Contextual Sound Event Classifier Models: Disentangle Acoustic Condition in Embedding Space

1st Sreenivasa Upadhyaya
Department of Computer Science, DTAI
KU Leuven
Geel, Belgium
sreenivasa.upadhyaya@kuleuven.be

2nd Wim Buyens
SoundTalks N.V.
Leuven, Belgium
wim.buyens@soundtalks.com

3rd Wim Desmet
Department of Mechanical Engineering, LMSD
KU Leuven
Leuven, Belgium
wim.desmet@kuleuven.be

4th Peter Karsmakers
Department of Computer Science, DTAI
KU Leuven
Geel, Belgium
peter.karsmakers@kuleuven.be

***Abstract*—Room acoustics significantly impact the performance of Sound Event Classification (SEC) models. To address the challenges posed by diverse real-world conditions, researchers have explored methods that incorporate knowledge of recording environments into model design and learning. Existing approaches typically focus on improving robustness by identifying acoustic condition invariant features. However, a gap remains in generalization to unseen environments. This article proposes leveraging available information about the acoustic condition to enhance SEC model robustness through (a) a contextual SEC model that incorporates known (and potentially dynamic) acoustic characteristics and (b) a multi-task learning scheme that disentangles acoustic conditions within the neural network's embedding representation. This approach allows the SEC model to adapt its behavior based on the acoustic environment. Compared to alternative methods, the proposed approach improved the weighted F1-score by 2.9% and reduced performance variation across validation folds by 5.14% in challenging, unseen acoustic conditions.**

***Index Terms*—Sound event classification, acoustics, deep learning, contextual classifier.**

I. Introduction

Sound Event Classification (SEC) refers to the task of identifying and classifying events within a sound signal. The field of SEC has widespread applications, including healthcare, security, environmental monitoring, and livestock monitoring. Microphones, the sensors behind sound capture, deliver rich data to derive insightful information about the monitored environment [1]. With the recent progress in Deep Learning (DL), several DL models have been proposed for SEC. More specifically, models that include convolutional, recurrent layers and/or transformer layers are popular and effective in SEC, as can be observed from the recent DCASE[1] challenge results of task 4.

In real-life applications, the performance of SEC models is affected by the presence of noise and room acoustic conditions, which differ from those in the training set [2]. The model robustness can generally be improved by extending the training data to include more variability in for example room acoustics. However, expanding the dataset can be costly, time-consuming, and impractical for covering all possible scenarios. Therefore, the literature has studied methods to improve the model robustness in SEC without requiring additional physical measurements.

A simple approach, that has proven to be effective in boosting performance, is to augment the data set with synthetic examples that are created by convolving (clean) data with Room Impulse Response (RIR) filters [3], [4]. Other methods to improve the model robustness employ adapting learning schemes to enforce the extraction of acoustic features that are agnostic to variations in acoustic conditions. For example, the work in [5] presented a learning method that suppresses specific environmental factors by trying to achieve condition invariant features by element-wise affine transformations between sound event features and auxiliary information from corresponding room impulse response. This successfully reduced performance degradation caused by the reverberation of the room.

Echo-aware feature refinement using spatial cues of the unknown environment obtained through measuring acoustic echoes was proposed in [6]. The feature refinement in domain adversarial training achieved generalization of features across conditions, improving the robustness of the SEC in varying acoustic conditions. The work in [7] proposed the

This work was supported by a Baekeland PhD grant of the Flanders Innovation & Entrepreneurship agency (VLAIO), Belgium (HBC.2019.2216).

[1]https://dcase.community/challenge2024/

Room Acoustic Adversarial Neural Network (RAANN) training scheme by exploiting knowledge of underlying acoustic metrics describing the properties of the underlying recording conditions. RAANN searches for a set of features that optimize the SEC performance across a range of acoustic conditions while minimizing the predictive performance of the regression function that estimates the room acoustic metrics. This improved the performance for acoustic conditions that were harder than those seen during the learning phase.

Another set of articles describes the use of domain adaptation [8] to adapt SEC models to better match the acoustic conditions of the target environment. Even though such methods have been effective in improving the performance of the target environment, they require data recorded in the target domain, and a subsequent training step to specialize the models for the target environment. This might be a bottleneck in practical settings.

