura, K., Abe, T., Ueno, N., & Brunnström, J. (2021, October). MeshRIR: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods. In 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (pp. 1-5). IEEE.

[16] Szöke, I., Skácel, M., Mošner, L., Paliesek, J., & Černocký, J. (2019). Building and evaluation of a real room impulse response dataset. IEEE Journal of Selected Topics in Signal Processing, 13(4), 863-876.

[17] Murphy, D. T., & Shelley, S. (2010, November). Openair: An interactive auralization web resource and database. In Audio Engineering Society Convention 129. Audio Engineering Society.

[18] Roy James. Room Impulse Responses. https://github.com/RoyJames/room-impulse-responses (retrieved 9/4/2023).

[19] Stamatiadou, M. E., Thoidis, I., Vryzas, N., Vrysis, L., & Dimoulas, C. (2021). Semantic crowdsourcing of soundscapes heritage: a mojo model for data-driven storytelling. Sustainability, 13(5), 2714.

[20] Kuttruff, H., & Mommertz, E. (2012). Room acoustics. In Handbook of engineering acoustics (pp. 239-267). Berlin, Heidelberg: Springer Berlin Heidelberg.

[21] Lartillot, O., & Toiviainen, P. (2007, September). A Matlab toolbox for musical feature extraction from audio. In International conference on digital audio effects (Vol. 237, p. 244).

[22] Vryzas, N., Kotsakis, R., Liatsou, A., Dimoulas, C. A., & Kalliris, G. (2018). Speech emotion recognition for performance interaction. Journal of the Audio Engineering Society, 66(6), 457-467.

[23] Mierswa, I., & Klinkenberg, R. (2018). RapidMiner Studio (9.2)[Data science, machine learning, predictive analytics]. Retrieved from rapidminer. com.

[24] Phan, H., Maaß, M., Mazur, R., & Mertins, A. (2014). Random regression forests for acoustic event detection and classification. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(1), 20-31.