OPEN ACCESS Freely available online

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 (2024 Nov.).
https://doi.org/10.17743/jaes.2022.0174.

Room Acoustic Adversarial Neural Network for Robust Sound Event Classification

SREENIVASA UPADHYAYA,[1, 2, *] WIM BUYENS,[2]
(sreenivasa.upadhyaya@kuleuven.be) (wim.buyens@soundtalks.com)
WIM DESMET,[3] AND PETER KARSMAKERS[1, 4]
(wim.desmet@kuleuven.be) (peter.karsmakers@kuleuven.be)

[1] *KU Leuven, Department of Computer Science, DTAI, Kleinhoefstraat 4, B-2440 Geel, Belgium*
[2] *SoundTalks NV, Interleuvenlaan 15/c, B-3001 Leuven, Belgium*
[3] *KU Leuven, Department of Mechanical Engineering, LMSD, Celestijnenlaan 300, Leuven, Belgium*
[4] *Flanders Make, DTAI-FET, B-3001 , Leuven, Belgium*

The variation in the acoustic condition of a room presents a major hurdle in the performance robustness of sound event classification. Room impulse response characterizes the way in which a sound wave is propagated from source to receiver and the overall perceptual quality and intelligibility of the recorded sound. This study presents the Room Acoustic Adversarial Neural Network (RAANN) method that can make sound event classification more robust to changes in acoustic condition by exploiting knowledge regarding the room acoustics during learning. With RAANN, the weighted F1 score for the classification task improved by 1.54 percentage points, and the standard deviation in performance dropped from 1.74 percentage points to 1.07 percentage points for acoustic conditions that were harder than those seen during the learning phase. The Clarity Index over the first 25 ms emerged as a good metric for the acoustic estimation in the RAANN training.

0 INTRODUCTION

With the progress in data acquisition and storage technologies as well as the data analysis techniques in machine learning, continuous monitoring of surroundings using a wide array of sensors has flourished in the recent years. Microphones, the sensors behind audio capture, deliver rich data that has ample perspective to derive insightful information of the environment being monitored. Furthermore, they exhibit interesting properties. Next to being contactless and relatively cheap, they also do not need a direct line of sight. This enables the entire surroundings to be monitored with a single sensor [1]. Sound event classification (SEC) refers to the task of automatically categorizing the active sound events into defined semantic classes in an audio recording [2]. SEC algorithms are of extreme importance to derive meaningful insights from the captured sound.

SEC is applied to a variety of use cases including monitoring urban scene [3], wildlife [4], domestic sound [5, 6], machine health [7, 8], conversational human-technology interfaces [9], and animal health [1]. Even though the fundamental task of SEC remains the same, the challenges [10] differ from application to application. Certain cases using portable sound sensors, for instance, mobile phones, have near-field recordings that provide a better SNR, but the variations in the background noise characteristics are also higher. Applications in group monitoring systems like animal health have far-field recordings [1] where the SNR is generally low. Sound monitoring in tunnel-like conditions have effects of high reverberation, reducing the clarity of the events and often distorting it heavily. The presence of background noise and interference creates ambiguity in the prediction of output event class and present a potential threat in urban scene monitoring [3] and environmental sound classification [11] tasks. Background noise conditions and variations in room acoustical conditions can be broadly termed as the two major causes of performance degradation in SEC systems.

Methods to improve the noise robustness in SEC [12, 13] and automatic speech recognition systems [14–16] has been widely studied and proven to be very effective in boosting the performance. Most past research has aimed to enhance robustness to reverberation through signal processing–based solutions such as dereverberation, source separation, and beamforming.

*To whom correspondence should be addressed, email: sreenivasa.upadhyaya@kuleuven.be.

In this work, the focus is on improving the robustness of the sound event classifier to variations in room acoustic conditions. The nature of the room implies the way in which sound waves propagate from their source to the receiver. It impacts the perceptual quality of the recorded sound. The variations in the acoustic condition of a room present a major hurdle in the performance robustness of the sound event classifier [17].

Data augmentation by introducing reverberation on the audio significantly reduced the word error rate [18] and improved classification robustness of audio events [19]. In [20], convolution of the audio with a set of room impulse responses (RIRs), containing both real-world and simulated RIRs, effectively improved the equal error rate in the case of a speaker identification task. In [21], data augmentation using generative adversarial networks for robust speech recognition was employed to obtain a boost in performance. The study in [22] shows significant gains in speaker recognition performance, when the RIR-based data augmentation was used in training an auto-encoder for signal enhancement and a speaker recognition model. Although the RIR-based data augmentation improved the robustness across conditions, the knowledge of the underlying acoustic condition that generates the augmented dataset is not utilized.

Humans have the ability to discern sound features/artefacts that are linked to the environment and those that characterize a specific sound event. Their brains have evolved to analyze these subtle acoustic nuances, enabling them to compensate for it and, in turn, effectively understand the sound [23, 24]. This work aims to help the learning mechanism to let deep neural networks also have such performance, by using a training strategy similar to Domain Adversarial Neural Network (DANN) [25]. Unlike the DANN scheme, the proposed method aims to achieve model robustness using only the available training data and does not depend on the data from the new, unseen environments. Hence, the proposed approach is not to be considered as a transfer learning scheme but a method that targets feature robustness across changing acoustic conditions.

Instead of a binary classifier (source/target domain classes) in a regular DANN scheme, the authors propose to use a regression model that exploits the inherent structure in room acoustic metrics to derive room acoustic agnostic (internal) features. For instance, in situations with increased reverberation, sound events will be stretched out more and can be repeated in case of short impulsive events. A regression model might capture such structure, while a classifier scheme with classes that are encoded with ad hoc numbers might not. If such structure in the data, which is linked to the acoustic condition, is picked up by the regression model, it might also be more effectively removed from the features (while retaining the SEC accuracy).

There exist a set of "easy" metrics that quantify the effect of a room acoustic condition on a sound event. Preferably, such metrics quantify changes that have an impact on the classifier performance. When the performance for a given baseline classifier is correlated with the room acoustic metrics, this could mean that they indeed indicate significant, for classification, changes in the sound that are related to the environment, which makes it harder for the classifier to have good performance [26]. Of course, some changes caused by RIR filtering might distort the sound events too much such that sound classification performance is severely reduced [27] and might even be difficult for humans to classify.

In this study, the authors compare and evaluate the room acoustic metrics to quantify the complexity of the classification task across acoustical conditions. Furthermore, they present the Room Acoustic Adversarial Neural Network (RAANN) method that exploits knowledge regarding the RIRs applied to the input for enhanced robustness to changes in room acoustic conditions. The remainder of the paper is organized as follows. SEC. 1 describes the methods, including the metrics to describe the room acoustic conditions and the deep learning (DL) model architecture. The experimental dataset and results are discussed in SEC. 2. The conclusions are given in SEC. 3.

1 METHODS

1.1 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), Direct-to-Reverb Ratio (DRR), and Clarity Index (CI) [28], that quantify the impact of room acoustic conditions on the original sound. For this study, the popular DRR, RT60, and CI metrics were selected because they broadly indicate the properties of the RIR [26].

1.1.1 RT60

Reverberation, commonly referred to as reverb in the field of audio engineering, is the persistent presence of sound after its initial production. Reverb occurs when a sound or signal is reflected, resulting in multiple reflections accumulating and gradually diminishing as the sound is absorbed by various surfaces within the space. These surfaces can include objects like furniture, people, and the air itself. The most noticeable aspect of reverberation is when the sound source ceases, yet the reflections persist, gradually decreasing in amplitude until they fade away entirely. Reverberation can cause distortion in the audio signal, making it harder to distinguish between different events or sounds. This distortion can mask important features of the audio signal, making it more challenging for classification algorithms to accurately identify the audio events. Fig. 3 shows an example of an audio event where the spectrogram of the filtered version of the audio event is smeared out (blurred) and destroys the intricate information present in the spectrogram.

