TECHNISCHE UNIVERSITÄT ILMENAU

Fakultät Elektrotechnik und Informationstechnik

Fachgebiet Elektronische Medientechnik

Masterarbeit

Room acoustic analysis and acoustic room classification using binaural recordings in real rooms

presented by

Sayantan Gope
Matrikel - 61009

Professor:

Prof. Dr.-Ing. Karlheinz Brandenburg

Supervisor:

Dr.-Ing. Stephan Werner

Ilmenau, den August 1, 2022

Acknowledgements

I would like to express my deepest gratitude to Professor Karlheinz Brandenburg and my supervisor Stephan Werner for their guidance, as this thesis would not have been possible without them. I would also like to thank Lukas Treybig, Ulrike Sloma and Florian Klein for their help with this thesis as well as their advice and feedback. I am also thankful to my family and friends for their support. Special thanks to Aakash Mehta for his help with the report and the various procedures involved in writing and submitting a thesis. Lastly, I would like to express my sincere thanks to Tereza Suchánková and Tomáš Suchánek for their help with the report and presentation, and their support throughout this thesis.

Zusammenfassung

Jeder Raum hat eine einzigartige akustische Signatur und Charakteristik. Sie spielen eine große Rolle für die akustische Wahrnehmung des Zuhörers. Das Verständnis der Raumakustik kann dabei helfen, Räume effizient zu gestalten und sie für die vorgesehenen Zwecke besser geeignet zu machen. Die akustischen Eigenschaften eines Raums können mit dem Room Impulse Response (RIR) oder Binaural Room Impulse Response (BRIR) gemessen werden und sind für verschiedene Räume sowie für die Positionen der Lautsprecher und Empfänger einzigartig. Ziel dieser Masterarbeit ist es, mit Hilfe von Deep Learning ein Convolutional Neural Network (CNN)-Modell zu erstellen, um Räume anhand ihrer akustischen Eigenschaften durch BRIR-Aufnahmen zu klassifizieren. Durch Datenerweiterung werden die Eingabedaten auch verwendet, um reale Musik- und Sprachsignale zu emulieren. Solche Musik- und Sprachsignale sowie rein akustische Informationen in Form von Impulsaufzeichnungen werden vom Klassifikator als Eingabe verwendet. Ziel dieser Arbeit ist es auch, die Entscheidungen der CNN-Architektur zu bewerten und die Funktionsweise der CNN zu verstehen. Zu diesem Zweck wird auch Layer-wise Relevance Propagation (LRP) eingesetzt. Das trainierte neuronale Netzwerk zeigt vielversprechende Ergebnisse bei der Klassifizierung von Räumen auf der Grundlage ihrer akustischen Eigenschaften. Es erreicht eine nahezu perfekte Genauigkeit von 99% bei Impulsaufnahmen und 100% bei emulierten Musik-/Sprach-Eingangssignalen nach weniger als 5 Stunden Trainingssitzung. Die Testergebnisse werden in Form einer Konfusionsmatrix visualisiert. Die wenigen ungenauen Klassifizierungen geben einen Einblick in die Denkweise des neuronalen Netzes. Es zeigt sich, dass die Ergebnisse des Klassifikators mit den Ergebnissen anderer Arbeiten übereinstimmen, die auf der menschlichen Wahrnehmung basieren.

Abstract

Every room has a unique acoustic signature and characteristics. It plays a big role in the acoustic perception of the listener. Understanding room acoustics can help design efficient rooms and make them more suitable for their intended purposes. The acoustic characteristics of a room can be measured by the RIR or BRIR, and are unique for different rooms as well as for the positions of the speakers and receivers. This master thesis is aimed at using Deep Learning to create a CNN model to classify rooms based on their acoustic characteristics through BRIR recordings. Through data augmentation, the input data is also used to emulate real-world music and speech signals. Such music and speech signals, and pure acoustic information in the form of impulse recordings are used as input by the classifier. It is also the aim of this thesis to assess the CNN architecture decisions and understand the "black box" working of the CNN. LRP is also utilized towards this goal. The trained neural network shows promise at classifying rooms based on their acoustic characteristics. It achieves near perfect accuracy of 99% on impulse recording input and 100% accuracy on emulated music/speech input signal after less than 5 hours training session. Test results are visualized in the form of a confusion matrix. From this, the few inaccurate classifications provide insight into the thinking of the neural network. It is seen that the classifier findings correspond with findings of other papers based on human perception.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Room acoustics is a sub-discipline of acoustical engineering that studies the behaviour of sound fields in an enclosed or semi-enclosed space [5]. While this space is filled with air in most cases, it can also include spaces or cavities filled with other fluids through which sound can propagate. Each space based on its contents and its nature behaves differently when there is sound propagating in it. The size of the room and space for sound to propagate, placement of objects, their reflective or absorptive nature, intended position of listeners can all make a large difference in how a room sounds. One key aspect of room acoustics is that due to the enclosed nature of the space there are several reflections from the boundary surfaces that make up the sound field in addition to the direct propagation of sound from the source to the listener as shown in Fig. 1.1. Further, sound of different frequencies also behaves differently in a room.

Applications for room acoustics can include noise control from various sources of unwanted noise which is undesirable in a particular room. Knowledge of room acoustics can be used to suppress any unwanted noise by making use of strategically placed sound absorbing objects. This can improve the acoustics of sound in a room by reducing echoes and reverberation of sound.

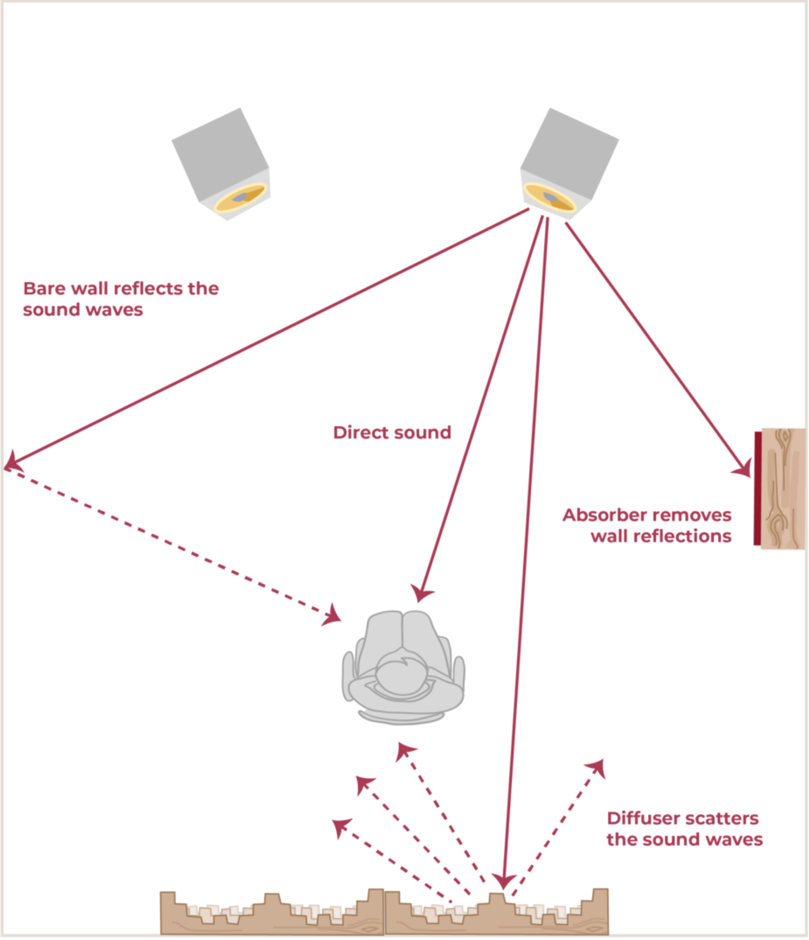

Bare wall reflects the sound waves
Direct sound
Absorber removes wall reflections
Diffuser scatters the sound waves

Figure 1.1: Sound propagation in an enclosed room [1]

The other type of application of room acoustics is using auditory information. Unlike the previous application, in this case listeners are interested in certain auditory information, which is mostly speech or music. This application of room acoustics is the area this thesis falls under.

1.1 Motivation

Understanding the behaviour of sound can help make rooms more suitable for different tasks. A classroom should be designed for clean crisp speech transmission so that the lecturer is audible and understandable to the students. Echoes or a long reverberation can hinder the intelligibility of the speech. A concert hall on the other hand should prolong and enrich the sound of the orchestra to make it sound more impressive. As such, understanding and predicting the acoustic behaviour of a room can be used for numerous applications such as creating a better sound profile in Augmented Reality (AR) applications [6], locating and tracking an acoustic source [7], noise control in reverberant rooms [8][9], or for designing new rooms to be more suitable for their intended auditory use.

For many of these applications, knowledge of room acoustics and the behaviour of sound field is vital. An audio recording when played in different types of rooms can be have audible difference. Understanding exactly what that difference is and how it occurs can prove to be useful in implementing a wide variety of acoustic applications such as ones mentioned before.

The challenge in this thesis is to make comparisons of classification results among sets of data where the input data itself differs. As such, the classification results need to be assessed in context with the weight of differences in type of spectral information used as input, variation in length of input files, nature of the audio files etc. Further analysis is made regarding the relevance of various acoustic parameters to the classification task, such as reverberation time.

1.2 Topic Description

The acoustic characteristics of a room can be measured by its RIR or BRIR, which are unique for different rooms as well as the position of the speakers and receivers. The objective of this thesis is to use Machine Learning Algorithm (MLA)s for classifying rooms based on BRIR recordings. Classification is to be made with impulse recordings from various types of rooms. Further the input data is augmented by convolving impulses with dry audio tracks to imitate real-life applications. Classification results will be assessed to see the difference speech or music information can make in the accuracy to bolster information of a room's acoustic characteristics.

1.3 Research Question

In this thesis, the following question is to be answered:

> ”How can rooms be classified based on prerecorded impulse recordings?”

Further, the subtask arises:

> ”To find which factors most differentiate the acoustic characteristics of various types of rooms.”

The focus of this thesis will be on the selection of suitable networks for the MLA as well as performing the classification of rooms based on prerecorded dataset of impulse recordings in different rooms. With data augmentation, we add speech and music information to the input to provide a practical use case for this thesis. Lastly, we try to explain the decisions of the CNN. Various factors can affect the decision making of the CNN classifier such as room acoustic parameters, audio duration, data augmentation steps such as noise padding etc.

This research will help broaden the understanding of room acoustics and make it easier to identify and predict the possible acoustic behaviour of a room. This helps in finding out what auditory purposes are suitable in a particular room. It makes it easier to design rooms to fit specific requirements. This knowledge can also be used by AR to provide a more immersive experience for the users.

1.4 Chapter Overview

Chapter 2 ”Scientific Background” provides the scientific fundamentals in the field of acoustical engineering and room acoustics in general which forms the basis of this thesis. It goes into detail about the study and various applications of room acoustics and possible use cases of this thesis. Section 2.2.1 goes into detail about Virtual Acoustics which is a prominent usecase of this thesis. Another such usecase is described in Section 2.2.2 which explores the process of recreating the acoustic features of certain desirable rooms and heritage sites. Further it follows State Of The Art (SOTA) in the fields of CNN and interpreting and understanding Deep Neural Network (DNN).

Chapter 3 ”Implementation” contains an explanation of the entire research process going into the decisions made with respect to designing the CNN. Section 3.1 and 3.2 explore the methods used for data augmentation and the design of the CNN architecture is provided in Section 3.3. Section 3.5 goes over the scrapped ideas and failures. The software and python libraries used in this code are shown in Section 3.6.

Chapter 4 ”Experiment” explains the experimental setup. Section 4.1 describes the dataset used for training and testing. It also gives an overview of the rooms that are the classification targets. Hardware specifications for the device used for experiments is provided in Section 4.2. Section 4.3 explores the experimental setup and procedure.

Chapter 5 ”Results” provides the results of the classification. The raw data is processed and tabulated here. Further, graphical representation is also found here. Section 5.1 provides the results of training on subset and Section 5.2 provides the results of training on the full dataset. Tabulated results are shown in Sections 5.1.3 and 5.2.3.

Chapter 6 ”Evaluation” provides an analysis of the classification results. It provides a brief discussion of the results obtained and the acoustic characteristics that could have led to the decision the CNN made. Section 6.3 analyses the results obtained from the CNN classifier and the insights obtained into the “black box” decision-making process.

Chapter 7 ”Conclusion” presents the conclusions drawn from the classification and analysis. Section 7.1 goes into the possible approaches that can build upon this thesis.

Chapter 2

Scientific Background

A variety of studies have been carried out in the field of room acoustics. Several approaches have been studied when classifying rooms, such as based on their reverberation effect. Various classifiers have been proposed for this purpose such as a Gaussian Mixture Model (GMM)-Universal Background Model (UBM) classifier using Mel-frequency Cepstral Coefficient (MFCC)s [10], a Gaussian Naïve Bayes Classifier (NBC) based on Negative-Side Variance (NSV) used by Moore et al. [11]. An NBC was also used in [12] using Frequency-Dependant Reverberation Times (FDRT). Papayiannis et al. in their paper ”Discriminative feature domains for reverberant acoustic environments” [13] show that spectral and energy decay features like FDRTs can be used for room classification and can be extracted from Acoustic Impulse Response (AIR).