Although these methods have achieved significant performance gains, they still fall short of human-level generalization [9], [10]. In applications such as livestock monitoring, acoustic conditions are highly dynamic due to factors like animal growth and environmental changes. Therefore, a contextual model capable of adapting its behavior based on the surrounding acoustic conditions is expected to be beneficial.

In this work, a novel method to learn contextual SEC models is proposed. For this purpose, it is assumed that for every recording, the RIR [11] of the room where the recording was made is available. With this RIR, the proposed learning scheme targets enhancing the internal embedding representation of the SEC model to have an improved generalization to unseen environments. Note that each classification requires a sound recording and a RIR. However, unlike domain adaptation models, no model retraining is required to adapt the model to an unseen environment.

The remainder of the paper is organized as follows. Sec. II describes the methods, including the metrics to describe the room acoustic conditions and the deep learning model architecture. The experimental dataset and results are discussed in Sec. III. The conclusions are given in Sec. IV.

II. Methods

A. SEC processing pipeline

The model architecture employed in this study is outlined in Fig. 1. The framework uses log mel-spectrogram as the input representation, which is a popular choice in deep learning based SEC models [12]. This is followed by a pretrained feature encoder (such as VGGish [13]) that generates high level feature maps which are then fed into label predictor block.

B. Learning contextual SEC models

Conventionally trained SEC models have poor generalization when trained on a dataset with a certain set of acoustical conditions and tested on a dataset with a distributional shift caused by a change in acoustic conditions. Instead of focusing on identifying acoustic condition invariant features, this research proposes to disentangle acoustic condition as a distinct factor within the neural network's embedding representation. By explicitly accounting for acoustic conditions, represented by some simple metric like Direct to Reverb Ratio (DRR) [14], the robustness of the model to novel acoustic environments could be enhanced. For instance, if acoustic conditions are factored out and represented as an embedding factor, and the training data includes rooms with both high and low DRR values, the model could generalize to intermediate DRR values. Moreover, the model may extrapolate from high to very low DRR values with little or no additional training data.

To achieve this goal, a multi-task learning framework is proposed, where the primary SEC task is supported by an auxiliary acoustic condition estimator. In this setup, both the SEC performance (classification task) and the estimation of an acoustic metric (regression task), such as DRR (related to the room where the sound is recorded), are optimized. Both tasks share a common embedding representation, which is designed to facilitate the discrimination of different sound events conditioned on the acoustic environment.

Assume a data set $\{(\mathbf{X}_i, \mathbf{y}_i, \mathbf{C}_i, \mathbf{d}_i)\}_{i=1}^{n}$ where $\mathbf{X}_i \in \mathbb{R}^{f \times t}$ and $\mathbf{C}_i \in \mathbb{R}^{f \times t}$ are time-spectral representations of a sound fragment and corresponding RIR respectively, with f the number of spectral components and t the number of time frames, $\mathbf{y}_i \in \{0,1\}^c$ a one-hot encoded vector that indicates the class label of the event where c is the number of classes, and $\mathbf{d}_i \in \mathbb{R}^m$ where m is the number of room acoustic metrics used in the acoustic estimator. The room acoustic metrics are min-max normalized to a range of 0 and 1.

The model architecture given in Fig. 1 includes an event feature encoder G_f, an acoustic condition encoder G_a, a classifier model G_y, and a regression model G_d, which estimates some metric describing the acoustic condition. The parameters in the model are optimized using the following objective,

$$\min_{G_f, G_y, G_d} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} \mathbf{y}_{i(j)} \; log(G_y\left(G_f\left(\mathbf{X}_i\right) \parallel G_a\left(\mathbf{C}_i\right)\right)) + \frac{\lambda}{n} \sum_{i=1}^{n} \sum_{k=1}^{m} \left| \mathbf{d}_{i(k)} - G_d\left(G_f\left(\mathbf{X}_i\right) \parallel G_a\left(\mathbf{C}_i\right)\right) \right| \quad (1)$$

where trade-off parameter λ balances the importance of both terms in the objective. $\mathbf{d}_{i(k)}$ corresponds to the k^{th} room acoustic metric for sample i, $\mathbf{y}_{i(j)}$ corresponds to the true label in one hot encoded format of class j for sample i, and $\parallel$ represents the concatenation operation. The classification loss L_y (left term) is based on a Categorical Cross Entropy (CCE) objective, and the regression loss L_d is based on a Mean Absolute Error (MAE) objective, as in the previous study [7].