RT [29], often denoted as RT60, is a metric used to quantify the duration it takes for sound to decay within an enclosed area after the sound source has ceased. RT60 is specifically defined as the time it takes for the sound pressure level to decrease by 60 dB, measured immediately following the abrupt termination of the test signal. RT60 is commonly expressed as a single value when measured using a wide band signal covering the frequency range of 20 Hz to 20 kHz. However, since it varies with frequency, a more accurate description can be provided in terms of frequency bands (e.g., 1 octave, 1/3 octave, 1/6 octave, etc.). This metric is typically presented as a time measurement in seconds, signifying the time required for the signal to diminish by 60 dB from its original level. Measuring a 60-dB decay can be challenging, especially at lower frequencies due to the influence of ambient noise.

To simplify the measurement, it is often acceptable to measure a 20-dB drop and then multiply the time by three or measure a 30-dB drop and multiply the time by two, under the assumption of a linear decay. In this work, the point of the 30-dB drop in the energy decay curve was measured, and the corresponding time was doubled to obtain the RT60 value [30].

1.1.2 Direct-to-Reverb Ratio

The DRR is a property of the room that quantifies the balance between the initial direct sound arriving from a sound source to a listener's ears and the subsequent reverberant sound in an acoustic environment. This ratio plays a crucial role in determining the perceived clarity and intelligibility of sound within a room. A high DRR indicates that the direct sound dominates, offering clear and distinct auditory information, which is essential for tasks like speech communication or music perception. Conversely, a low DRR signifies that the reverberant sound significantly contributes to the acoustic environment, potentially leading to reduced speech intelligibility.

The DRR can be calculated from the RIR and is typically calculated using the following mathematical formula,

$$DRR = 10\log_{10}\left(\frac{P_d}{P_r}\right), \tag{1}$$

where the DRR is expressed in decibels, P_d is the power of the direct sound component, and P_r is the power of the reverberant sound component.

1.1.3 Clarity Index

The CI [31] is a property of the room and is a crucial metric for gauging the perceptual lucidity and comprehensibility of sound in an acoustic setting. It assumes a pivotal role in evaluating the efficacy of audio playback systems, especially in venues like concert halls, lecture halls, and auditoriums, where clear and easily understood communication is imperative. The CI considers various elements, including the initial direct sound, early reflections, and subsequent reverberation in the room, offering valuable insights into the degree to which speech or music can be grasped by the audience. A higher CI value denotes superior speech intelligibility and an enriched auditory experience, while a lower value implies diminished clarity and the potential for difficulties in comprehending spoken words or musical intricacies. As a result, the CI proves indispensable in both the planning and assessment of acoustic environments, contributing to the optimization of room acoustics across a spectrum of applications and ensuring that the intended auditory content is conveyed with the desired clarity and faithfulness.

Formula for calculating the CI is as follows,

$$CI = 10\log_{10}\left(\frac{L_p}{L_s}\right), \tag{2}$$

where L_p is the level of the direct sound arriving at the listener's ears in the first T ms and L_s is the level of the late sound or the sound that arrives after a certain time delay (typically around T ms) due to reflections and reverberation in the room. In automatic speech recognition systems, the CI metric, with T as 50 ms ($C50$), was the room acoustics representation that correlated most strongly with speech recognition performance [32]. However, the optimal choice of T could depend on the use case.

1.2 Room Acoustic Adversarial Neural Network

Conventionally trained models have poor scaling when trained on a dataset with a certain set of acoustical conditions and tested on the dataset with distributional shift (e.g., caused by a change in acoustic conditions). To achieve the goal of acoustic condition invariant classifier, the authors introduce the RAANN training scheme, which infuses the knowledge of underlying acoustic properties corresponding to the input audio events into the learning process. As a consequence, the classifier based on these features is expected to be more robust to changes in room acoustics.

For this purpose, ideas from DANN are reused. DANNs aim at removing the data shift observed in some target domain when compared to a source domain by an adversarial interplay between as SEC and source/target domain binary classifier. Note that for the target domain, typically no class labels are available. The SEC and domain classifier operate on the same learnable set of features. Features are learned such that a) the SEC has the highest classification performance on the source data for which labels are present and b) the domain classifier classification performance is as bad as possible, hence aiming to have one that fails to discriminate based on the features from which domain (room) the sound originated.

In the proposed RAANN, the discriminator is replaced by a regression function, which, based on the features, estimates room metrics [in this work, this will be one of the RT60, DRR, or CI over 25 ms (C25)]. While DANN is used in a transfer learning setting, RAANN is not. The input to the RAANN is a dataset where, for each sound example, the class label and room acoustic metrics are known. Based on this information, RAANN searches for a set of features that optimize the SEC performance while minimizing the predictive performance of the regression function that estimates the room acoustic metrics. The latter indicates

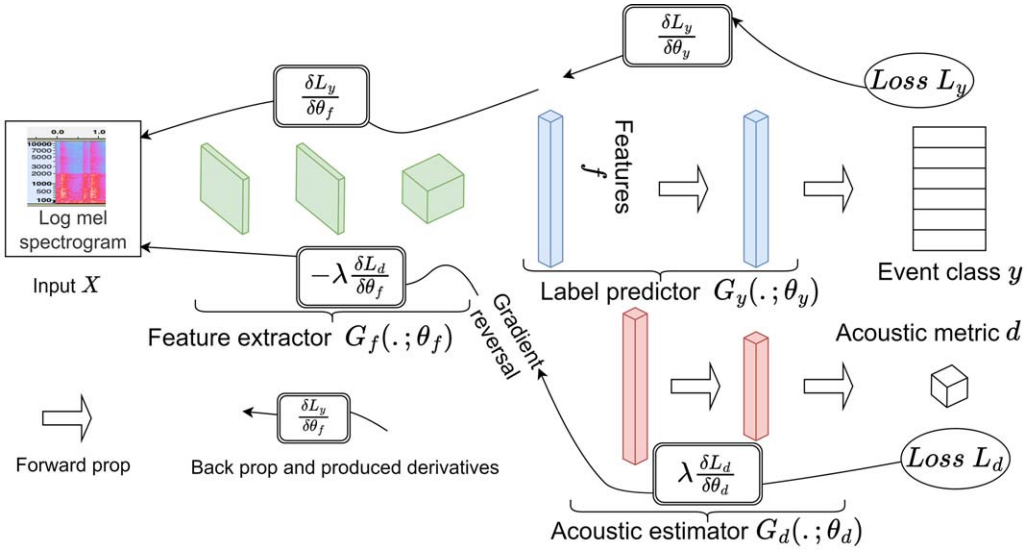

Fig. 1. RAANN learning methodology.

that room-specific information is removed from the features (see Fig. 1). Unlike DANN, RAANN does not require any target domain data. The RAANN block is realized as a multilayer perceptron–based regression network. The regression network accepts the flattened features from the feature extractor block and has three hidden dense layers of 512 neurons each. The output layers is comprised of a single neuron with linear activation corresponding to the estimated room acoustic metric.

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

The three network parameters are optimized based on the following objective:

$$\min_{G_f, G_y, G_d} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} \mathbf{y}_{ij} \, log(G_y\left(G_f\left(\mathbf{X}_i\right)\right)) - \alpha \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} |\mathbf{d}_{ij} - \max\left(-d_{tr}, \min\left(1 + d_{tr}, G_d\left(G_f\left(\mathbf{X}_i\right)\right)\right)\right)|, \quad (3)$$

where G_f, G_y, and G_d are the feature extraction, the classifier, and regression models, respectively. The left term corresponds to the classification loss (L_y) and the right term corresponds to the regression loss (L_d). Categorical cross-entropy and mean absolute error (MAE) are used as the classification task loss and regression loss, respectively. MAE was selected because it reduces the impact of outlier samples [33] (see Appendix A.1). The trade-off parameter α balances the importance of both losses. The constant parameter d_{tr} is introduced to limit the estimated acoustic parameter. Setting d_{tr} to ∞ disables the constraining. This parameter enables to control the learning process by limiting the contribution of the outlier samples with large errors (see Appendix A.4).