Apart from room classification, deep learning has been used in the past to learn about other features of reverberant rooms such as reverberation time [14][15], Early-To-Late Reverberation Time [15] etc. In [16], five DNN model architectures were compared for classifying reverberant rooms in which the highest accuracy was provided by a Convolutional Recurrent Neural Network (CRNN) architecture that incorporates an attention-mechanism.

When dealing with room acoustic parameters, blind recreation [17][18] would require acoustic simulation of the room. This process can be unsuitable for real time parameter calculation for a variety of reasons such as presence of loud environmental noise or being a time-consuming process. As such, acoustic parameter measurements such as those by Werner et al. [19] can be used instead. This is used by Dziwis et al. [6] in the field of AR to perform machine learning based room classification by estimating room characteristics, using images as the input. Building on the work shown above, this thesis uses Melspectrogram as input features and pre-recorded impulse response is used as input to the CNN classifier.

2.1 Spatial hearing in enclosed spaces

There are several descriptors for spatial hearing in room acoustics. Room acoustics is typically studied through prerecorded or estimated impulse responses at key locations, as the impulse response of a room contains all the information about its acoustics [5]. From impulse responses, we can calculate numerical descriptors of room acoustics.

Among these, the most commonly used parameter is Reverberation Time (T60), which is defined as the time interval for the sound pressure level in a room to drop 60 db. Early Decay Time (EDT) quantifies the decay of early reflections of the propagating sound. C50 is a measure of clarity of speech measured in Decibels (dB) similar to C80 and D50. Another method to estimate the ratio of early and late energy is called Centre time, t_s. While these descriptors quantify the quality and clarity of the sound propagating in a room, they do not give a description of the spaciousness of the room. Spaciousness can be quantified by Inter Aural Cross Correlation (IACC). An initial goal in this thesis was to use precalculated acoustic parameters such as those mentioned above as input to the CNN classifier, however this idea was later scrapped as shown in Section 3.5.

2.2 Applications of Room Acoustics

In [5], Masovic et al. state,

> ”The applications of room acoustics spread in two main directions. The first group of applications involves primarily noise control of various sources of unwanted sound which may be present inside or, less commonly, outside the room.”

> ”Quite different is the second group of applications, in which sound fields in rooms carry a specific auditory information, typically music or speech, which is of interest to its receivers.”

This group of applications include a myriad purposes. While the receiver is typically a human listener who is present in the room with the audio, it is far from the only use-case. One field of application is the recreation of the acoustics of a desirable room or enclosure in a separate location. As previously stated this can be applied in the field of AR to create audio profiles. Further, recreating the acoustics of a room is used in Virtual Acoustics or Auralization. In some cases, the acoustics of existing rooms can be used to design acoustics of rooms that do not exist, or that we do not have access to using Convolution Reverb. This is examined in more detail in Sections 2.2.1 and 2.2.2 respectively.

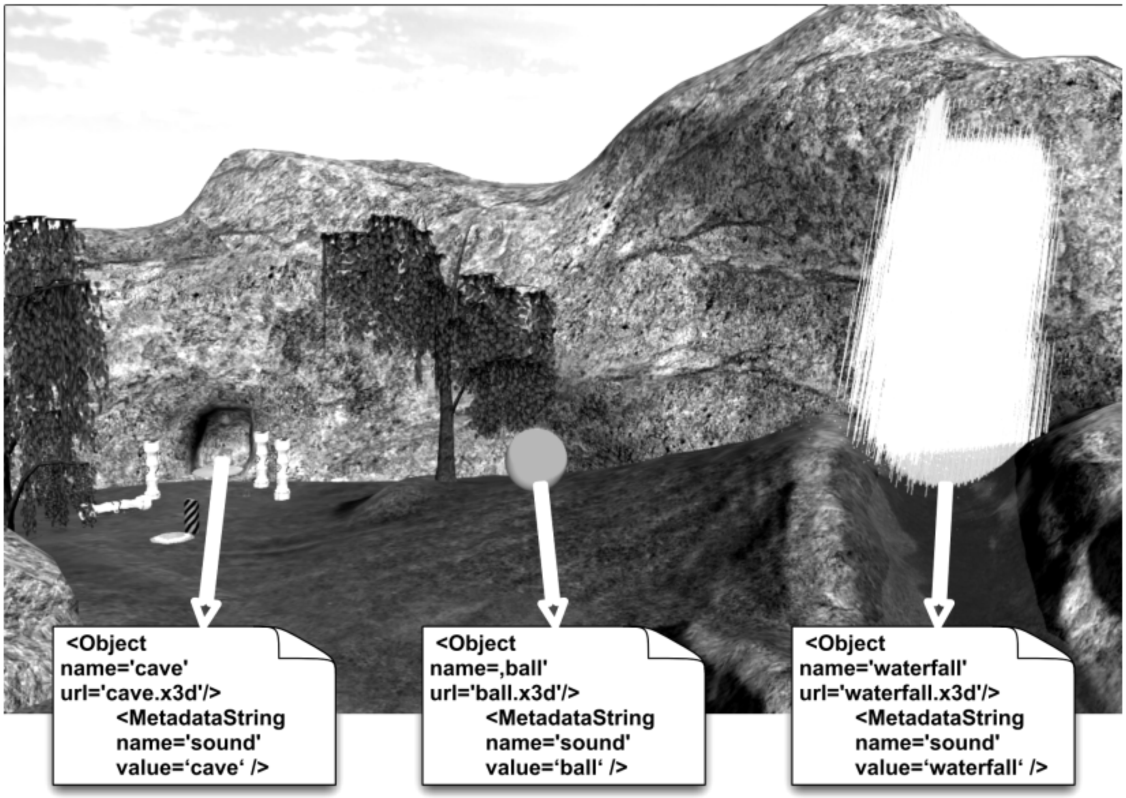
Such applications of room acoustics, while not directly related to this thesis, are areas where the findings of this thesis can be used to increase the effectiveness and create simpler implementations. Virtual Acoustics, explored in Section 2.2.1, is an area where understanding of room acoustic features can drastically improve performance. As this thesis aims to find the most relevant acoustic parameters for room acoustics, this knowledge can lead to a more focused approach to virtual acoustic applications. Room Acoustic Synthesis, as shown in Section 2.2.2, on the other hand is a process that is utilized in this thesis as a data augmentation method. Section 3.1 describes how this process is used to emulate music and speech signals in the target rooms by convolving dry audio tracks and impulse recordings from these rooms. The following sections provide the background and development for both of these procedures.

2.2.1 Virtual Acoustics

Virtual Acoustics (VA) is the general term for “physically-based or perceptually-based modeling of sound sources, room acoustical systems and human spatial hearing primarily by means of Digital Signal Processing (DSP)” [20]. The beginnings of artificial speech production can be traced all the way back to the 1700s when von Kempelen and Kratzenstein created so called ”talking machines” [21] [22]. However, the idea of VA was developed much later in the 1960s with DSP and improvements in computer technology.

The three main components of VA are source, room and listener modeling. The first real-time consumer application featuring these three was introduced by Foster et al. [23] who synthesized three dimensional sound cues over headphones using a powerful signal processor called the Convolvotron, which was based on earlier attempts by Wenzel et al. [24].

In recent past, a lot of development has taken place in the field of VA to produce virtual spaces with accurate acoustics. In [2], Fröhlich et al. proposed semantically annotated ”smart objects” that know how they should sound as shown in Fig. 2.1. By using this method, they intended to create a more realistic auditory experience by having multiple sound sources instead of a single independent source. These smart objects can be made more detailed by focusing more on the most relevant acoustic parameters, whose relevance to room acoustics is analysed in this thesis.



```
<Object
name='cave'
url='cave.x3d'/>
        <MetadataString
        name='sound'
        value=‘cave‘ />
```

```
<Object
name=,ball'
url='ball.x3d'/>
        <MetadataString
        name='sound'
        value=‘ball‘ />
```

```
<Object
name='waterfall'
url='waterfall.x3d'/>
        <MetadataString
        name='sound'
        value=‘waterfall‘ />
```

Figure 2.1: Semantically enriched smart objects [2]

2.2.2 Room Acoustic Synthesis

When sound is reproduced for playback, it is not enough to just play the audio. The environment that it is played in makes a huge difference to the perception of the listener. That is why locations like certain prestigious concert halls, theatres and opera houses are famous. It is desirable to recreate the experience of being in such places and to do that, it is necessary that the sound source is faithfully reproduced and “the room-acoustical percept observed in the recording room should be reproduced in the playback room” [25].

There are many methods utilised when trying to reproduce the sound field such as Wave Field Synthesis (WFS) [26], ambisonics [27] or just playing back the audio over stereo speakers. The reverberation effect of a room can be modeled by simple convolution of an input signal with an impulse response recorded in the target room [28], which is a method utilised in this thesis as well. The problem with WFS and ambisonics methods are that they assume that the sound field is being reproduced in a room without reflective boundaries to ensure that the room does not affect the reproduced audio. This is seldom the case in reality however, and the boundaries of the room and reflective surfaces add the room's own reverberation to the acoustic reproduction being played which creates additional coloration [29] or increase reverberation time [30].

To account for the influence of the playback room, several approaches have been tried such as adaptive filtering in the wave domain [31] and crosstalk-cancellation for listening room compensation [32]. de Vries et al. [33] state that "the acoustics of any room can be added to dry recorded sources by convolving their signals with a distribution of virtual mirror image sources created by WFS, using a 'surround' configuration of loudspeaker arrays". However they find that to not be the most practical approach and propose a more appropriate method based on impulse response measurements. Grosse et al. [25] propose a method to reduce the detrimental influence of the playback room by optimizing IACC coefficient within critical bands as well as reverberation time of the recording room.

It is of vital importance to save the sound profile of prestigious locations in the same vein as preserving heritage sites. The Past Has Ears (PHE) project [34] is a research consortium dedicated to such acoustic heritage preservation. They use cultural heritage sites in France, Italy and United Kingdom as case studies and have developed methods for acoustic inquiry of these heritage sites. One such study that is a part of PHE project is by Eley et al. [35]. They have developed a real-time auralization architecture to recreate "the acoustic conditions within the Cathedral of Notre-Dame de Paris for a choir ensemble". Members of a medieval choir participated in a subjective test of the quality of the rendering. They also studied the interactions between musicians and the acoustics of such acoustic locations. From other studies, it is seen that reverberation time has the most measurable effects on the performance of musicians [36] [37].

2.3 Deep Learning

Modern research and developments in engineering and IT related fields have largely been possible to due developments in machine learning. Using machine learning models, Artificial Intelligence (AI) and deep learning to process and learn data has pushed the pace of improvement dramatically. Over time several types of network architectures have been developed. This includes DNNs in which data is processed by moving forward through the network. A Recurrent Neural Network (RNN) similarly is a Feed Forward Neural Network (FFNN) but with an added time twist, having features link to past nodes through which information can flow back to previous sections of the network. However, RNNs sometimes run into the issue of losing weights in the long term. A Long Short Term Memory (LSTM) was designed to counter this issue by learning long term dependencies and patterns to make better decisions. Further, CNNs can be combined with RNN and LSTMs for various tasks such as automatic image captioning [38].

Such deep learning techniques have grown in popularity thanks in large to improvements made in the field of image classification [39]. A lot of these innovative breakthroughs have occurred from submissions in an annual computer vision competition, namely the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). One of the most popular and widely used network architecture for image processing applications, VGG16, was first introduced at ILSVRC 2014 where it produced the best results in image classification and localization. The VGG16 was then published by Simonyan et al. [40] in 2015. The ILSVRC has also seen several other notable submissions such as the GoogLeNet model by Szegedy et al. [41] from Google in 2014, ResNet model by He et al. [42] in 2015. Both of these models achieved the best results for object detection in their respective years. Outside of image classification, deep learning has been used to perform a variety of operations in the field of image processing with constant new improvements. Zhang et al. [43] created a feed forward denoising CNN to overcome the drawbacks of current state of the art techniques like BM3D [44] such as being time consuming. Dong et al. [45] use CNN for image super-resolution tasks to create high resolution images for low resolution inputs. FaceNet model was created by Schroff et al. [46] for facial detection, recognition and clustering.

A growing use of deep learning is in the field of Natural Language Processing (NLP). In [47], Yoav Goldberg describes NLP as “a collective term referring to automatic computational processing of human languages”. Deep learning has been used in NLP for various tasks such as text classification and sentiment analysis [48], caption generation [49], machine translation [50] and document summarization [51]. For speech recognition tasks, Abdel-Hamid et al. [52] implemented a hybrid DNN/Hidden Markov Model while Graves et al. [53] use an LSTM.

2.3.1 Deep Learning in Audio Processing

Then comes the use of deep learning in audio processing. There are a myriad of tasks in the field of audio processing that can be accomplished with the use of AI and deep learning. A lot of these applications are in the field of Music Information Retrieval (MIR) such as automatic music tagging, musical genre classification [54], beat tracking [55], music recommendation, onset detection [56], music transcription to generate sheet music etc. For speech signals, speech recognition and audio transcription are commonly used by smart assistants such as Siri, Cortana, Alexa and Google Assistant on phones or devices like Amazon Echo. Many companies use MLAs to provide real world applications. Spotify uses machine learning techniques like collaborative filtering and content modelling to provide recommendations, understand voice commands with speech recognition, provide lyrics with caption generation and many more of its features [57]. Shazam and SoundHound perform audio fingerprinting for music recognition to identify the song from an audio clip.