As seen in Fig. 1, during model learning the L_y loss influences the G_y network, and L_d has an impact on G_d. They both impact the learning of G_f. Note that for G_a, a fixed model is used to translate the time-spectral representation of a (RIR) to a lower dimensional representation. Such representation can be learned using an autoencoder setup as used in

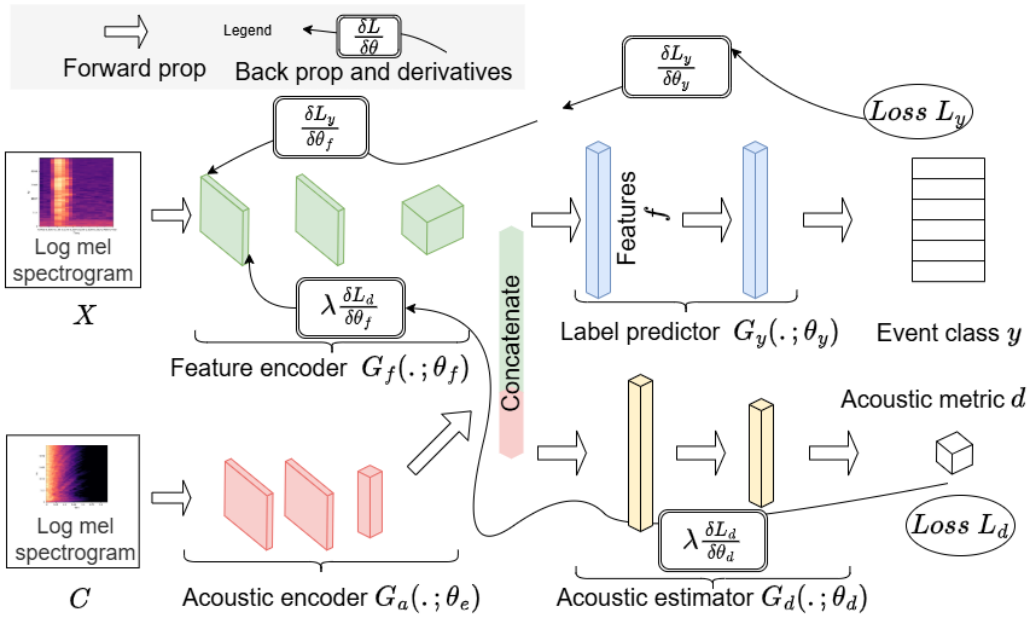

Fig. 1. Learning framework for contextual SEC.

[15]. By concatenating event features with acoustic condition encodings, factors related to the acoustic environment are incorporated into the embedding space. However, embeddings from different environments may exhibit significant overlap. To address this, when the embedding representation is used to estimate a room-specific acoustic metric, the learning algorithm encourages a representation where distributions for different acoustic conditions are more distinct. Given that the acoustic metrics are continuous, a regression-based approach is employed. Additionally, the event classifier learns to interpret features differently based on the acoustic condition. Notably, during inference, the acoustic estimator model is not utilized.

C. Room acoustic metrics

RIRs characterize the way sound gets propagated from the source to the receiver and indicates the overall perceptual quality and intelligibility of the recorded sound. The property of the room recording conditions like dimensions, building materials, distance of the source from the receiver, presence of obstacles, and reflecting surfaces, play a vital role in shaping the nature of the RIR. The RIR of a room can be measured and used to derive insightful metrics (including Reverberation Time (RT60), DRR, Clarity Index (CI), Center time (CT)) [14] that quantify the impact of room acoustic conditions on the original sound. Note that the framework can be used for any other continuous acoustic metric as well as for a combination of metrics.