1.3 Evaluation Criteria

Typically, the performance of the audio event classification system is measured in terms of accuracy, precision, recall, and F1 score [34]. These metrics are derived from the values of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) [see Eq. (4)]. Precision gauges the accuracy of the positive predictions, while recall assesses the model's capacity to capture all pertinent instances of the positive class. F1 score amalgamates precision and recall into a singular score, offering a well-balanced assessment of a model's overall performance. An elevated F1 score signifies a model excelling in both precision and recall, providing a balance between accurately identifying positive instances and minimizing false predictions.

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\ Pr &= \frac{TP}{TP+FP} \\ Re &= \frac{TP}{TP+FN} \\ F_1 &= \frac{2\ Pr\ Re}{Pr+Re}. \end{aligned} \quad (4)$$

However, in case of multiclass classification, the weighted average F1 score [35] is a performance metric commonly employed, especially when dealing with imbalanced datasets. Unlike traditional F_1 score computation, which treats each class equally during averaging, the weighted $F_1^{(w)}$ score considers the varying class sizes by assigning weights w_i based on the number of instances for each class [see Eq. (5)]. This approach computes F1 score for each label and returns the average, considering the proportion for each label in the dataset.

$$F_1^{(w)} = \sum_{i=1}^{c} w_i (F_1)_i, \quad (5)$$

where $(F_1)_i$ represents the F_1 score when class i is put against a single group composed out of all other classes and w_i is the proportion of examples for class i in the dataset.

2 EXPERIMENTS AND RESULTS

2.1 Original Dataset

The audio events used for this study were taken from the Real World Computing Partnership (RWCP) dataset [36]. The dataset contains nonspeech sounds recorded in an anechoic room. The anechoic nature of the audio events makes them ideal for simulating different acoustical conditions on the recordings. From RWCP dataset, 80 events from each of the 50 preselected sound event classes were used. Sound event classes were selected as in [37]. Each sound event was recorded at a sampling frequency of 16 kHz, and the event length was 1 s. This dataset is named as the original (ORIG) dataset because it contains unmodified events from the RWCP dataset. Hence, the ORIG dataset contains the data from anechoic conditions. Subsequently, the original data was augmented to create multiple modified versions by convolving it with a wide range of RIRs that span the desired ranges of the room acoustics.

Different sources of RIRs were used in the augmentation process, and they are described below.

2.1.1 Python RIR–Simulated RIRs

These RIRs are simulated using the Python RIR [38] generator utility. This is a Python-based [39] RIR generator package, which generates RIRs for a specified set of configuration parameters based on the image source method [40]. The key configuration parameters that can be modified in Python RIR are room dimension, sound source position, receiver position, and target RT60 values. The receiver position was fixed at (5, 5, 2) in a room with length, width, and height as (8, 8, 4) m, respectively. The source position was sampled within the room randomly to obtain the RIRs with target acoustic metrics. The package uses the image-source method for RIR generation. The maximum possible order was used to achieve the desired RT60 value specified in the input configuration parameters. The receiver is assumed to be omnidirectional, and the surfaces are uniformly sampled as per the Sabine-Franklin's formula. The reflection coefficients of the surrounding walls are applied broadband. A total of 40,000 RIRs were generated in total, covering different combinations of the above parameters to get a wide variety of RIRs.

2.1.2 Echo Thief RIRs

Echo Thief (ET) [41] 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 the simulated ones, these RIRs include the effect of real acoustic spaces with different materials and interiors, which brings in more diversity in the data. A total of 74 RIRs were collected.

Table 1. Dataset to DRR mapping in ET RIR dataset.

Dataset Name	DRR (dB) [Min to Max)	Number of RIRs
ET_0	[−5, 14)	30
ET_1	[−15, −5)	44

2.2 Generated Datasets

The authors' objective is to evaluate the robustness of the RAANN method when it is used in acoustic conditions that are more challenging compared with those from the training set. For this purpose, four training and test set combinations were generated in which the test set always contains more challenging conditions compared with the training set.

Firstly, the ET set of RIRs was split in two parts as is given in Table 1: the subset ET_0, which has RIRs that have DRR values in the interval [−5, 14), and subset ET_1, which has more challenging DRR values in the interval [−15, −5). The split could have been on any of the three room acoustic metrics. In preliminary experiments (given in Appendix A.2), it was seen that RT60, DRR, and C25 all correlated well with the classifier classification performance.

Secondly, the data in ORIG was split in 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 three 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 an RIR that a) for training, is sampled from the Python RIR–simulated RIRs (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 test are generated.

Note that events present in the training set will not be present in the test set. However, the acoustic conditions represented in ET_0 are available both in the training and test set. The acoustic conditions from ET_1 are only available in the test set (and, hence, were not seen during training). In Fig. 2, two scatter plots indicate which acoustic conditions are available in the test set. The first scatter plot shows the DRR value in function of RT60 for each RIR. The second scatter plot gives the C25 in function of RT60. The points in orange asterisks represent the conditions that are available in both training and testing from the ET RIRs. The blue triangle points represent the conditions only present during testing. The dots in green represent the conditions from the simulated RIRs that were part of the training set. The overlap between the metrics of real RIRs and simulated RIRs is limited by the simulation setup that was used to generate the RIRs.

2.3 Input Representation

The input representation used is a log mel-spectrogram [31]. Time domain audio signals are transformed to the frequency domain using the short-time Fourier transform. Subsequently, it is converted by several mel filter banks to a lower dimensional mel representation. Mel is a perceptual scale for the audio frequencies. Mel spectral features

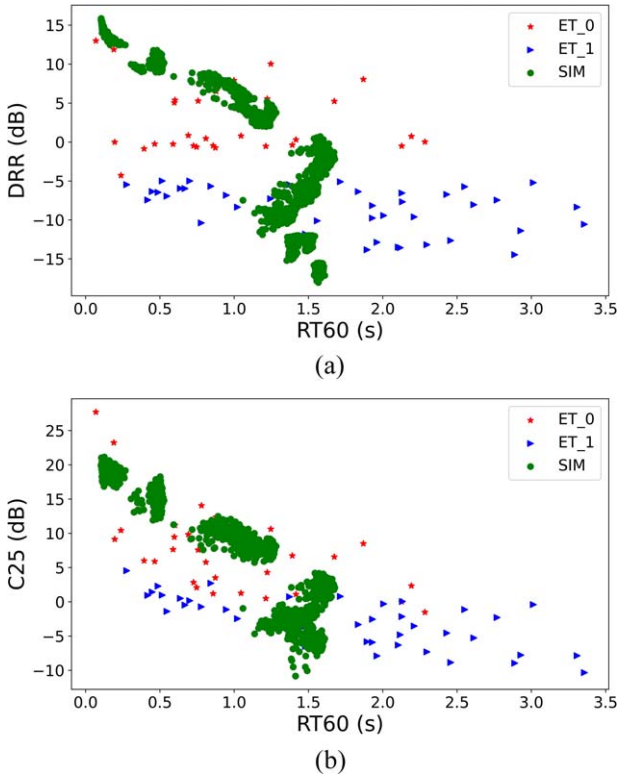

Fig. 2. Room acoustic metric distribution of conditions used in ET and SIM datasets: RT60 vs. DRR (a) and RT60 vs. C25 (b).

are a popular choice for input representation in DL-based acoustic event detection [42].