In the past, audio machine learning applications used to rely on traditional DSP techniques to extract features, such as extracting phonemes for audio analysis to understand human speech [58]. With deep learning however, audio data is not handled in its raw form. Audio files are converted into spectrograms which are used as images when provided as input to CNN architectures. In that sense, audio processing using deep learning can be seen as an extension of image processing. Similar to image classification, audio classification is a vital task. Hershey et al. [59] utilize several popular image classification CNN architectures such as VGG, ResNet, AlexNet etc to perform large-scale audio classification. Lu et al. [60] propose a robust algorithm for audio classification and segmentation. Audio classification tasks also include environmental sound classification [61] and emotion detection in speech [62] or music.

2.4 PyTorch

When it comes to deep learning frameworks, there are three options that stand out namely Keras, TensorFlow and PyTorch [63] [64]. While all three are great options, they have a few differences. Keras [65] is an open source neural network library written in Python. It is a high level Application Programming Interface (API) that can run on top of TensorFlow and is designed to be user friendly, modular and provide fast experimentation. Keras has the simplest architecture and is used for smaller datasets as it is comparatively slower. It was integrated into TensorFlow in mid-2017. TensorFlow [66] is an open source library for dataflow programming which was developed by Google. It is a symbolic math library that provides both high and low-level APIs. TensorFlow is fast and provides high performance. It has a well-documented framework and abundance of trained models. TensorFlow has great visualization allowing developers to track the training process better. Finally, PyTorch [67] is an open source machine learning library for Python designed by Facebook's AI research group. It is a low-level API which is known for its simplicity, flexibility and more manageable coding. PyTorch is a fast library with similar pace to TensorFlow, and can handle large datasets. Unlike the other two options, PyTorch has much more debugging capabilities. Due to the fast execution, high performance and debugging options, PyTorch was selected for use in this thesis.

2.5 Explaining DNN decisions

As we see in Sections 2.3 and 2.4, deep learning is a widely used scientific tool responsible for many advancements in various fields such as computer vision, audio processing and natural language processing. Even though AI and deep learning can perform these tasks with accuracy on par with humans, it is extremely important to understand their working and verify what they are learning from. A neural network needs to learn proper features and problem solving, rather than exploiting artifacts in the data [68]. Such interpretability is extremely important in fields where automated tasks are required such as self driving cars [69] or medicine [70], where absolute certainty is required that the model has learned the correct features.

Several studies have been made in order to verify and validate the performance of neural networks [71]. In [72], Bach et al. perform Pixel-wise Decomposition to understand the contribution of a single pixel to the classifier in an image classification task, in order to understand the black box working of classifiers such as ImageNet. Montavon et al. [3] give an overview of techniques for understanding complex machine learning model decisions, with a focus on DNN.

2.5.1 Sensitivity Analysis

One approach is Sensitivity Analysis, which is based on a model's local gradient. As such, it explains the local relation of the function rather than the importance to the function itself. Sensitivity analysis has been regularly used in scientific tasks such as in classification and medical diagnosis of cancers [73]. Zurada et al. [74] use it in order to minimize redundancy in input data. It is even used in image classifications tasks. Montavon et al. [3] show an example of sensitivity analysis used in image classification using the MNIST database to explain the decisions of a DNN as shown in Fig. 2.2.

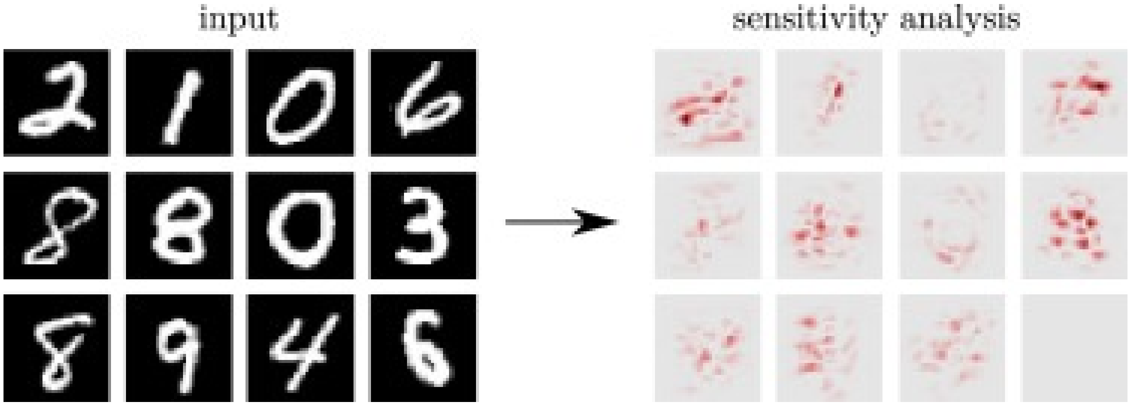
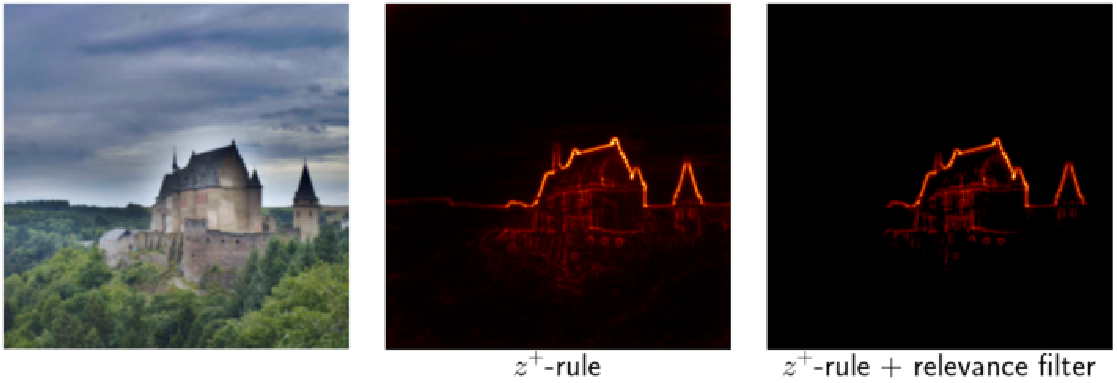


Figure 2.2: Sensitivity analysis of DNN classification on MNIST database by [3]

2.5.2 Layer-wise Relevance Propagation

LRP is a backward propagation technique to find out the relevance of different parts of the input to the model decision at every layer of the neural network. LRP was introduced by Bach et al. [72] and is based on a conservation property where each neuron has a share of the model's decision and passes it on to its predecessors in the previous layer. In [75], Montavon et al. provide a tutorial for the use of LRP in image classification tasks which is widely used. Fabi et al. [4] made an easy to use LRP implementation with an additional relevance propagation filter for better heatmaps to display results clearly.



z^+-rule

z^+-rule + relevance filter

Figure 2.3: Layerwise relevance propagation with relevance filter by [4]

LRP is applicable in a broad range of fields. Sturm et al. [76] use it in the field of cognitive neuroscience to interpret the model decisions and provide insight into neurophysiological phenomena. It is widely used for image classification [4] [72], identifying facial expressions [77], and in NLP to categorize text documents [78]. Given its broad applicability, we try to adapt LRP for an audio classification task in this thesis and identify relevant features for classification of rooms.

Chapter 3

Implementation of the System

To perform acoustic room classification, this thesis started on a base set of ideas and changes were made as required as progress was made towards the completion of this task. The main goal has been to provide understanding of what parameters play an important role towards differentiating the identifying features or feel of a particular room. A CNN model pipeline was created to perform such a classification and by understanding the "black box" working of the classifier, this would allow us to tackle the goal of identifying the most important acoustic features. There were various challenges along the way such as working with several input types, adapting an image classification algorithm to work for audio classification etc. Some ideas were scrapped in favour of others in the development stage that appear more promising to the goal of the thesis.

More details about the implementation and thesis setup are provided in the following sections of the chapter. The workflow diagram for this classification task is shown in Fig. 3.1. To summarize, two types of input data are used in this classification, impulse recordings and dry audio convolved impulses. The reasoning behind this and data preparation and augmentation steps involved in classification are described in detail in Sections 3.1 and 3.2.

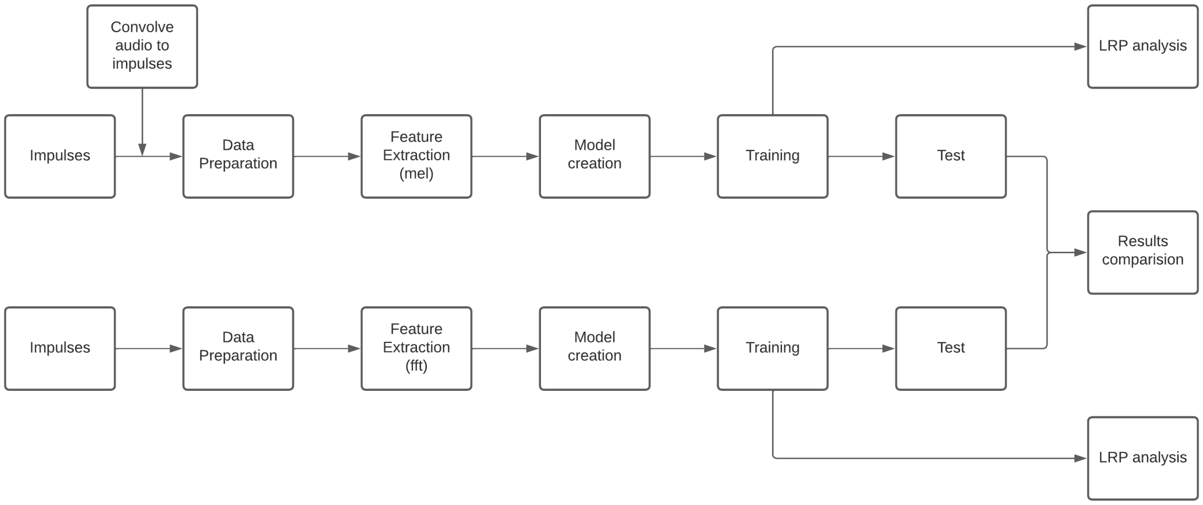


Figure 3.1: Workflow of room classification task and analysis

The prepared input data is provided to the neural network in Section 3.3 and the output is provided in Chapter 5 and analysis is made from the results in Chpater 6. LRP is used as a means of understanding the CNN model's decision making. Section 3.5 describes the challenges faced in the implementation of the room classification task, along with scrapped ideas and why it was decided to leave such ideas out. Finally, the software use and code implementation is shown in Section 3.6.

3.1 Convolved Audio Path

As there are two different types of input used, the implementation of the thesis has been divided into two paths as shown in Fig. 3.1. From the figure, it can be seen that the two paths differ in terms of inputs and data preparation and augmentation, though they are identical once input is provided to the neural network. As such, the inputs and data preparation steps are discussed in Sections 3.1-3.2 and the common steps following that are shown in Sections 3.3 and 3.4.

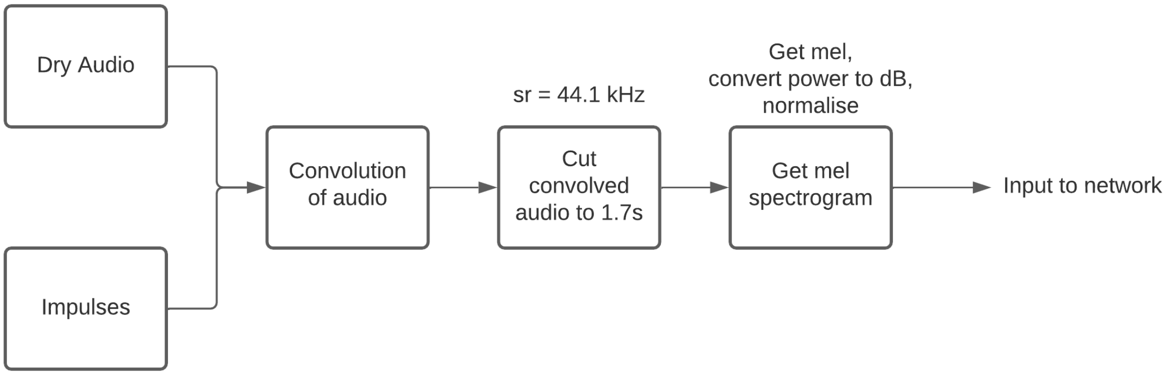


Figure 3.2: Convolved audio path workflow

The first path is the so called “Convolved Audio” path. This path uses impulse recordings convolved with dry audio tracks as input. The idea behind this is to emulate a real world input that can be made by anyone on the simplest recording setups such as their mobile phones. Dry audio tracks contain only music or speech information without providing any information about the acoustics of the room in which they were recorded. The impulse recordings on the other hand represent acoustic characteristics of a room. As such, by adding dry audio tracks we add music and speech signals thus emulating a real world input, without overwriting the acoustic characteristics of the rooms we are trying to classify.