III. EXPERIMENTS AND RESULTS

A. Clean sound events and RIRs

The sound events used in this study were taken from the Real World Computing Partnership (RWCP) dataset [16]. The dataset contains non-speech sounds recorded in an anechoic room. From the RWCP dataset, 80 events from each of the 50 pre-selected sound event classes were used [7]. Each sound event was recorded at a sampling frequency of 16 kHz, and the event length was 1 s. This dataset is named as ORIG dataset because it contains unmodified events from the RWCP dataset. Subsequently, the original data was augmented to create multiple modified versions by convolving it with a wide range of RIRs which are described below,

1) Simulated RIRs (SIM)
   These RIRs are simulated using the Python RIR-generator utility [17] based on the image source method. The key configuration parameters that can be modified are room dimension, sound source position, receiver position, and target RT60 values. A total of 40,000 RIRs were generated in total, covering different combinations of the above parameters to get a wide variety of RIRs.
2) Echo Thief RIRs (ET)
   Echo Thief [18] is a collection of RIRs measured in real-world conditions. This is a library of RIRs of unique spaces from around North America, including caves, skateparks, stairwells, underpasses, glaciers, fortresses, and more. Unlike simulated ones, these RIRs include the effect of real acoustic spaces with different materials and interiors, which brings more diversity in the data. A total of 74 RIRs were collected.

B. Generated datasets

Four training and test set combinations were generated, in which the test set always contains more challenging conditions compared to the training set. Firstly, the ET set of RIRs was split into two parts: the subset ET_0 has RIRs that have DRR values in the interval $[-5, 14)$, and the subset ET_1 that

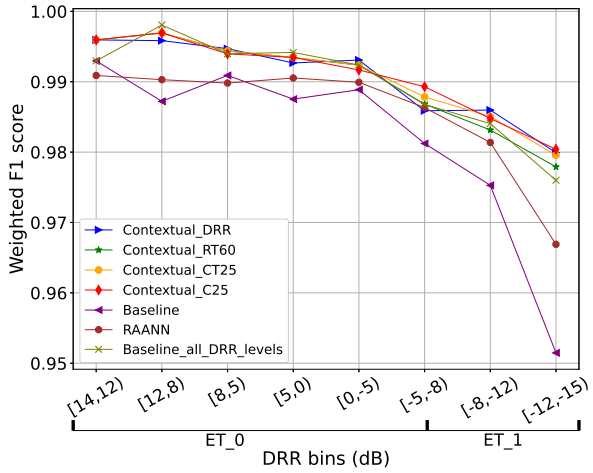

Fig. 2. Model performance across different DRR conditions.

has more challenging DRR values in the interval $[-15, -5)$. The split could have been on any of the three room acoustic metrics: DRR, RT60, or C25. In preliminary experiments [7], it was seen that all three acoustic metrics were well correlated with classifier classification performance. Secondly, the data in ORIG was split into four folds where, in each fold, the number of examples per event class is balanced. In this way, four different partitions of training (that has 3 folds) and test (the remaining fold) are created. Thirdly, to mimic different recording conditions each clean event is replaced by a version of the event itself convolved with a RIR that a) for training is sampled from the SIM and ET_0 sets, and b) for testing is sampled from the ET_0 and ET_1 sets. As a result, four combinations of training and testing are generated.

C. Model architectures and learning parameters

The feature encoder (G_f) (see Fig. 1) is initialized by a room invariant event feature encoder using the method described in [7]. The acoustic condition encoder (G_a) is obtained from a pre-trained CNN autoencoder [15]. G_a is a fixed encoder network that generates a 64 dimension feature vector for an input RIR spectrogram. The label predictor was comprised of two fully connected layers with 512 neurons each. The output event label predictions are generated by a softmax activation on the outputs of the last dense layer with 50 neurons. The models are trained with an Adam optimizer, with an initial learning rate of $1e-2$, and a batch size of 64.

D. Results

In Fig. 2 different models learned by alternative learning schemes were compared to each other in terms of their weighted average F1 score ($F_1^{(w)}$) [7], calculated using 4 fold cross-validation (with folds as defined in Sec. III-B), across different DRR ranges. For each of the models the same model architecture as described in Sec. III-C was trained with the considered learning procedure.