Conversion from time domain audio signal to mel-spectrogram was done using the mel-spectrogram function in librosa [43]. Mel-spectrogram is converted to the log domain using the *power_to_db* function. Fmin, the lowest frequency to be considered to derive the mel spectral bins, was set to 50 Hz to avoid very low frequency bias due to the ambient noise in the audio files. Short-time Fourier transform window length was set to 512 samples (32 ms) with a hop length of 160 samples (10 ms). The number of mel-spectrogram bins was set to 64. This implies that every 10 ms of raw audio samples, a mel-spectrogram of length 64 is generated. Each raw audio event is converted to a mel-spectrogram of size 100×64, where 100 represents the number of time frames (of 10 ms) and 64 corresponds to the number of mel frequency bins. Fig. 3 shows an audio event and the log mel-spectrograms of the original event and its filtered version.

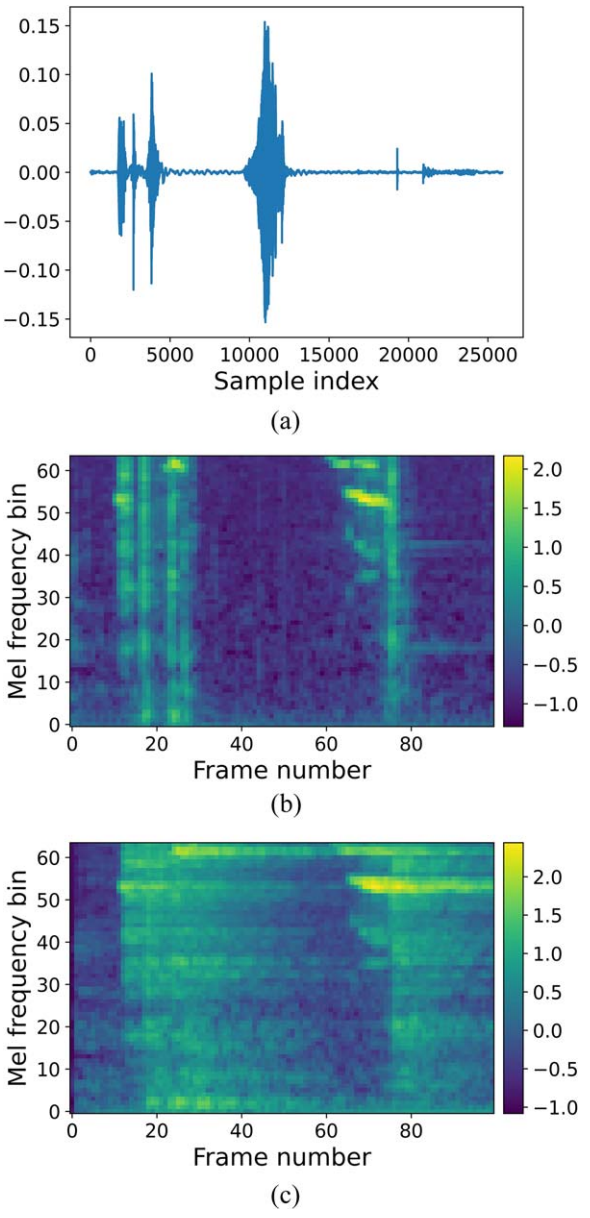

Fig. 3. Sample representation of an audio event: Time domain (a), log mel-spectrogram of the audio event (b) and log mel-spectrogram of the event filtered by an RIR (c).

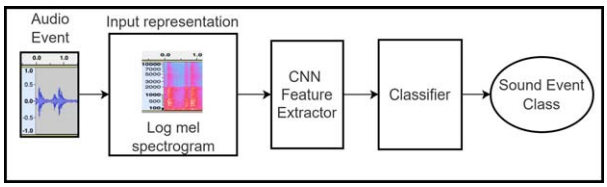

Fig. 4. CNN-based SEC model.

2.4 SEC Processing Pipeline

The model architecture employed in this study is outlined in Fig. 4. From the log mel-spectrogram, a feature extractor, based on a pretrained Convolutional Neural Network (CNN), calculates meaningful feature maps. These feature maps are then input into the classifier block, implemented using a set of fully connected layers and a softmax layer, to make the output label predictions.

The core of this model is a VGGish [44] CNN feature extractor model for sound event detection. VGGish is used as a feature extractor, which extracts semantically meaningful 128-dimensional embeddings. The VGGish network is pretrained on the YouTube audio dataset [45]. This is fed to a classification network comprised of two dense layers with 512 neurons. Finally, an output dense layer with 50

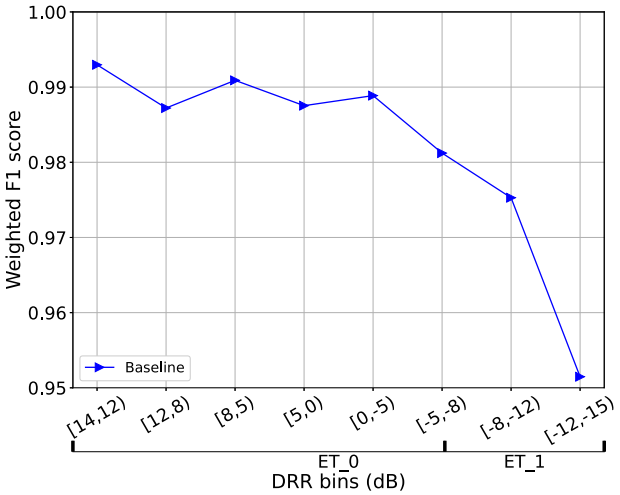

Fig. 5. Weighted F1 score for the baseline SEC model across different DRR conditions available in the ET test set, comprising of ET_0 (available during training and testing) and ET_1 (available only during testing).

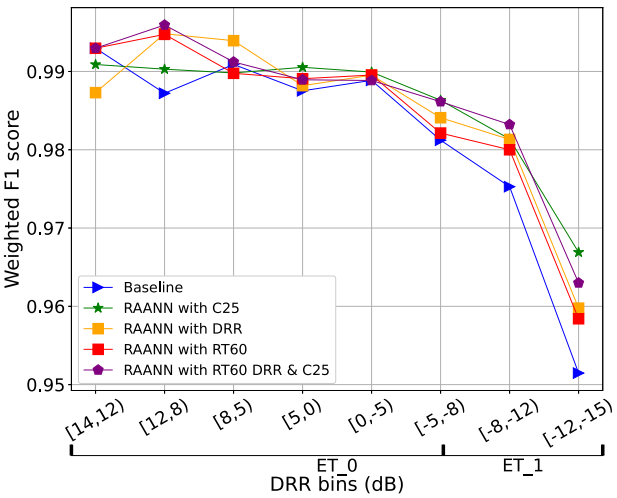

Fig. 6. Performance comparison of RAANN training with various room acoustic metrics.

neurons, one for each class, with SoftMax [46] activation gives the output of the classification task. This model will be referred to as the baseline model in the remainder of the work.

2.5 Model Training

The VGGish feature extractor is finetuned in the training process. The SEC model is trained with an Adam optimizer, with an initial learning rate of $1e-2$. The training is done on mini batches of size 64. The model is trained for 50 epochs with categorical cross entropy as the loss function and Adam as the optimizer. The training and experimentation are done on a server with AMD EPYC 7402P processor and NVIDIA GeForce RTX 2080 Ti GPUs. The models were implemented in Python3 using the TensorFlow v2 Keras DL framework [47].