Fig. 3.2 shows the data preparation steps for the Convolved Audio path. After the impulses and dry audio tracks are combined by convolution the audio is cut to two different lengths for comparision, of 1.7 and 5 second duration. A 1.7 seconds duration is taken to match the duration of impulse recordings used as input in Section 3.2 for which the reason is stated there, while an audio length of 5 seconds much better emulates a real world input. Comparision is made between these two audio lengths to see the effect of audio duration on the classification. These two operations constitute the data augmentation steps for this path. Once data augmentation is done, the next step in the data preparation is to perform feature extraction. Since the input of this path represents music and speech signals, we obtain melspectrogram features of the audio as input to the CNN. The mel scale was proposed by Stevens et al. [79] as a subjective unit of pitch which is similar to how humans would hear. In the mel scale, sounds of equal distance to each other sound equidistant to a listener. As such, melspectrogram features are commonly used as input to a neural network model.

3.2 Impulses Path

The second path is the “Impulses” path. This path utilizes a dataset of impulses recorded in several rooms of a variety of types, which is described in Section 4.1. These impulse recordings represent the acoustic characteristics of a room in a pure form. Due to the fact that some rooms have longer echoes than others, the impulse recordings for the rooms are of different length from each other. However, for the sake of simplicity and coherence, inputs to the CNN need to have the same dimensions in a batch of training data. After considering the spectral information of the impulses and acoustic parameters such as reverberation time, it is seen that only one room (H1562) has late reverberations going past 1.7 seconds length. As such, 1.7s is taken as the duration for impulses. As a part of data preparation, all impulses are either cut or padded to that duration. This step is of vital importance, and it leads to certain data preparation methods not being a viable option as seen in Section 6.2.2.

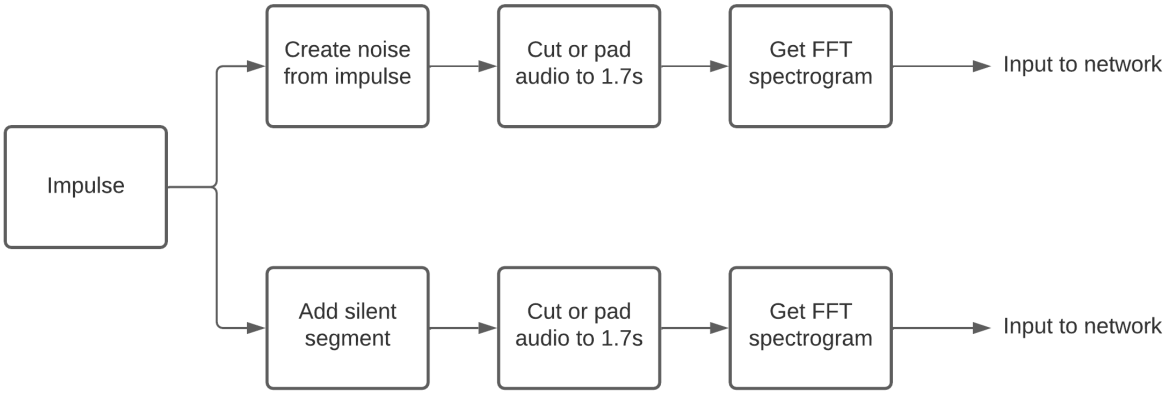


Figure 3.3: Impulses path workflow

As shown in Fig. 3.3, the data augmentation for this path involves cutting or padding the impulse recordings to the length of 1.7 second duration. Only one of the rooms, namely H1562 which is the foyer of the Helmholtz building in TU Ilmenau, has an impulse recording longer than 1.7s as it is a big open area and highly reverberant, so the audio is cut to the required length. For the other rooms, the impulse needs to be padded to the required length. To achieve this, there are two possible options which are noise padding and adding a silent segment at the end.

Adding a silent segment at the end of the recording is a simple task, but noise padding requires more meticulousness. It is important that the noise part does not have its own room acoustic features but that of the room the recording is from. To do that, the noise is created from the impulse recording itself. A small clip is taken from the last 50 ms of the impulse and randomized and added at the end of the impulse over and over until the required length is reached. The effect of noise padding and silence padding is shown side by side in Fig. 3.4 in the spectrogram of the same impulse recording.

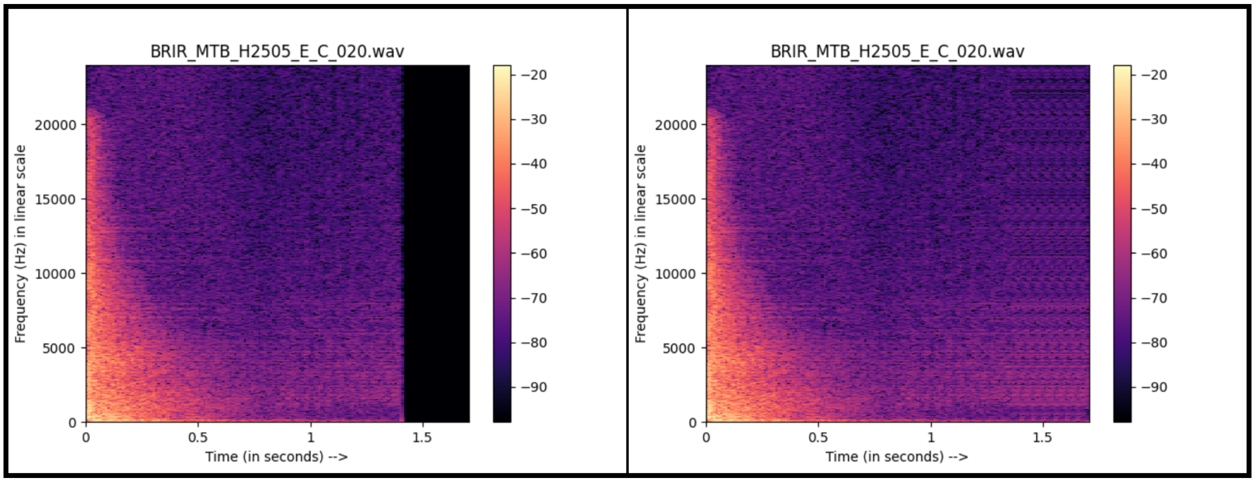


Figure 3.4: Spectrogram showing silence padding vs noise padding

After these data augmentation steps are performed, the next step is feature extraction. Unlike the first path, impulses on their own are quite short with most of them being well under 0.5s duration. For this reason, Fast Fourier Transform (FFT) magnitude spectrogram is used instead of melspectrogram. The parameters used for the Mel-spectrogram in the “convolved audio” path and FFT magnitude spectrogram and the reasoning behind it is explored in Section 4.3.

3.3 Neural Network

The neural network used in this thesis is a CNN that is written in Python 3.8 with the use of PyTorch and its Torchaudio libraries. It is a rather simple model with four convolutional layers, with Rectified Linear Units (ReLU) activation and batch normalization is performed at each layer. ReLU is the most commonly used activation function as it is simple, fast, flexible and performs well. The four convolutional layers are followed by an adaptive average pooling layer. Finally a linear layer flattens the output and provides the classification as one out of the available classes, which are seven rooms in this case.

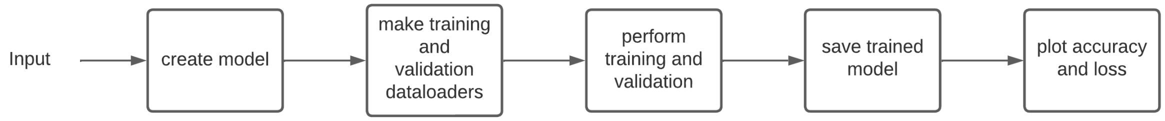

Input
create model
make training and validation dataloaders
perform training and validation
save trained model
plot accuracy and loss

Figure 3.5: Workflow of CNN classifier

From Fig. 3.5 it is seen that the next step after model creation is to create dataloaders for training and validation. When training a neural network model, the typical practice is to provide the input data in batches and the model updates itself after every batch. Further, reshuffling the input data after every epoch helps reduce overfitting of the model. Overfitting refers to a model that has been trained too much on one set of data such that it performs very well when the input is part of the data it was trained on, but suffers on any input outside that. A Dataloader is an iterable object that performs these operations for us in an easy API.

The training data is split in a 7:3 ratio using the random_split() function from PyTorch to get training and validation datasets. While it can be an option to try to split the dataset while maintaining a balanced spread of rooms, positions of setup and rotation angle, it would be a tedious task and would have to be manually balanced. Even then, it might lead to a dataset that only has certain positions and angles etc., and not be a proper representation of the entire dataset. As such, randomly splitting the dataset will provide a mostly evenly balanced spread while saving time. This data is then provided to the model by training and validation dataloaders during the model training phase. Once training is complete, the trained model is saved by saving the model's state_dict() for future use in testing and LRP implementation. A model's learnable parameters such as weights and biases are stored in the model's parameters which can be saved using the state_dict(). Further, the model's accuracy and loss is plotted for analysis of results.

3.4 Layer-wise Relevance Propagation

Following the model training, LRP is performed as a means to interpret the neural network's "black box" understanding. In LRP the inputs are passed through the trained model which was previously saved, but they are propagated backwards through the model starting at the output layer going all the way up the model architecture to the input layer.

One key property utilized in LRP is the conservation of relevance, $\sum_j R_j = \sum_k R_k$ where j represents the neurons of the current layer and k represents the neurons of the following layer. Thus, the relevance can be formulated as,

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k \tag{3.1}$$

where z_j_k represents the contribution of neuron j towards neuron k. Further, as this model utilizes ReLU activation, the activation of neurons j and k in consecutive layers can be described as,

$$a_k = max(0, \sum_{0,j} a_j w_k) \tag{3.2}$$

as the ReLU activation function passes the input value directly if it is positive and passes 0 if it is negative. Thus by combining eqn. 3.1 and 3.2 the previous relevance rule of the LRP function can be modified as,

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k \tag{3.3}$$

Equation 3.3 is a general propagation rule known as LRP-0 rule. From this general rule, more specialized rules such as LRP-$\in$ and LRP-γ are derived [72]. The output of LRP is a plot showing the relevance values across the spectrogram used as input. This plot is then used as a mask on top of the input spectrogram to display the relevant portions of the spectrogram as highlighted.

3.5 Challenges and Scrapped Ideas

Over the course of the thesis, many challenges were present. The first challenge was the sheer amount of data augmentation options that were present. On many steps along the way multiple options to utilize were encountered such as padding with noise or silence, selecting the optimal audio length and the optimal features to use as input. In many cases like this, rather than choose one option, both options were considered and tested against each other to find out the optimal solution as seen with the two audio lengths and pitting noise padding against silence padding. Further, there was a hardware memory size limitation that was encountered during the training. The FFT features used as input in the impulses path were bigger is size than the mel features used in convolved audio path. This leads to being limited to a batch size of 8 for the second path due to memory constraints. As such to keep it fair, both paths use a batch size of 8.

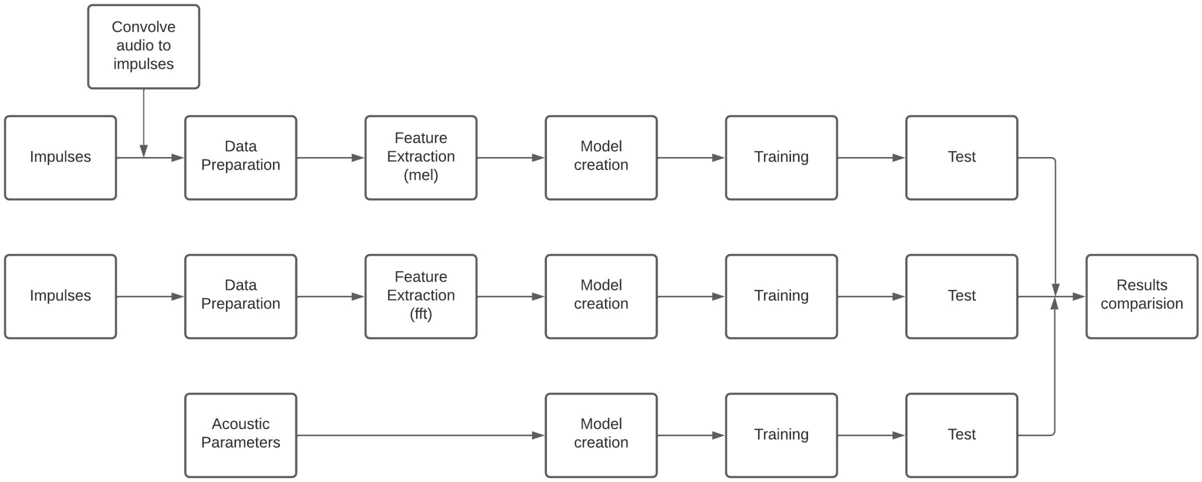


Figure 3.6: Original workflow proposed for the CNN classifier

Apart from that, changes were made to the core structure of the thesis itself. As seen in Fig. 3.6 there were originally three paths rather than the current two. The third path was meant to use pre-calculated acoustic parameters from the impulse recordings, including T60, EDT, t_s, C50, D50 and IACC as input.

However, while testing the working of the neural network on a subset of the input dataset, it was observed that the results of this classifier highly coincide with the results seen from Treybig et al. [80] in their paper where they calculated these acoustic parameters and analysed them. Given this similarity it was determined that a greater understanding could be obtained of the model's learning and the important acoustic features of a room by comparing the results from the CNN classifier to results obtained from subjective testing by Treybig et al. As such, combined with the difficulty of using tables of acoustic parameters as input and comparing that with audio input in the other two paths, the "Acoustic parameter" path was dropped in favour of comparision of results and LRP was added as an additional means of understanding the CNN decision.