First, a *baseline* model was trained using a traditional supervised learning scheme. Next, the *baseline_all_DRR_level* model was learned in a similar fashion using the same folds as in the previous experiment, but now events in the training set were also convolved with RIRs that are present in the test set (though the events themselves are never shared between training and test). The latter represents a best-case scenario where the data from the unseen test data is recorded in conditions that were also present during training. It is important to note that no noise was added to the simulated data, which could explain the high baseline performance even at relatively low DRRs.

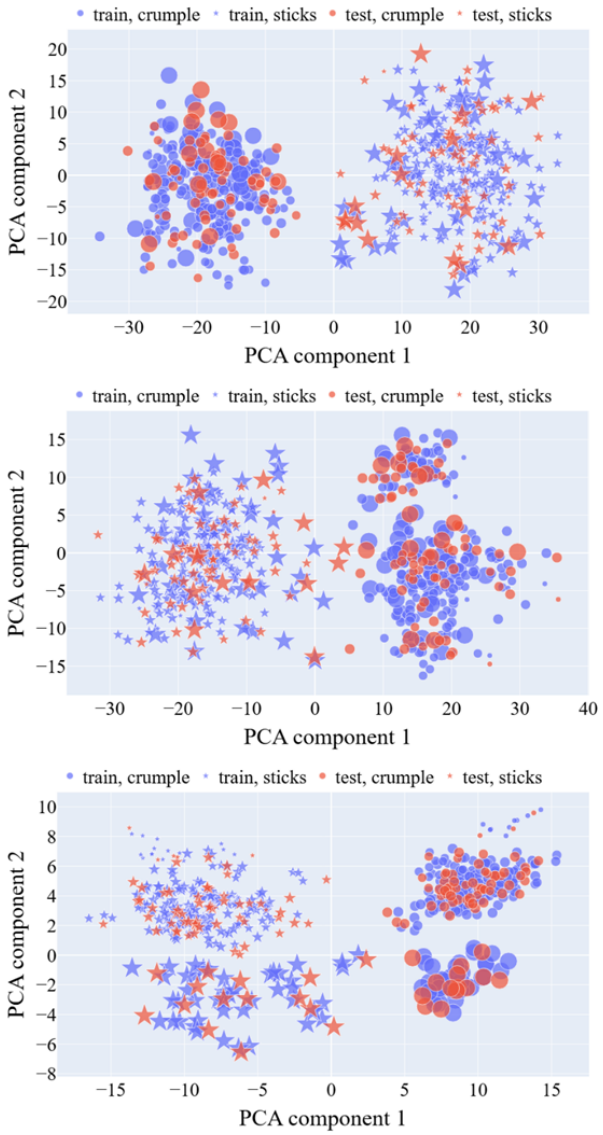

Fig. 3. PCA projected embeddings for the baseline (top), contextual SEC without metric estimation (middle), and contextual SEC (bottom). The larger the size of the data marker, the lower is the DRR of the underlying acoustic condition.

The complexity for the SEC task increases with decreasing DRR values as the signal energy of the reflected component of the sound increases. As can be observed from Fig. 2 the *baseline* model and *baseline_all_DRR_level* model have

similar performance until DRR range [0, -5) dB after which the performance, as expected, of the *baseline* model drops compared to the *baseline_all_DRR_level* model.

The *contextual* SEC models learned using the proposed strategy have a significantly lower performance drop as the DRR decreases. Even though all the different acoustic metrics that were tested in the contextual SEC learning framework (see Sec. II-C) improved the performance, the CT with 25 percentile of the energy (CT25) gave the best performance improvement. The value of λ in the model training loss (see Eq. 1) was empirically selected to be 0.25.

Overall, the performance improved in all DRR ranges. The performance variation across the validation folds also decreased considerably. Importantly, for the last DRR range ([-12,-15) dB) the performance improved by 2.9%, and performance variation across the validation folds reduced by 5.14% compared to the *baseline* model. The *contextual_CT25* model consistently outperforms the RAANN classifier model that was trained to have room invariant features in its internal representation. Note that this model was also used as a starting point for the event feature encoder used in all variants of the *contextual* SEC models.