2.6 Baseline Model Results

Using the four training and test set combinations (as described in SEC. 2.2), a baseline model architecture (as described in SEC. 2.4) was four times trained in a traditional way (as described in SEC. 2.5) and tested. In Fig. 5, the average performance of the models on the independent test set was visualized across DRR bins. The complexity of the acoustic environment increases as one moves from left to right on the DRR bin axis. The results suggest that a more complex acoustic environment makes it harder for the classifier to perform classification. Considering the model performance across DRR bins, the weighted F1 score stays closer to the 99% mark until the DRR values of -5 dB. This is expected because these conditions were also seen during training. After this, a decline in performance is observed. Overall, there is a decrease in F1 score of about 4 percentage points moving from the highest DRR condition to the lowest DRR condition.

2.7 RAANN

In this paper, the authors aim to bridge the gap in performance when using the model in conditions that are different (more challenging) from those used in training. For this purpose, they use the proposed RAANN training that exploits knowledge of the acoustic environment that generated the training data during training.

The RAANN model training (see SEC. 1.2) can be done with an acoustic parameter of choice that can be derived from the available data. In this work, the model training was explored with RT60, DRR, and C25 parameters. In Fig. 6, average classifier performance across the four training and test set combinations is given for the baseline and the three RAANN alternatives. Although the performance gain compared with the baseline was observed with all the three room acoustic metrics, C25-based RAANN provided the maximum improvement. The C25 was chosen instead of C50 and C80 because C25 provided better results empirically. In Fig. 7, the RAANN C25 alternative is compared with the baseline by showing the box plot representations of the classifier performance across different runs.

RAANN aims to generalize the features and make it more invariant to changes in the underlying acoustic conditions. At milder acoustic conditions represented by values of DRR more than 0 dB, the weighted F1 score is closer to the baseline performance. But as the DRR decreases and the complexity of the acoustics increases, it can be seen that the proposed model outperforms the baseline and the decline in the performance is lower. The weighted F1 score improved by 1.54 percentage points for the last bin corresponding to DRR values between –12 and –15 dB. It is also observed that in this most challenging condition, the standard deviation for RAANN is much less (1.07 percentage points) compared with the baseline (1.74 percentage points).

In RAANN, it is proposed to create an internal feature representation that makes it difficult to estimate the room acoustic metrics of the room in which the sound was

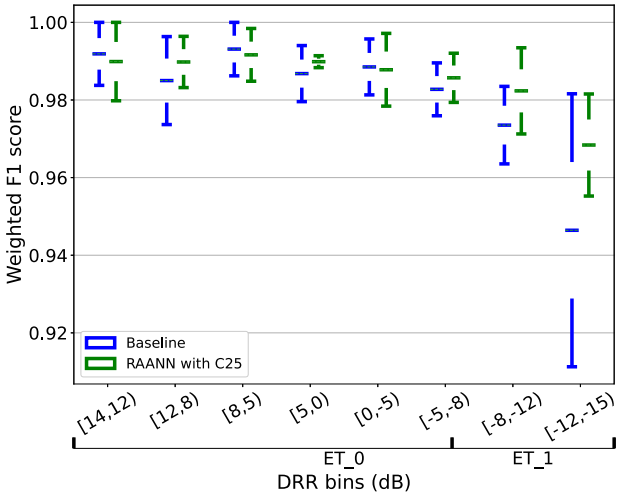

Fig. 7. Comparison of the distribution of weighted F1 scores for the baseline (left/blue) and RAANN (right/green) model across various DRR conditions from the ET test dataset. In each distribution, the median is indicated by the middle dash, and the range (minimum and maximum) is indicated by the bottom and top dashes, respectively.

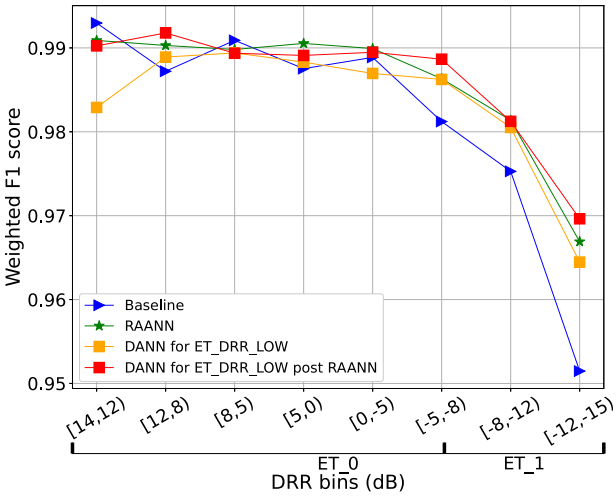

Fig. 8. Weighted F1 scores for the DANN training post RAANN training across various DRR conditions on the ET test dataset.

recorded. Instead of estimating room acoustic metric values, all different rooms can get assigned a room code. In this case, the room acoustic metric estimator can be replaced by a classifier that, based on the internal representation tries to classify the room code. In Appendix A.3, RAANN was found to outperform the latter approach.

2.8 Domain Adaptation Post RAANN

Domain adaptation is the procedure of modifying a predictive model, initially trained on a source domain, so that it can effectively function in a target domain despite potential disparities in data distribution between the two domains. In this section, RAANN is compared with a domain adaptation scheme. The domain adaptation is realized using DANN training with ET_1 as the target domain and unlabeled input samples from the target domain are used during training. The labelled source domain dataset used is the same as the one used for training the RAANN. Hence, when DANN is applied, knowledge about the most challenging room acoustic in ET_1 is available, whereas it was not available during RAANN training.

DANN training was performed with and without RAANN training as a pretraining step, and the results are compared in Fig. 8. The performance gain when DANN is applied after the RAANN training is very marginal compared with the gain with domain adaptation without the RAANN pretraining. This is expected because the RAANN pretrained model is already trained to be less dependent on the acoustic conditions and, hence, the impact of the domain adaptation is almost negligible. This observation strengthens the case as how the RAANN can be used to achieve generalization of the model.

3 CONCLUSION

The presence of reverberation and filtering effects due to the room acoustics give a signature modification to the characteristics of the recorded sound. The nature and amount of distortion introduced depend on the parameters of the room like dimensions, construction materials, and interiors as well as the distance of the source from the receiver. These effects cause significant challenges for the problem of SEC. The authors demonstrated the effects of the RIRs on an audio event classification scenario.

In this work, RAANN is proposed, wherein the knowledge of the room acoustics is utilized to improve the generalization of the feature extractor across different acoustic conditions. When RAANN was applied to more challenging acoustic conditions compared with those used in training, the overall weighted F1 score improved by 1.54 percentage points, and standard deviation reduced by 0.67 percentage points compared with the baseline.

4 ACKNOWLEDGMENT

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

5 REFERENCES

[1] S. Upadhyaya, D. Berckmans, W. Desmet, et al., “Significance of Having a Large Sound Dataset for Pig Cough Classification,” in *Proceedings of the 2nd U.S. Precision Livestock Farming Conference*, pp. 595–602 (Knoxville, TN) (2023 May).

[2] S. Adavanne, H. M. Fayek, and V. Tourbabin, “Sound Event Classification and Detection With Weakly Labeled Data,” in *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events*, pp. 15–19 (New York, NY) (2019 Oct.). https://doi.org/10.33682/fx8n-cm43.

[3] A. F. R. Nogueira, H. S. Oliveira, J. J. M. Machado, and J. M. R. S. Tavares, "Transformers for Urban Sound Classification—A Comprehensive Performance Evaluation," *Sensors*, vol. 22, no. 22, paper 8874 (2022 Nov.). https://doi.org/10.3390/s22228874.

[4] L. Nanni, G. Maguolo, and M. Paci, "Data Augmentation Approaches for Improving Animal Audio Classification," *Ecol. Inform.*, vol. 57, paper 101084 (2020 May). https://doi.org/10.1016/j.ecoinf.2020.101084.