3.6 Software Implementation and Coding

The code for this thesis was written entirely using Python 3.8 in a virtual environment in PyCharm 2022.1.3 (Community Edition) running on Ubuntu 20.04.4 LTS operating system. The relevant python libraries and their function in this classifier project are listed below.

- NumPy v1.22.1, SciPy v1.7.3 and Matplotlib v3.5.1

  Commonly used python libraries for mathematical operations, scientific operations and plotting.

- Seaborn v0.11.2

  Used to neatly present test results in the form of a Confusion Matrix.

- Scikit-learn v1.0.2

  Free machine learning library with tools for predictive data analysis and data mining.

- Librosa v0.8.1

  Python package for music and audio analysis. It provides many tools widely used in the field of Music Information Retrieval.

- Torch v1.11.0

  Python library for PyTorch which is an open source machine learning framework as described in 2.4. It provides high level tensor computation accelerated with GPU and framework for DNN.

- Torchaudio v0.11.0

  It is a library for audio and signal processing using PyTorch. It is an audio processing counterpart to the Torchvision library and was published by Yang et al. in [81].

- FFmpeg

  A widely used open source software library that contains tools for processing image, video and audio streams and multimedia, used in this project for presentation purposes.

Chapter 4

Experiment

The CNN classifier was trained and tested on a dataset of impulse recordings created at Technische Universität Ilmenau. The details about the dataset used as input is provided in Section 4.1. Information about the hardware that was used to run the neural network can be seen in Section 4.2, and the experiment is described in detail in Section 4.3.

4.1 Dataset

The dataset used in this thesis is created by Treybig et al. [80] containing 20,520 audio files. This dataset consists of impulse recordings from seven rooms of various types including a private living room, seminar rooms, listening lab, entrance foyer and a large auditorium. These recordings were taken using three separate recording setups: A Dummy Head (DH), Motion Tracked Binaural (MTB) microphone array and Spatial Decomposition Method (SDM) KEMAR dummy head. Five RL906 ME loudspeakers from Geithain were set around the recording setup in a standard 5.0 surround configuration with a 2m radius.

In total, there are seven rooms for which there are recordings. Each room acts as a class for the CNN classifier to pick between. Description of each room is provided below. With the exception of LR001, all of these rooms are located in TU Ilmenau.

- H1539b, Audio/Video Lab:

  It is a small room with acoustic optimization used as an audiovisual lab

- H1562, Helmholtz building foyer:

  It is the entrance foyer to the Helmholtz building in TU Ilmenau. It is a big open hall and is highly reverberant.

- H2505, Seminar room:

  It is a medium sized room used for seminars, and it is highly reverberant.

- HL, Hörlabor or Listening lab:

  It is a medium sized room that follows the room acoustic requirements of ITU-R BS.1116-3 standard

- HU103, Audimax auditorium:

  This is an auditorium called Audimax located in the Humboldt building in TU Ilmenau. It is a large, reverberant room like most auditoriums.

- LR001

  It is a private living room.

- ML2-102, Editorial room in Media Lab 2:

  It is a small room with acoustically optimized ceiling and two window fronts with medium reverberation.

Recordings were made in each room in three separate positions: middle of the room, adjacent to a wall and close to a corner, except for the living room where recordings were only made in the middle of the room as shown in Fig. 4.1. In each position, the setups were rotated, and recordings were made at every 5 degree rotation.

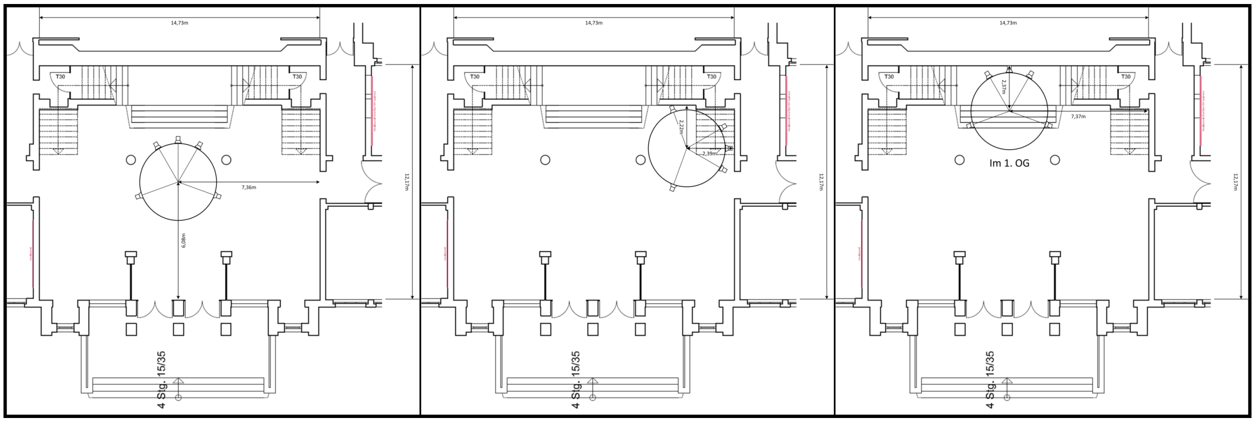


Figure 4.1: Floor plan of room H1562 showing the three recording positions

The audio files that are convolved with these impulse recordings come from the Lexicon Dry Tracks II dataset. From this dataset, five audio files were selected such as male and female vocal, vocal group, claps, television music. These five files have varied content and allow us to mimic speech and music signals when convolved with the impulse recordings. As these audio files are dry tracks, they do not bring their own room acoustic features. Thus convolving them with the impulse recordings which represent acoustic characteristics of a room, emulates this music and speech signal as having been recorded in one of the rooms from the TU Ilmenau dataset. This is known as convolutional reverb.

4.2 Hardware used

The hardware used to run the neural network training and testing is my personal laptop, namely a Lenovo Legion 5. The hardware specifications are shown below in Table 4.1.

	Model
Device	Lenovo Legion 5
CPU	AMD Ryzen™ 7 4800H Processor
RAM	16GB DDR4 3200 Hz
GPU	GeForce RTX 2060 with 6GB VRAM
CUDA	Version 11.4

Table 4.1: Hardware specifications

4.3 Experimental setup

The TU Ilmenau dataset contains 20,520 impulse recordings for use as input. For the “Convolved audio” path, the five dry audio tracks from the Lexicon dataset need to be convolved with every impulse recording. This means that though there are 20,520 input files for the “Impulses” path, there are 102,600 input files for the convolved audio path. This full set is divided into the Training and Test datasets by random division in a 7:3 ratio. The total size of the training and test sets for each path is given below in Table 4.2.

Path	Training set	Test set
Convolved audio	71,820	30,780
Impulses	14,364	6,156

Table 4.2: Size of the input dataset

Before training on the full dataset, the working of the CNN classifier was tested on smaller subsets of the input dataset. A subset was created with 1,026 impulses, and correspondingly 5,130 convolved audio files. It was a carefully crafted subset, while making sure the diversity of recording positions and speaker positions were kept intact. Further, to check the effect of having five times more input files for the "convolved audio" path than the "impulses" path, a subset was created for the convolved audio path such that there would be a similar number of input files of around 1000, which in this case was 1,140 inputs. The results from these initial tests showed high accuracy and they overlapped with the findings of Treybig et al. [80].

The first step in the experiment is to convolve the dry audio tracks with impulses for the "convolved audio" path and perform noise and silence padding for the "impulses" path. Then these inputs are cut to the desired lengths, 1.7s and 5s for convolved audio and 1.7s for impulse recordings. All of the inputs are sampled at 44.1 kHz sampling rate. The next step is feature extraction. For mel features, the following parameters are used:

- n_mels = 128
- n_fft = 2048
- hop length = 512

As the goal is to understand the effect of music/speech signals, impulses, duration and padding method, standard values were used for the melspectrogram parameters which were taken from librosa documentation for melspectrograms. Similarly, standard values also taken from librosa documentation were used for the FFT features as shown below:

- n_fft = 2048
- hop length = 512
- window length = 128

The original shape of each file is (no. of channels, audio length x sampling rate), and after extracting mel features, the shape of each mel feature file is (no. of channels, mel freq_bands (n_mels), time_steps). Time steps can be obtained as,

$$Timesteps = (sample\_rate/hop\_length) * duration \tag{4.1}$$

From this formula, we get mel features of shape (1, 128, 147) for 1.7s convolved audio and (1, 128, 431) for 5s convolved audio. For FFT features, the shape is (no. of channels, fft freq_band ($\frac{n\_fft}{2} + 1$), time_steps). Thus the shape of the FFT features is (1, 1025, 147) for impulses path. This bigger size led to the limitation of having a batch size of 8.

Once the feature extraction is done, the model training can be done. The training set is divided into a training and validation set in a 7:3 ratio by the dataloader. The CNN model is then trained for 50 epochs with a batch size of 8. The accuracy and loss plots are graphed and the model's weights and biases are saved for the test later.

After training, the training set is passed through the saved CNN model again to perform LRP. Then, model inference is run on the saved model with the Test set. During this testing, the accuracy, precision, recall and F1 scores are calculated and the prediction results are plotted in a confusion matrix for better understandability. The results of the training and testing will be shown in Chapter 5. The results are evaluated and analysis of the CNN model's decisions are provided in Chapter 6.

Chapter 5

Results

The results from the CNN classifier are presented in this chapter. The training was performed in stages. First, a subset of inputs was created containing 1k inputs to test the working of the CNN classifier. Once the coding stage of the thesis was completed and tested, model training was performed with the classifier on the full TU Ilmenau dataset. As such, this chapter has been split into two sections. Section 5.1 provides results from training on the smaller subset while Section 5.2 provides results from training on the full dataset. The results provided in this chapter are analysed and the findings of this thesis explained in Chapter 6.

5.1 Training results on subset

This subset contains 1,026 impulse recordings. Special care was taken in the preparation of this subset such that the total number of inputs is close to a standard 1000 inputs while maintaining the acoustic diversity of the dataset. That means that the inputs were sampled so that all recording positions and room positions are equally represented as they were in the TU Ilmenau dataset. Further, given there are 5 dry audio tracks to be convolved with each impulse recording, the “convolved audio” path has 5,130 input files. This section provides the tabulated and plotted test results for every different input type, namely convolved audio of 1.7s and 5s duration, noise and silence padded impulse recordings on this subset.

Further, in the beginning the following training parameters were used:

- batch size = 100
- no. of epochs = 50

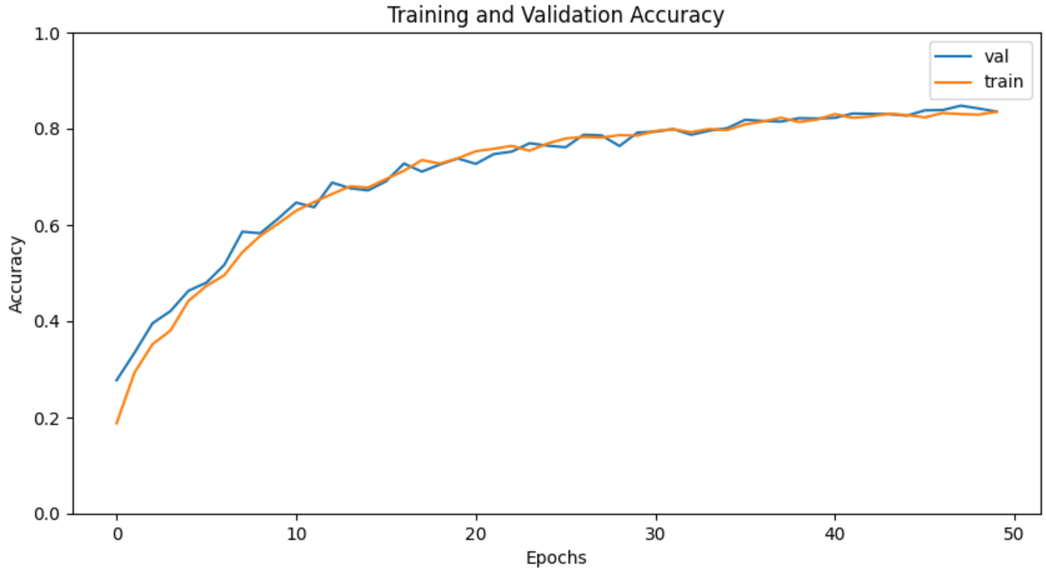
But as stated in Section 4.3, due to hardware limitations and the bigger size of the inputs in “impulses” path, the batch size was later limited to 8. While having different batch sizes adds an element of unbalance when comparing the training results, the results are still shown here as they allow us to see the effect that batch size has on training. Further, the training on these smaller subsets allow us to learn some specific behaviour that cannot be noticed on the full dataset trained model. It gives insight into the working on a CNN model that is in the middle of the learning process. These intermediate results are shown below, and the results of training on the full dataset is provided in Section 5.2.