In Fig. 3, embeddings of two example event classes (crumple and sticks), convolved with different RIR ranges, are projected using Principal Component Analysis (PCA). The visualization compares the *baseline* model with the *contextual_CT25* model. Events convolved with RIRs from similar Direct-to-Reverberant Ratio (DRR) conditions are represented by similarly sized data markers. The embedding space of the *contextual_CT25* model exhibits a different structure compared to the *baseline* model. Notably, despite no explicit mechanism enforcing order based on the CT25 acoustic metric, the embeddings appear to be organized according to their corresponding CT25 values. This structured organization suggests that interpolation (and potentially extrapolation) to unseen acoustic conditions is feasible.

IV. Conclusions and future work

Room acoustics introduce distinctive filtering effects that alter the characteristics of recorded sound, impacting SEC performance. The proposed learning scheme for contextual SEC models leverages knowledge of room acoustic conditions, enabling the SEC model to adapt its behavior based on the environment. This approach improves SEC performance across different DRR ranges and outperforms alternative methods that focus on RIR-agnostic features. The results highlight the potential of contextual SEC models in enhancing robustness to varying acoustic conditions.

References

[1] S. Upadhyaya, D. Berckmans, W. Desmet, W. Buyens, and P. Karsmakers, "Significance of having a large sound dataset for pig cough classification," in *Proc. USPLF Conference*, May. 2023, pp. 595–602.

[2] D. Emmanouilidou and H. Gamper, "The effect of room acoustics on audio event classification," in *The 23rd International Congress on Acoustics*, 09 2019.

[3] S. Upadhyaya, W. Buyens, E. Vranken, W. Desmet, and P. Karsmakers, "Assessment of data augmentation and transfer learning for making pig cough classifier robust to changing farm conditions," in *Proc. IEEE ICMLA*, 2023, pp. 952–957.

[4] L. Schmidt and N. Peters, "Device generalization with inverse contrastive loss and impulse response augmentation," in *Proc. DCASE*, 09 2023.

[5] J. Lee, D. Lee, H.-S. Choi, and K. Lee, "Room adaptive conditioning method for sound event classification in reverberant environments," in *Proc. IEEE ICASSP*, 2021, pp. 870–874.

[6] M. Yasuda, Y. Ohishi, and S. Saito, "Echo-aware adaptation of sound event localization and detection in unknown environments," in *Proc. IEEE ICASSP*, 2022, pp. 226–230.

[7] S. Upadhyaya, W. Buyens, W. Desmet, and P. Karsmakers, "Room acoustic adversarial neural network for robust sound event classification," *J. Audio Eng. Soc.*, vol. 72, no. 11, pp. 754–766, Nov. 2024.

[8] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[9] S. K. Zieliński, H. Lee, P. Antoniuk, and O. Dadan, "A comparison of human against machine-classification of spatial audio scenes in binaural recordings of music," *Applied Sciences*, vol. 10, no. 17, 2020.

[10] Y. Tan, Y. Wu, Y. Hou, X. Xu, H. Bu, S. Li, D. Botteldooren, and M. D. Plumbley, "Exploring differences between human perception and model inference in audio event recognition," *ArXiv*, vol. abs/2409.06580, 2024.

[11] D. Sundström, A. Björkman, A. Jakobsson, and F. Elvander, "Room impulse response estimation using optimal transport: Simulation-informed inference," *Proc. EUSIPCO*, pp. 276–280, 2024.

[12] R. Serizel, V. Bisot, S. Essid, and G. Richard, *Acoustic Features for Environmental Sound Analysis*. Springer International Publishing, 01 2018, pp. 71–101.

[13] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. W. Wilson, "Cnn architectures for large-scale audio classification," in *Proc. IEEE ICASSP*, 2017, pp. 131–135.

[14] J. Bradley, "Review of objective room acoustics measures and future needs," *Applied Acoustics*, vol. 72, no. 10, pp. 713–720, 2011.

[15] M. Meire and P. Karsmakers, "Comparison of deep autoencoder architectures for real-time acoustic based anomaly detection in assets," in *Proc. IEEE IDAACS*, vol. 2, 2019, pp. 786–790.

[16] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. of the Second International Conference on Language Resources and Evaluation*. Athens, Greece: ELRA, may 2000.

[17] N. Werner, "audiolabs/rir-generator: V0.2.0," 2023.

[18] Echothief, "Echothief impulse response library," http://www.echoThief.com/, accessed Feb. 12, 2024.