[5] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound Event Detection in Domestic Environments With Weakly Labeled Data and Soundscape Synthesis," in *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events*, pp. 253–257 (New York, NY) (2019 Oct.). https://doi.org/10.33682/006b-jx26.

[6] G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, and P. Karsmakers, "DCASE 2018 Challenge - Task 5: Monitoring of Domestic Activities Based on Multi-Channel Acoustics," Tech. Rep. (2018 Jul.). https://doi.org/10.48550/arXiv.1807.11246.

[7] E. Brusa, C. Delprete, and L. G. Di Maggio, "Deep Transfer Learning for Machine Diagnosis: From Sound and Music Recognition to Bearing Fault Detection," *Appl. Sci.*, vol. 11, no. 24, paper 11663 (2021 Dec.). https://doi.org/10.3390/app112411663.

[8] S. Ding, S. Zhang, and C. Yang, "Machine Tool Fault Classification Diagnosis Based on Audio Parameters," *Results Eng.*, vol. 19, paper 101308 (2023 Sep.). https://doi.org/10.1016/j.rineng.2023.101308.

[9] S. Ö. Arik, M. Kliegl, R. Child, et al., "Convolutional Recurrent Neural Networks for Small-Footprint Keyword Spotting," in *Proceedings of the Interspeech*, pp. 1606–1610 (Stockholm, Sweden) (2017 Mar.). https://doi.org/10.21437/Interspeech.2017-1737.

[10] T. N. T. Nguyen, K. N. Watcharasupat, Z. J. Lee, et al., "What Makes Sound Event Localization and Detection Difficult? Insights From Error Analysis," in *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events*, pp. 120–124 (Online) (2021 Nov.).

[11] W. Mu, B. Yin, X. Huang, J. Xu, and Z. Du, "Environmental Sound Classification Using Temporal-Frequency Attention Based Convolutional Neural Network," *Sci. Rep.*, vol. 11, paper 21552 (2021 Nov.). https://doi.org/10.1038/s41598-021-01045-4.

[12] Y. Jeong, J. Kim, D. Kim, J. Kim, and K. Lee, "Methods for Improving Deep Learning-Based Cardiac Auscultation Accuracy: Data Augmentation and Data Generalization," *Appl. Sci.*, vol. 11, no. 10, paper 4544 (2021 May). https://doi.org/10.3390/app11104544.

[13] J. Salamon and J. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283 (2017 Mar.). https://doi.org/10.1109/LSP.2017.2657381.

[14] V. Kadyan, P. Bawa, and T. Hasija, "In Domain Training Data Augmentation on Noise Robust Punjabi Children Speech Recognition," *J. Ambient Intell. Humaniz. Comput.*, vol. 13, pp. 2705–2721 (2022 May). https://doi.org/10.1007/s12652-021-03468-3.

[15] H. K. Kathania, S. R. Kadiri, P. Alku, and M. Kurimo, "Using Data Augmentation and Time-Scale Modification to Improve ASR of Children's Speech in Noisy Environments," *Appl. Sci.*, vol. 11, no. 18, paper 8420 (2021 Sep.). https://doi.org/10.3390/app11188420.

[16] A. Pervaiz, F. Hussain, H. Israr, et al., "Incorporating Noise Robustness in Speech Command Recognition by Noise Augmentation of Training Data," *Sensors*, vol. 20, no. 8, paper 2326 (2020 Apr.). https://doi.org/10.3390/s20082326.

[17] D. Emmanouilidou and H. Gamper, "The Effect of Room Acoustics on Audio Event Classification," in *Proceedings of the 23rd International Congress on Acoustics*, pp. 102–109 (Aachen, Germany) (2019 Sep.). https://doi.org/10.18154/RWTH-CONV-239986.

[18] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224 (New Orleans, LA) (2017 Mar.). https://doi.org/10.1109/ICASSP.2017.7953152.

[19] 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 *Proceedings of the International Conference on Machine Learning and Applications (ICMLA)*, pp. 952–957 (Jacksonville, FL) (2023 Dec.). https://doi.org/10.1109/ICMLA58977.2023.00141.

[20] S. Wang, Y. Yang, Z. Wu, Y. Qian, and K. Yu, "Data Augmentation Using Deep Generative Models for Embedding Based Speaker Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2598–2609 (2020 Aug.). https://doi.org/10.1109/TASLP.2020.3016498.

[21] Y. Qian, H. Hu, and T. Tan, "Data Augmentation Using Generative Adversarial Networks for Robust Speech Recognition," *Speech Commun.*, vol. 114, pp. 1–9 (2019 Nov.). https://doi.org/10.1016/j.specom.2019.08.006.

[22] O. Novotný, O. Plchot, O. Glembek, J. H. Černocký, and L. Burget, "Analysis of DNN Speech Signal Enhancement for Robust Speaker Recognition," vol. 58, pp. 403–421 (2019 Nov.). https://doi.org/10.1016/j.csl.2019.06.004.

[23] J. C. Middlebrooks, "Sound Localization," in M. J. Aminoff, F. Boller, and D. F. Swaab (Eds.), *The Human Auditory System: Fundamental Organization and Clinical Disorders*, Handbook of Clinical Neurology, vol. 129, pp. 99–116 (Elsevier, Amsterdam, The Netherlands, 2015). https://doi.org/10.1016/B978-0-444-62630-1.00006-8.

[24] A. Neidhardt, C. Schneiderwind, and F. Klein, "Perceptual Matching of Room Acoustics for Auditory Augmented Reality in Small Rooms - Literature Review and Theoretical Framework," *Trends Hear.*, vol. 26, paper 233121652210929 (2022 May). https://doi.org/10.1177/23312165221092919.

[25] Y. Ganin, E. Ustinova, H. Ajakan, et al., "Domain-Adversarial Training of Neural Networks," *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35 (2016 Apr.).

[26] A. Tsilfidis, I. Mporas, J. Mourjopoulos, and N. Fakotakis, "Automatic Speech Recognition Performance in Different Room Acoustic Environments With and Without Dereverberation Preprocessing," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 380–395 (2013 Jan.). https://doi.org/10.1016/j.csl.2012.07.004.

[27] R. Petrick, K. Lohde, M. Wolff, and R. Hoffmann, "The Harming Part of Room Acoustics in Automatic Speech Recognition," in *Proceedings of the Interspeech*, pp. 1094–1097 (Antwerp, Belgium) (2007 Aug.). https://doi.org/10.21437/Interspeech.2007-112.

[28] N. Kaplanis, S. Bech, T. Lokki, T. van Waterschoot, and S. H. Jensen, "Perception and Preference of Reverberation in Small Listening Rooms for Multi-Loudspeaker Reproduction." *J. Acoust. Soc. Am.*, vol. 146, no. 5, pp. 3562–3576 (2019 Nov.). https://doi.org/10.1121/1.5135582.

[29] P. S. López, P. Callens, and M. Cernak, "A Universal Deep Room Acoustics Estimator," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 356–360 (New Paltz, NY) (2021 Oct.). https://doi.org/10.1109/WASPAA52581.2021.9632738.

[30] ISO, "Acoustics-Measurement of Sound Absorption in a Reverberation Room," *Standard 354:2003* (2003 May).

[31] S. S. Stevens, J. E. Volkmann, and E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *J. Acoust. Soc. Am.*, vol. 8, pp. 185–190 (1937 Jan.).

[32] P. Peso Parada, D. Sharma, P. A. Naylor, and T. van Waterschoot, "Reverberant Speech Recognition Exploiting *ClarityIndex* Estimation," *EURASIP J. Adv. Signal Process.*, vol. 2015, paper 54 (2015 Jul.). https://doi.org/10.1186/s13634-015-0237-7.