5.1.1 Convolved audio path

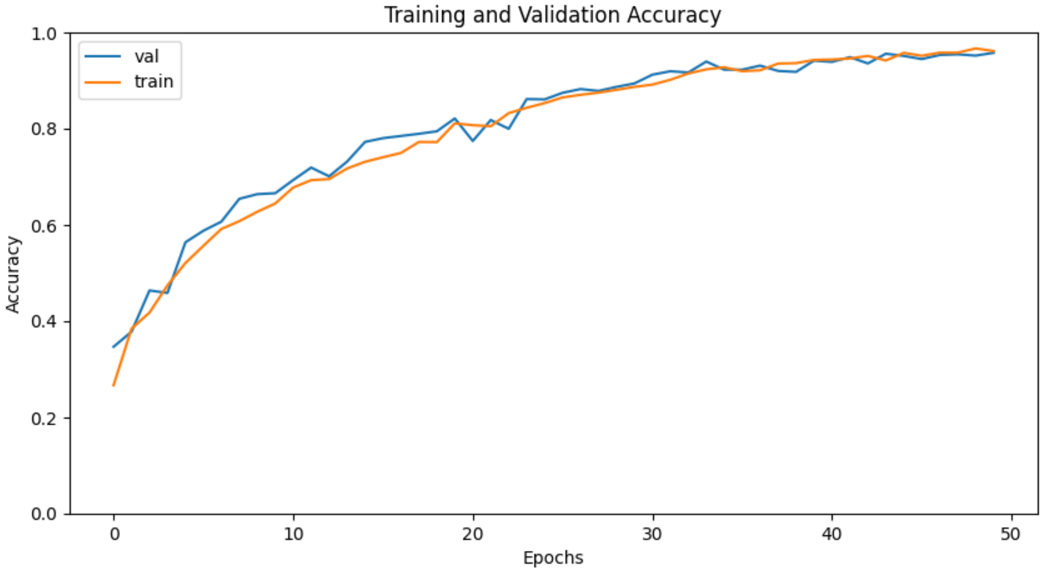
Training was initially performed for the 1.7s inputs at a batch size of 100. But after the batch size limitation was found, training was redone at batch size of 8. Figure 5.1 has these parameters:

- batch size = 100 vs batch size = 8
- no. of epochs = 50

Training results for 1.7s convolved audio input are as follows:



Accuracy curve for batch size = 100



Accuracy curve for batch size = 8

Figure 5.1: Training and validation accuracy curve for 1.7s convolved audio

- For batch size = 100, accuracy is 84%
- For batch size = 8, accuracy is 96%

The 5s convolved audio inputs were trained on a batch size of 8 for 50 epochs, and Fig. 5.2 shows training results for 5s audio.

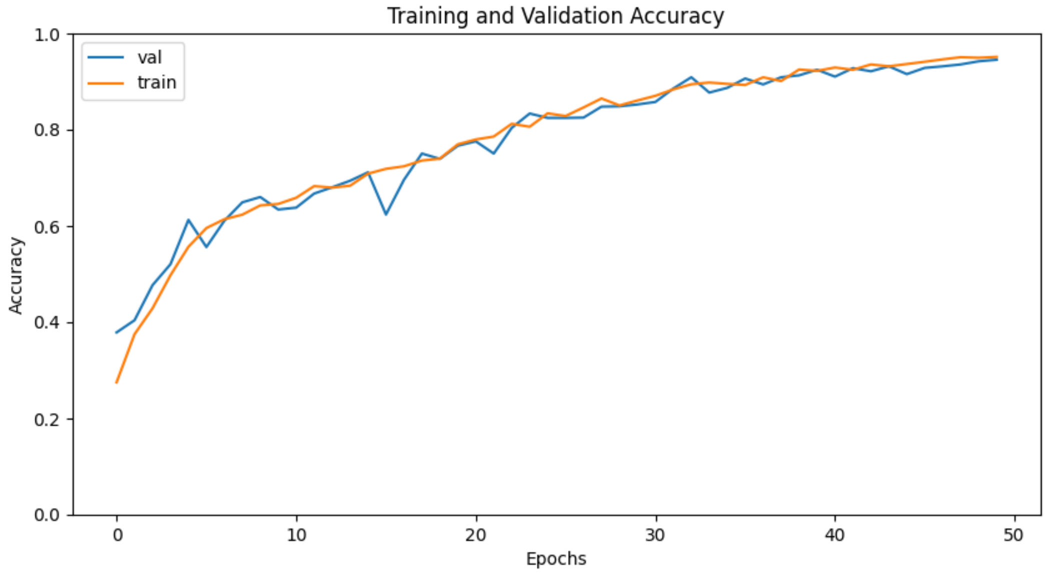
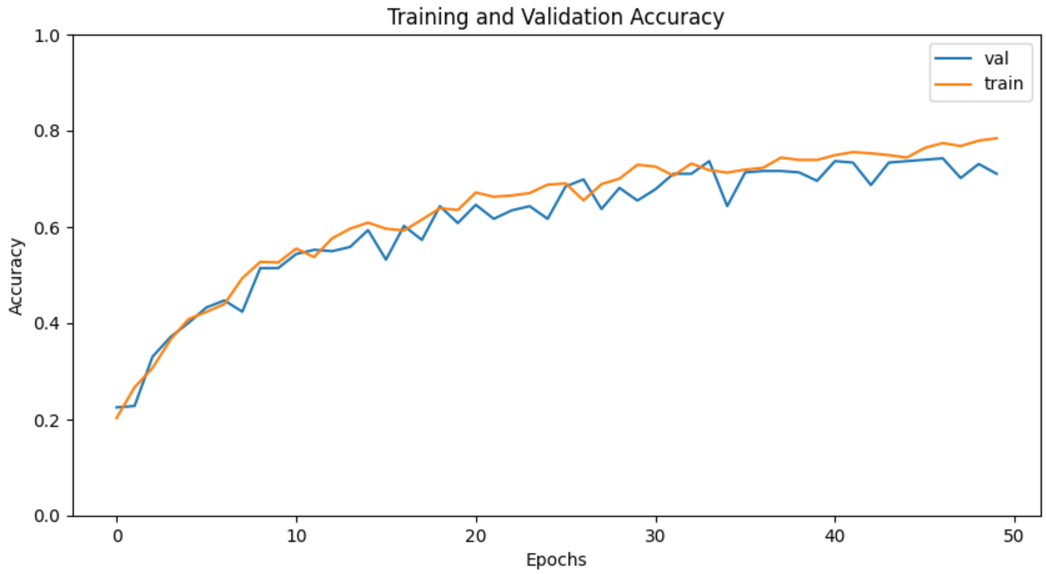


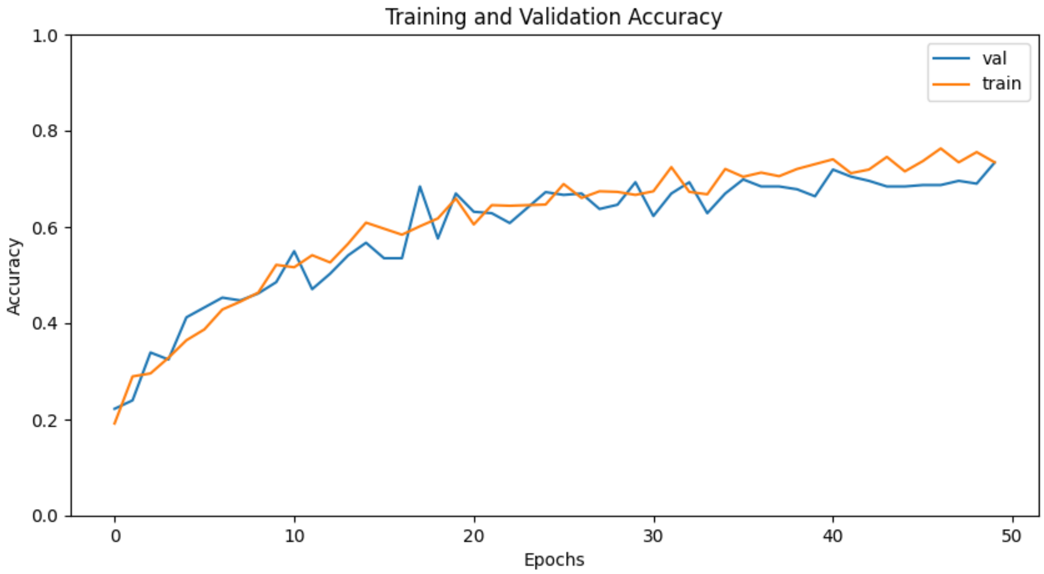
Figure 5.2: Training and validation accuracy curve for 5s convolved audio

The accuracy for 5s convolved audio is 95% compared to the 96% accuracy of 1.7s audio.

Due to the nature of the experiment, the "convolved audio" path has five times as many inputs as the "impulses" path. But to study the impact of the parity in number of inputs, a smaller subset was created for convolved audio path such that there were a similar number of inputs while still maintaining as much input diversity as possible. In this subset, there are a total of 1,140 inputs. Fig. 5.3 shows this result.



Accuracy curve for 1.7s audio

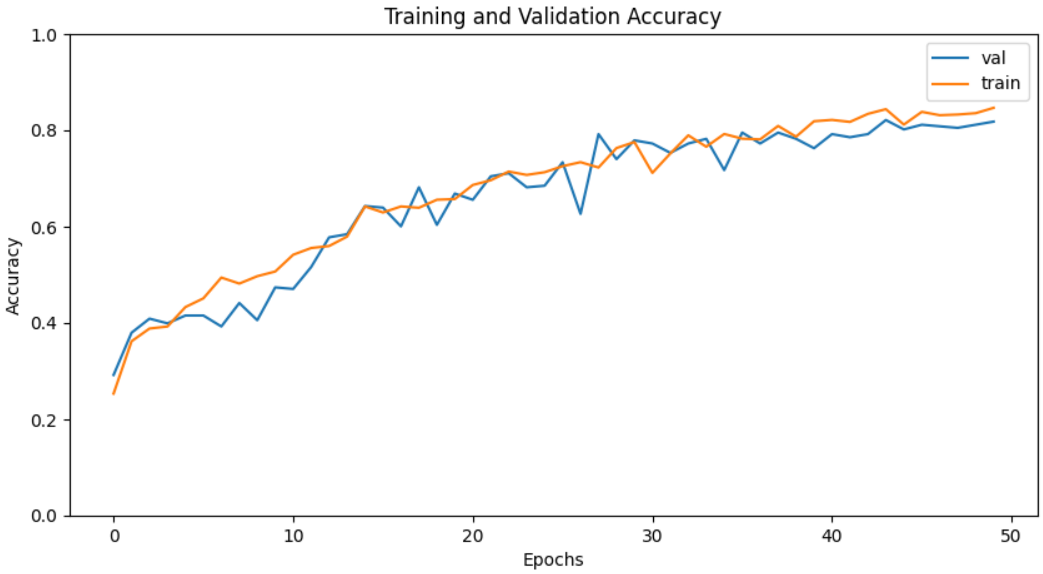


Accuracy curve for 5s audio

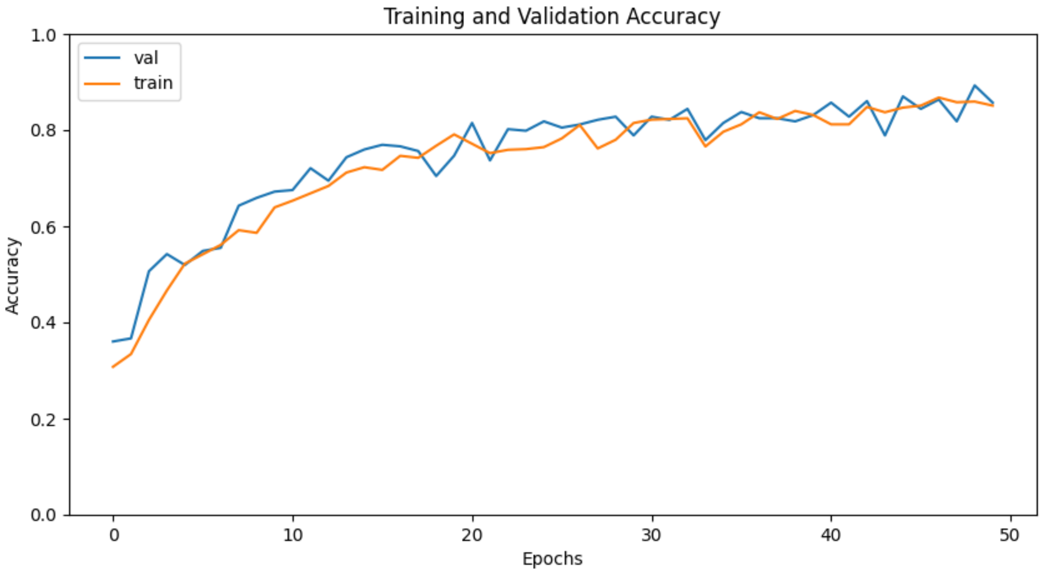
Figure 5.3: Accuracy curve for 1.7s and 5s convolved audio with 1k inputs

- For 1.7s audio, accuracy is 78%
- For 5s audio, accuracy is 73%

5.1.2 Impulses path



Accuracy curve for noise padding



Accuracy curve for silence padding

Figure 5.4: Accuracy curve for noise and silence padded impulses

Training for noise and silence padding was done with a batch size of 8 for 50 epochs and results are shown in Fig 5.4.

- For noise padded impulses, accuracy is 85%
- For silence padded impulses, accuracy is 85%

5.1.3 Tabulated results for subset

The tabulated results for this subset on all input types are provided in Table 5.1.

Trained model	Accuracy	Loss
1.7s audio, batch = 100	0.84	0.41
1.7s audio, batch = 8	0.96	0.14
5s audio, batch = 8	0.95	0.18
Noise padding, batch = 8	0.85	0.52
Silence padding, batch = 8	0.85	0.46

Table 5.1: Training results on subset

Further, training was done with 1k inputs for all input types as stated before. Results for training with equal number of inputs are provided in Table 5.2.

Trained model	Accuracy	Loss
1.7s audio, batch = 8	0.78	0.52
5s audio, batch = 8	0.73	0.63
Noise padding, batch = 8	0.85	0.52
Silence padding, batch = 8	0.85	0.46

Table 5.2: Training results on subset with equal inputs

5.2 Training results on full dataset

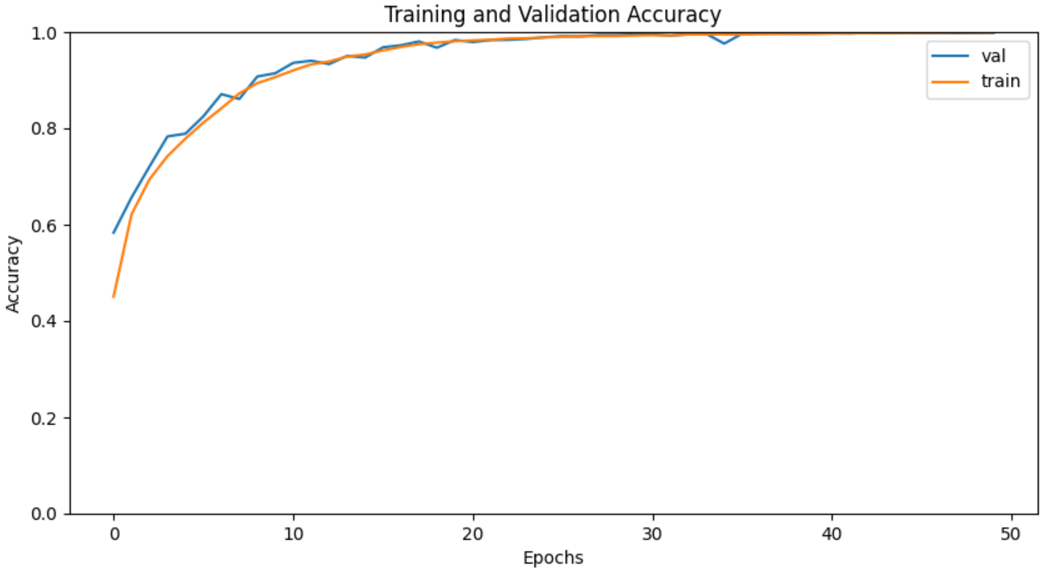
After the previously shown tests, training was performed on the full TU Ilmenau dataset. All four input types were trained with the same parameters. In this training session, no smaller subset was created for "convolved audio" path to match the size of the "impulses" path as the effect of an equal subset is already explored in Section 5.1.3. This section covers the results of training the CNN classifier on the full dataset. The following parameters were used during training:

- batch size = 8
- no. of epochs = 50
- no. of inputs are shown in Table 4.2

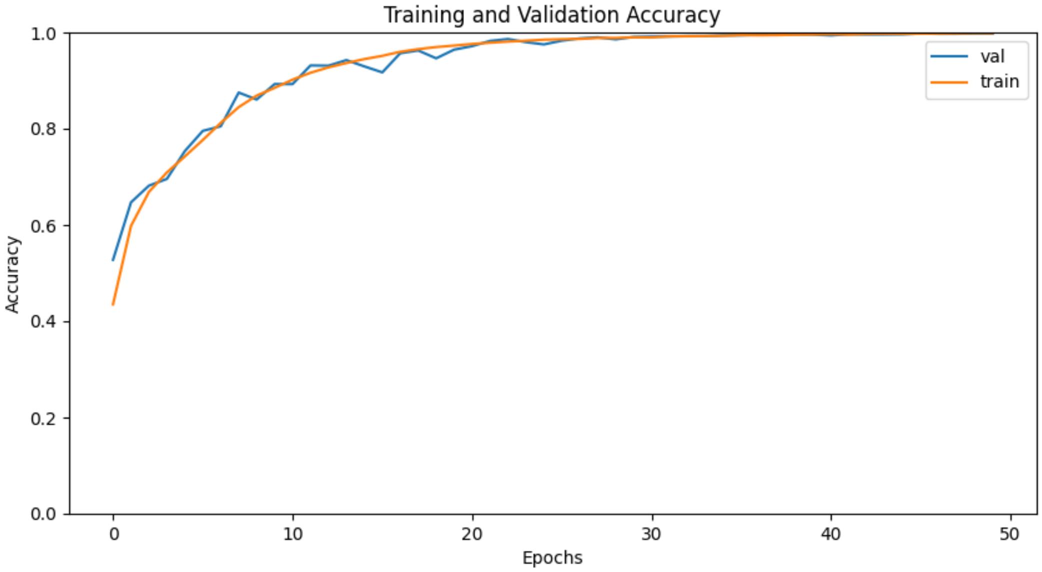
5.2.1 Convolved audio path

The training and validation accuracy for the "convolved audio" path is shown in Fig. 5.5.

- For 1.7s audio, accuracy is 100%
- For 5s audio, accuracy is 100%



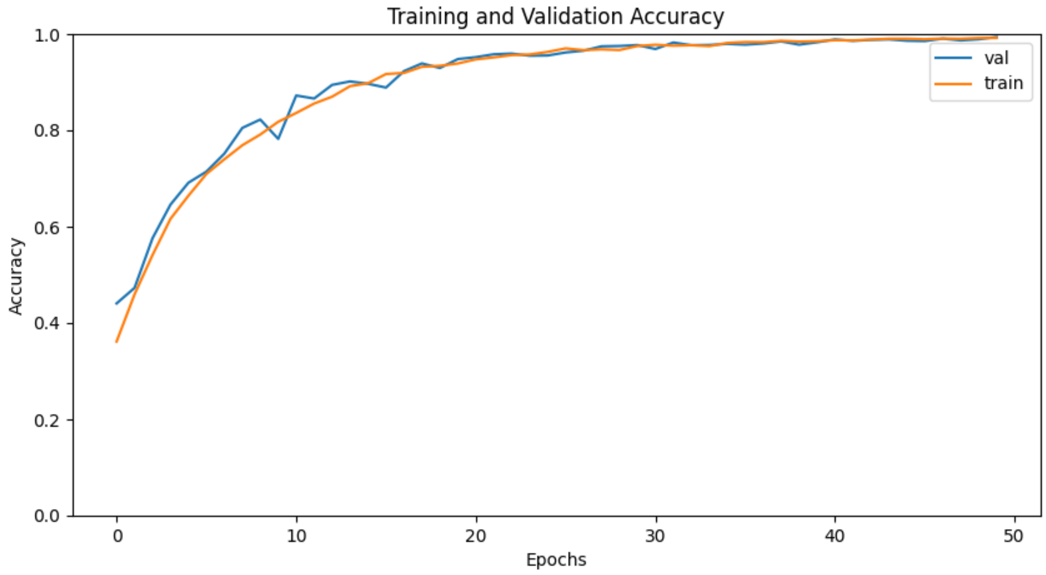
Accuracy curve for 1.7s audio



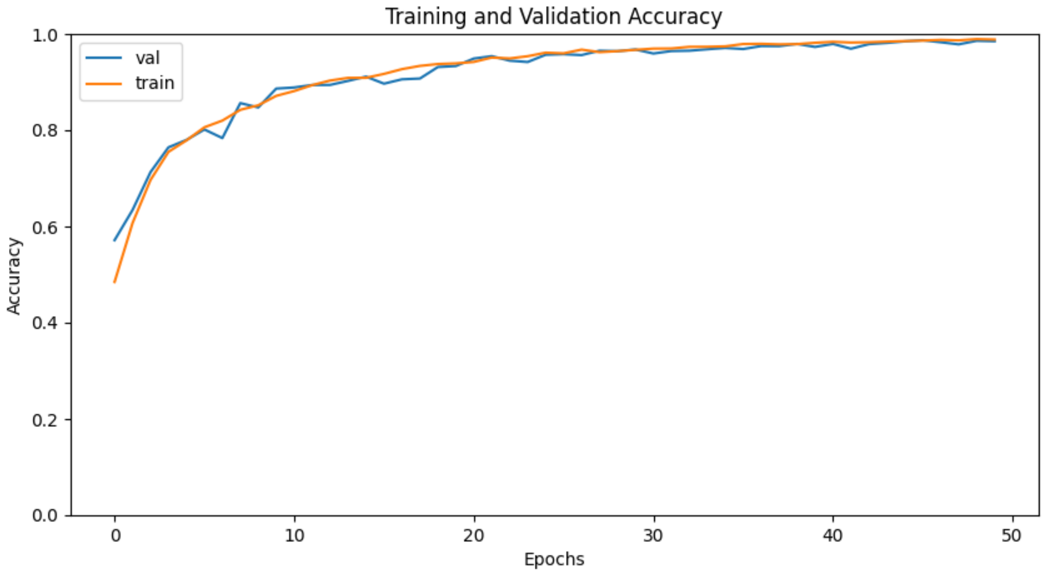
Accuracy curve for 5s audio

Figure 5.5: Accuracy curve for 1.7s and 5s convolved audio on full dataset

5.2.2 Impulses path



Accuracy curve for noise padding



Accuracy curve for silence padding

Figure 5.6: Accuracy curve for noise and silence padded impulses on full dataset

Fig. 5.6 shows the training and accuracy curve for the "impulses" path for noise and silence padded impulse recordings as inputs.

- For noise padded impulses, accuracy is 99%
- For silence padded impulses, accuracy is 99%

5.2.3 Tabulated results for full dataset

The results for the training on full dataset have been shown above in Sections 5.2.1-5.2.2. They are tabulated below for easier comparision in Table 5.3.

Trained model	Accuracy	Loss
1.7s audio, batch = 8	1.00	0.01
5s audio, batch = 8	1.00	0.01
Noise padding, batch = 8	0.99	0.04
Silence padding, batch = 8	0.99	0.05

Table 5.3: Training results on full dataset

The time required for training was 93 minutes for 1.7s audio and 347 minutes for 5s audio in the "convolved audio" path and 135 minutes for noise padding and 137 minutes for silence padding in the "impulses" path.

5.2.4 Layer-wise Relevance Propagation results

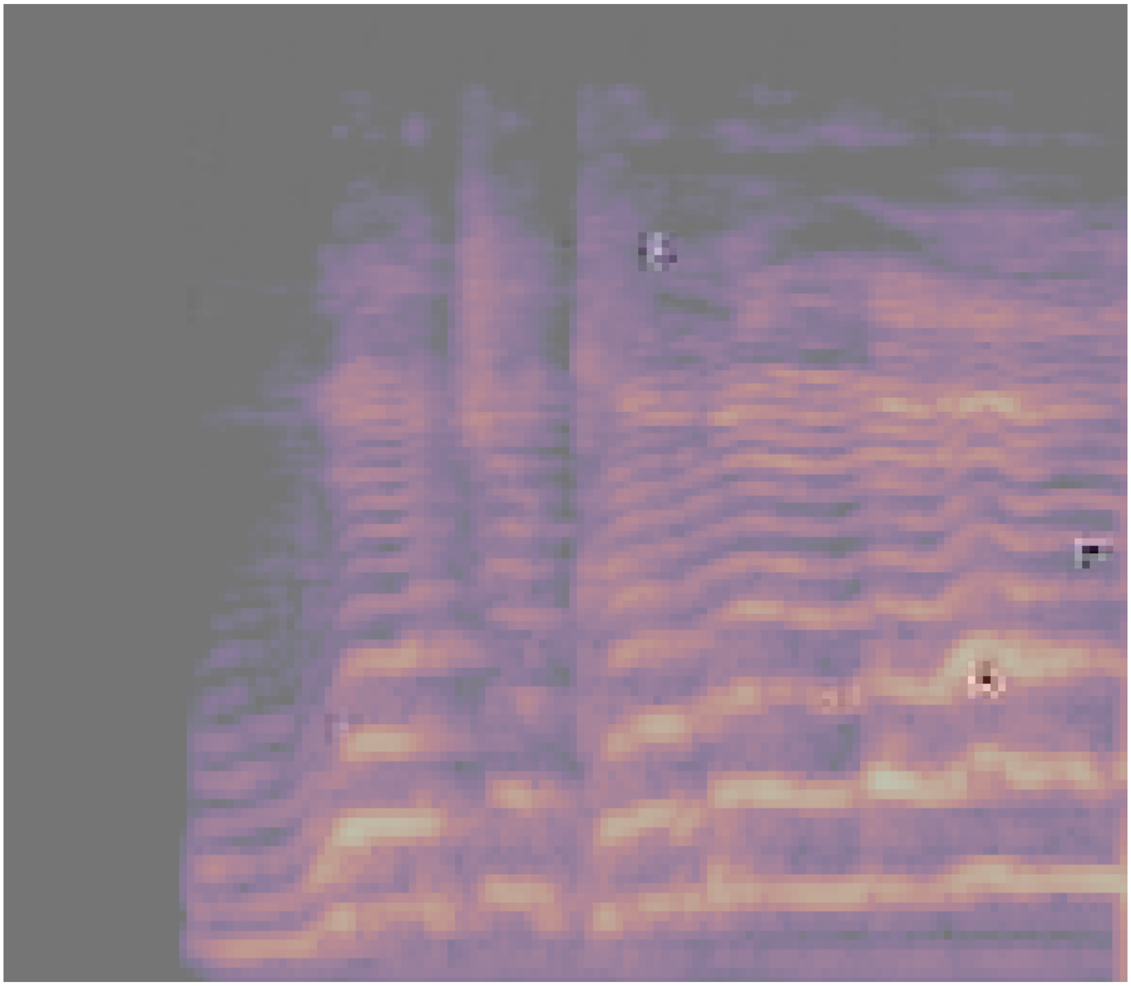


Figure 5.7: LRP output on impulse recording convolved with male vocal track

Fig 5.7 shows the output of LRP as performed on an input from the “convolved audio” path. It can be clearly seen that the highlighted portions that signify high relevance to the CNN model coincide with some regions of change or boundaries in the melspectrogram. However, this output does not provide enough information unlike in the case of computer vision tasks as shown in Fig. 2.3 where LRP shows a clear image of the relevant portions.