[33] A. Jadon, A. Patil, and S. Jadon, "A Comprehensive Survey of Regression Based Loss Functions for Time Series Forecasting," *arXiv preprint arXiv:2211.02989* (2022 Nov.). https://doi.org/10.48550/arXiv.2211.02989.

[34] J. Lever, M. Krzywinski, and N. Altman, "Classification Evaluation," *Nat. Methods*, vol. 13, pp. 603–604 (2016 Aug.). https://doi.org/10.1038/nmeth.3945.

[35] C. Mondal, M. K. Hasan, M. T. Jawad, et al., "Acute Lymphoblastic Leukemia Detection From Microscopic Images Using Weighted Ensemble of Convolutional Neural Networks," *arXiv preprint arXiv:2105:03995* (2021 May). https://doi.org/10.48550/arXiv.2105.03995.

[36] 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 *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*, paper 356 (Athens, Greece) (2000 May).

[37] J. Dennis, H. D. Tran, and E. S. Chng, "Image Feature Representation of the Subband Power Distribution for Robust Sound Event Classification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 2, pp. 367–377 (2013 Feb.). https://doi.org/10.1109/TASL.2012.2226160.

[38] N. Werner, "audiolabs/rir-generator: Version 0.2.0," *Zenodo* (2023 May). https://doi.org/10.5281/zenodo.4133077.

[39] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).

[40] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950 (1979 Apr.). https://doi.org/10.1121/1.382599.

[41] EchoThief, "EchoThief Impulse Response Library," http://www.echoThief.com/ (accessed Feb. 12, 2024).

[42] R. Serizel, V. Bisot, S. Essid, and G. Richard, "Acoustic Features for Environmental Sound Analysis," in T. Virtanen, M. D. Plumbley, and D. Ellis (Eds.), *Computational Analysis of Sound Scenes and Events*, pp. 71–101 (Springer, Cham, Switzerland, 2018). https://doi.org/10.1007/978-3-319-63450-0_4.

[43] B. McFee, C. Raffel, D. Liang, et al., "librosa: Audio and Music Signal Analysis in Python," in *Proceedings of the Python in Science Conference*, pp. 18–24 (Austin, TX) (2015 Jul.). https://doi.org/10.25080/Majora-7b98e3ed-003.

[44] S. Hershey, S. Chaudhuri, D. P. W. Ellis, et al., "CNN Architectures for Large-Scale Audio Classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135 (New Orleans, LA) (2017 Mar.). https://doi.org/10.1109/ICASSP.2017.7952132.

[45] S. Abu-El-Haija, N. Kothari, J. Lee, et al., "YouTube-8M: A Large-Scale Video Classification Benchmark," *arXiv preprint arXiv:1609:08675* (2016 Sep.). https://doi.org/10.48550/arXiv.1609.08675.

[46] B. Gao and L. Pavel, "On the Properties of the Softmax Function With Application in Game Theory and Reinforcement Learning," *arXiv preprint arXiv:1704:00805* (2017 Apr.). https://doi.org/10.48550/arXiv.1704.00805.

[47] F. Chollet, et al., "Keras," https://github.com/fchollet/keras (2015).

A.1 RAANN WITH DIFFERENT LOSS FUNCTIONS FOR ACOUSTIC ESTIMATOR

The estimation of room acoustic metrics is realized as a regression problem. In this context, a range of loss functions are possible to guide the model training. Mean square error and MAE are compared in Fig. 9. MAE gives equal weight to all errors, regardless of their magnitude, and, hence, it reduces the impact of outlier samples. With MAE, the model performs better.

A.2 ACOUSTIC METRIC COMPARISON FOR CLASSIFICATION COMPLEXITY

To understand the relation between an acoustic metric and the complexity of the classification task, the authors created ten datasets (derived from the ORIG dataset) for each of the acoustic metrics, namely RT60, DRR, and

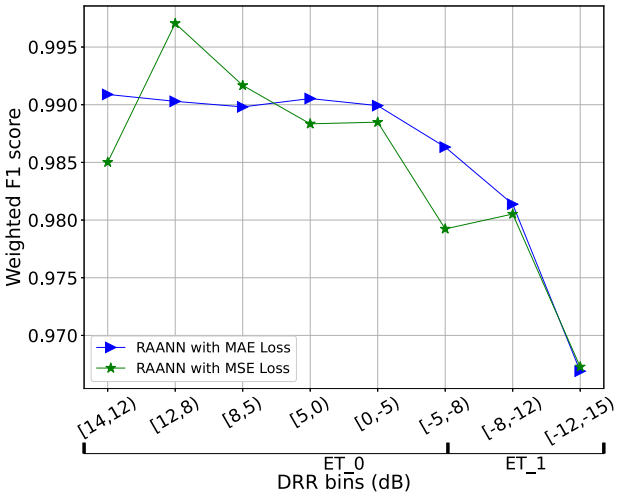

Fig. 9. RAANN performance with different loss functions for room acoustic metric estimator, across various DRR conditions on the ET test dataset.

C25, and the datasets are numbered and ordered according to increasing complexity of the acoustic condition. The 40,000 SIM RIRs are ordered according to their increasing complexity based on the metric. Ten groups of RIRs are formed from the ordered RIRs, where the first group has the lowest 4,000 RIRs, the second group has the next 4,000 RIRs, and so on. This grouping ensured that each group contains an equal number of RIRs contributing to the diversity of RIRs in a given value range of an acoustical metric.

The ten datasets for each of the acoustical metrics are derived by convolving the ORIG dataset with the previously created RIR groups. Acoustic condition specific models are trained with each one of the ten datasets. Fig. 10 shows the performance of these models on the test set of the ten datasets for all the three acoustic metrics. The models corresponding to the lower complexity of the RIR have a sharp drop in performance as the complexity increases. The models corresponding to the higher-complexity RIRs tend to perform better across conditions. This observation is very similar in RT60, DRR, and C25.

A.3 ACOUSTIC ESTIMATOR AS A MULTI-DOMAIN CLASSIFIER

By realizing the acoustic estimator as a multi-domain classifier, the authors assign a fixed output class to an input sample depending on the dataset where it belongs. Thus, the acoustic estimator becomes a classifier, and the task becomes like a domain classification task. The performance of this model in comparison with RAANN is shown in Fig. 11. This experiment reveals the advantage of RAANN and how the estimation of the acoustic metric facilitates the model to adapt across a wide range of conditions, unlike the multi-domain classifier implementation.

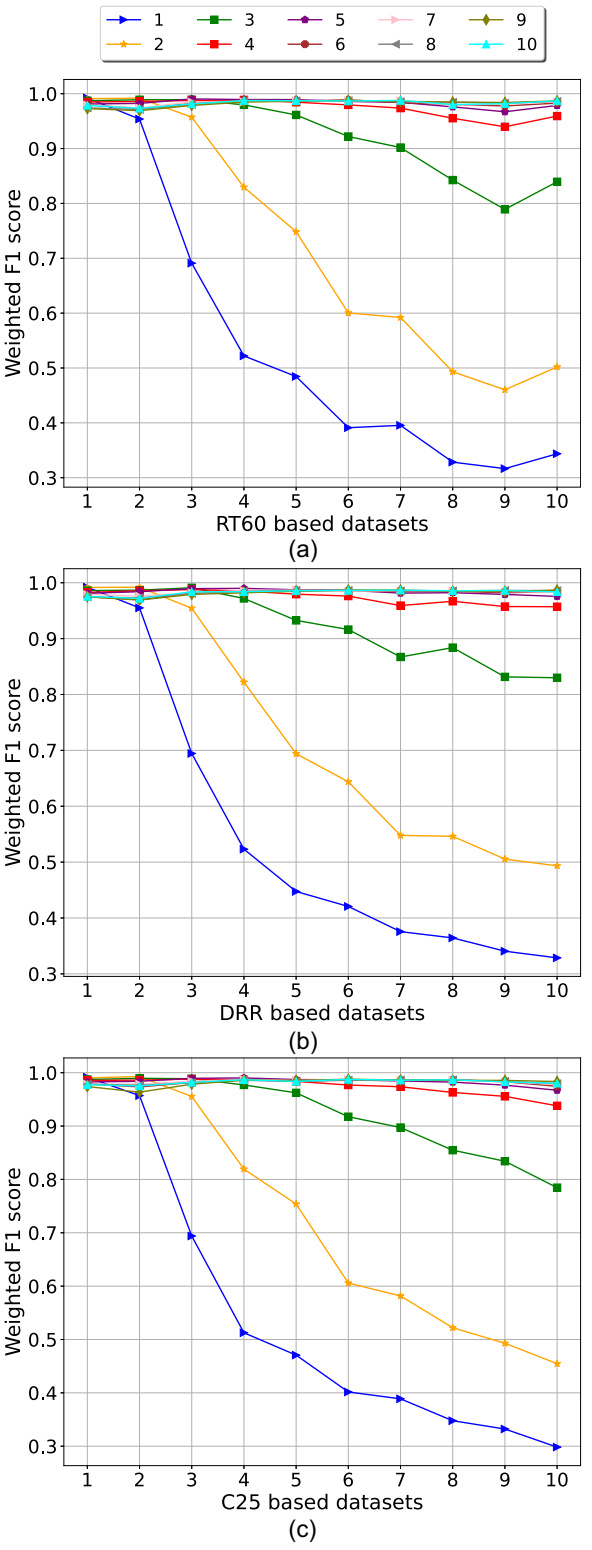

Fig. 10. Performance of group-specific models across datasets formed with groups of RT60 (a), DRR (b), and C25 (c).

A.4 CONSTRAINING THE ACOUSTIC METRIC ESTIMATION IN RAANN

The values of the target acoustic metric is min-max normalized to values between 0 and 1. The upper and lower limits represent the maximum and the minimum values of

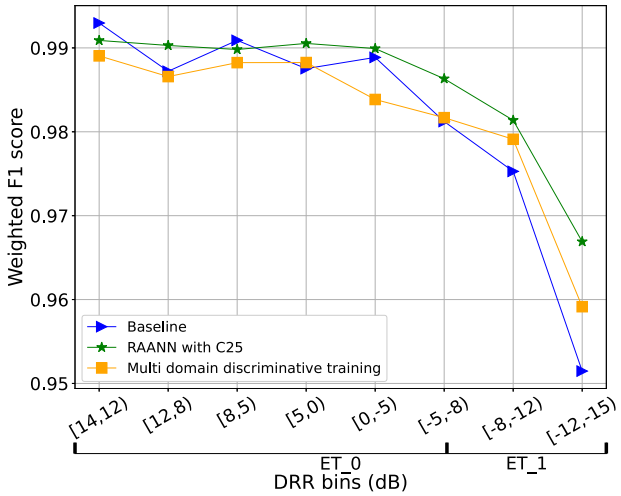

Fig. 11. Performance of room acoustic metric estimator in RAANN as a multi-domain classifier across various DRR conditions on the ET test dataset.

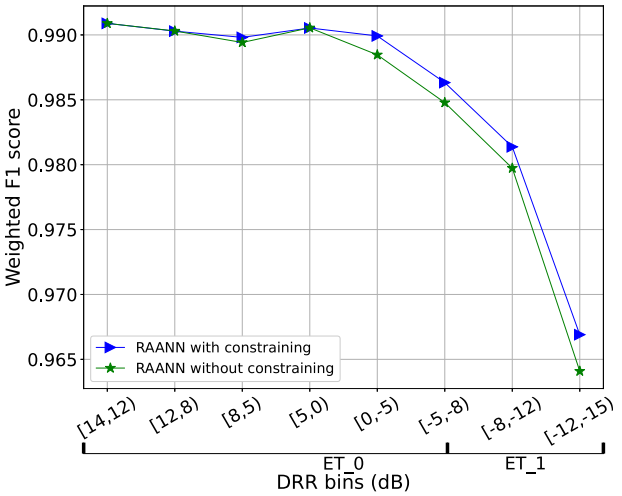

Fig. 12. Weighted F1 scores for constraining the room acoustic metric estimation in RAANN across various DRR conditions on the ET test dataset.

the acoustic metric in the datasets. The constraining is applied on the target metric and the estimated value of the acoustic metric is capped within the $(-d_{tr}, 1+d_{tr})$ where d_{tr} is a positive value. Lower value of d_{tr} implies a higher effect of constraining. The constraining reduces the contribution of the a sample to the gradient which has very high error. Hence, the effect of the outlier samples are reduced and increases the contribution of the nominal samples in model training. Empirically, a value of 0.3 for d_{tr} in Eq. (3) achieved higher improvements compared to the model without constraining (see Fig. 12).

THE AUTHORS


Sreenivasa Upadhyaya


Wim Buyens


Wim Desmet


Peter Karsmakers

Sreenivasa Upadhyaya earned his bachelor's degree in Electronics and Communication Engineering from Visvesvaraya Technological University (VTU), Karnataka, India, in 2012, and his master's degree in Artificial Intelligence from KU Leuven, Belgium, in 2019. For his master's thesis, he developed a deep learning–based method to estimate biological age using teeth images. He has 5 years of industry experience in signal processing, focusing on speech and video applications. Currently, supported by SoundTalks N.V. and a VLAIO doctoral scholarship, Sreenivasa is researching innovative acoustical classification models for livestock health monitoring. His work involves creating classification models that leverage room acoustical information to achieve robust performance under varying conditions. His primary research interests include acoustic pattern recognition, audio and video signal processing, and machine learning.

•

Wim Buyens was born in Brasschaat (Belgium) in 1976. He received the degree of Master of Science in Electrical Engineering from KU Leuven and pursued a Ph.D. in the Arenberg Doctoral School at KU Leuven on music preprocessing for cochlear implants. After his Ph.D., he joined SoundTalks and is currently Team Lead. Dr. Buyens has gained expertise in the area of audio signal processing, sound analysis, speech enhancement, machine learning, and cochlear implant sound coding.

•

Wim Desmet holds an M.Sc. and Ph.D. degree in Mechanical Engineering from KU Leuven. He is a Full Professor in engineering dynamics and mechatronics at the KU Leuven Department of Mechanical Engineering, where he is member of the Leuven Mechatronic System Dynamics (LMSD) research group. His major research interests include Digital Twin and Model-Based System Engineering developments for mechanics and mechatronics engineering. This involves advanced modeling, analysis, testing, monitoring, and control techniques in vibro-acoustics, aeroacoustics, and (flexible) multibody dynamics. Prime application domains involve sustainable engineering solutions for industrial machinery and manufacturing, green transportation, energy supply, and health care. He serves currently as Managing Director of KU Leuven.

•

Peter Karsmakers received an M.Sc. degree in artificial intelligence in 2004 and Ph.D. degree from the Department of Electrical Engineering, KU Leuven, in 2010. From 2010 to 2013, he was a postdoctoral researcher in the Mobilab research group from Thomas More. From 2013 to 2018, he worked as a postdoctoral researcher at KU Leuven where he cofounded the Advanced Integrated Sensing (ADVISE) research team. Currently, he is an Associate Professor within the Computer Science Department in the Declarative Languages and Artificial Intelligence (DTAI) section at KU Leuven and is a member of the Leuven.AI institute. Since 2022, he has been a principal investigator of Flanders Make@KU Leuven. His research interests include designing machine learning algorithms that consider application-specific constraints like the computing platform, need for physical consistency, and limited availability of annotated data. He worked on diverse industrial collaboration projects that involve monitoring applications using microphones, accelerometers, and radars.