Chapter 6

Evaluation

This chapter goes into detail of the assessment and evaluation of the results as shown in Chapter 5. After training is performed, the trained model and its weights and biases are saved. The test set described in Section 4.3 is then passed through the trained CNN classifier by utilizing the saved model weights and biases. The result of this classification test on the test set is shown in the form of a confusion matrix for better visualization of the test results. These confusion matrices are examined in this chapter to understand the “black box” decisions of the model, and as seen later in this chapter, the inaccurate classifications that are seen in the confusion matrix give important and meaningful insight into the thinking and decision making of the CNN classifier.

Section 6.1 evaluates the performance and adaptation of LRP in this thesis, and comments on the viability of LRP in the field of audio processing. Section 6.2 provides the test results in the form of confusion matrices and analyses the results. As a measure to evaluate results, the accuracy, precision, recall and F1 scores are also calculated. This section also summarises the takeaways and learning from this thesis.

6.1 LRP

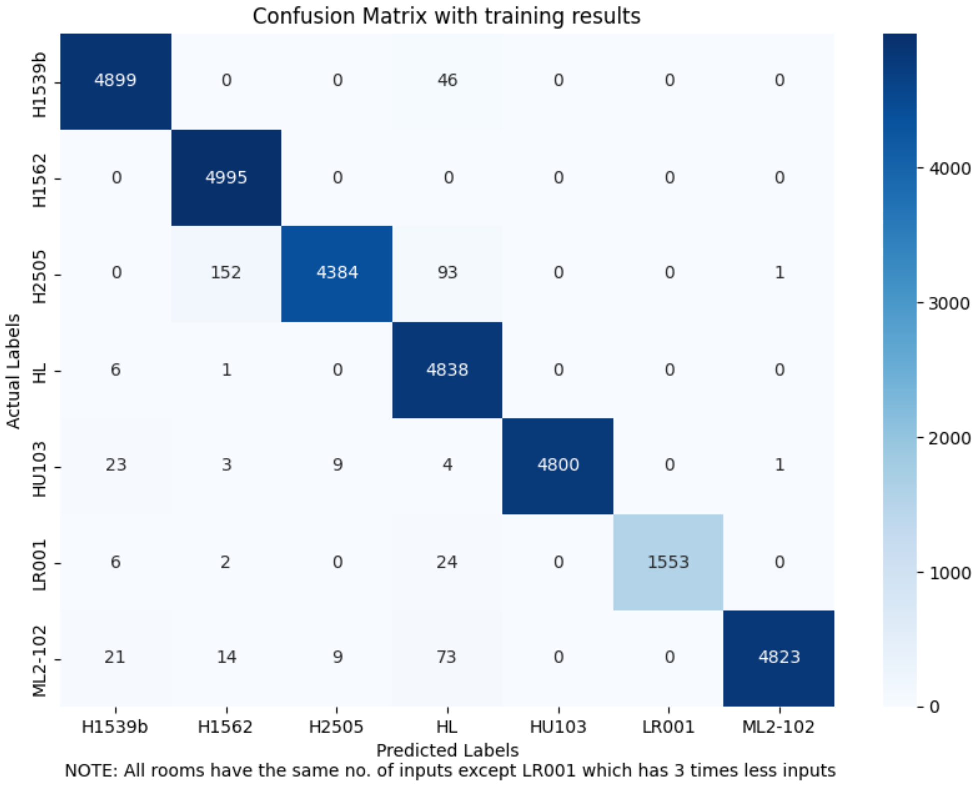
LRP as stated in Section 2.5.2 is a process introduced for computer vision tasks. This thesis attempts to adapt it to audio classification tasks with the basic idea that audio processing using Deep Learning is an extension of image processing. As can be seen in Fig. 2.3, in image classification tasks the relevant sections of the image are the visual boundaries of the object being classified. This means a neural network sees the outline of a castle or cat or any other object that is a classification target as the most relevant section of the image.

As seen in Fig. 5.7, the output of LRP for room classification based on a spectrogram as input does not show as extensive a region highlighted as in Fig. 2.3 which is for an image classification task. In an image the boundary of an object can be clearly seen as the relevant section for classifying the object, however it is not as clear cut for classifying audio from a spectrogram. In a spectrogram, there is no clear visual indicator that can aid in a classification task. While domain expertise can aid in identifying audio features from a spectrogram such as spoken words, differentiating vocals and instruments etc, it is harder to visually tell which room the acoustic characteristics belong to. Due to this basic difference in the nature of image and audio, LRP loses some of the immediate applicability even if audio classification is a similar process to image classification. In a spectrogram, every point on the graph is meaningfully connected to other points on the spectrogram. As such, relevance of a point is also tied to others unlike in an image where a pixel does not have any relation to the value of the neighbouring pixels. Despite this, Fig. 5.7 shows a few relevant sections that coincide with important boundary regions and provides some information on relevance in a spectrogram. In the future, this can be built upon to adapt LRP to be more suited to audio classification tasks and take into account the nature of spectrograms.

6.2 Test results

Finally, this section evaluates the results from the CNN classifier on the test set. Section 6.2.1 analyses the results from the "convolved audio" path while results for the "impulses" path are provided in Section 6.2.2.

6.2.1 Convolved audio path


Confusion Matrix with training results

NOTE: All rooms have the same no. of inputs except LR001 which has 3 times less inputs

Figure 6.1: Confusion matrix for 1.7s convolved audio

Fig 6.1 shows the test results for 1.7s convolved audio inputs as a confusion matrix.

For 1.7s convolved audio input:

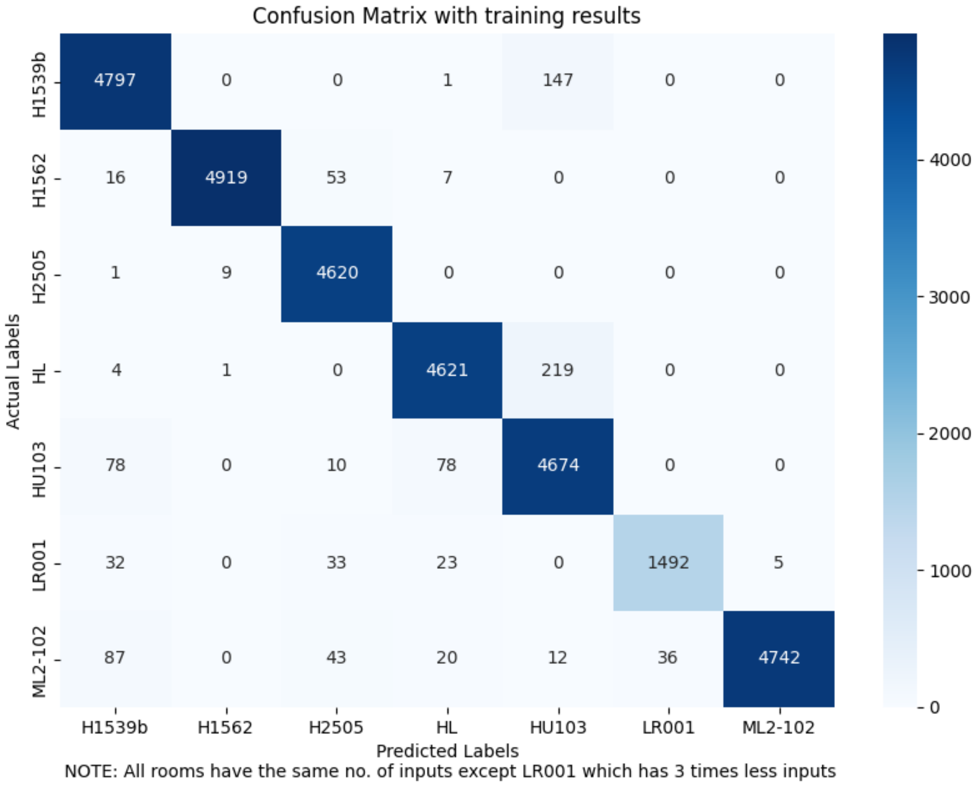
- Accuracy: 0.98, Precision: 0.98, Recall: 0.98, and F1 score: 0.98

Given the high accuracy of the training and tests, the accurate classifications do not tell us much more than the fact that the CNN classifier performs well at the room classification task. Instead, much more is learned about the thinking and decision making of the classifier from the inaccurate predictions. From Fig. 6.1, it can be seen that the only major inaccuracy is that room H2505 being mistaken for room H1562. From Section 4.1, it can be seen that they are both highly reverberant rooms. Keeping in mind that the “convolved audio” path is meant to emulate real-world speech and music signals, this lines up with the findings of [36] [37], where it is seen that reverberation time is the acoustic parameter that has the most measurable effect on music.

Fig. 6.2 shows the confusion matrix of test results for the 5s convolved audio inputs. For 5s convolved audio input:

- Accuracy: 0.97, Precision: 0.97, Recall: 0.97, and F1 score: 0.97

It is seen from that accuracy is slightly less for the 5s convolved audio inputs than inputs of 1.7s duration. This lines up with the findings from training on smaller subsets as well as seen in Tab. 5.1-5.2 as well as training on the full dataset as seen in Tab. 5.3.



Confusion Matrix with training results

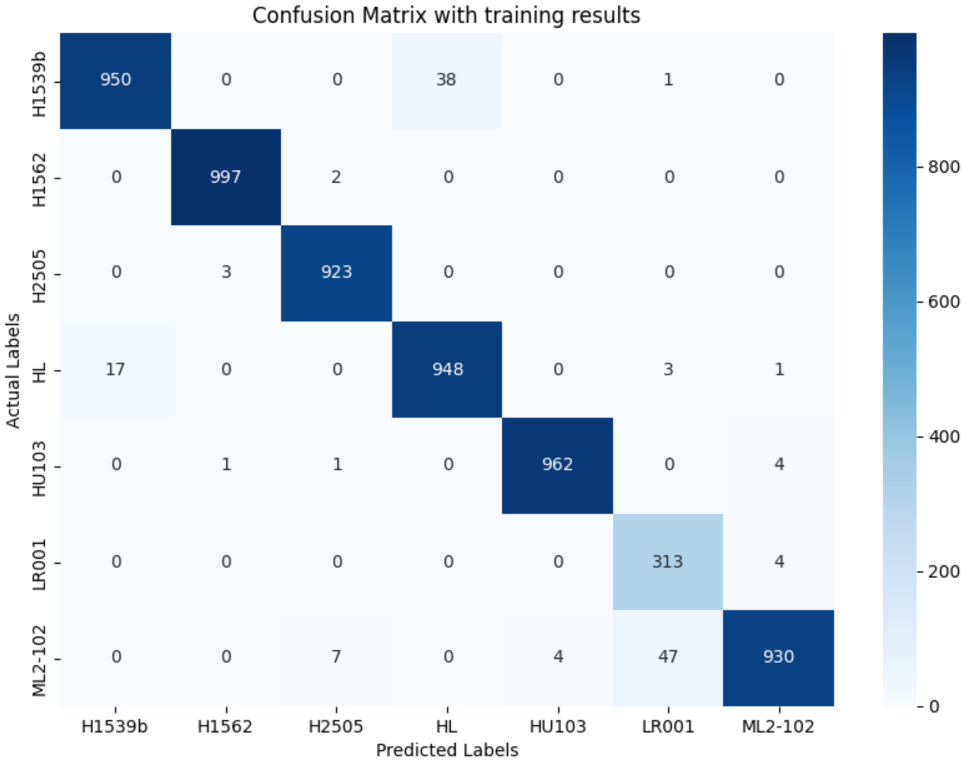
Actual Labels \ Predicted Labels	H1539b	H1562	H2505	HL	HU103	LR001	ML2-102
H1539b	4797	0	0	1	147	0	0
H1562	16	4919	53	7	0	0	0
H2505	1	9	4620	0	0	0	0
HL	4	1	0	4621	219	0	0
HU103	78	0	10	78	4674	0	0
LR001	32	0	33	23	0	1492	5
ML2-102	87	0	43	20	12	36	4742

NOTE: All rooms have the same no. of inputs except LR001 which has 3 times less inputs

Figure 6.2: Confusion matrix for 5s convolved audio

From Fig. 6.2 it can be seen that the major inaccuracies in classification is rooms H1539b and HL being mistaken for room HU103. From Section 4.1, it is seen that all three rooms are similar in that they are acoustically optimized.

6.2.2 Impulses path



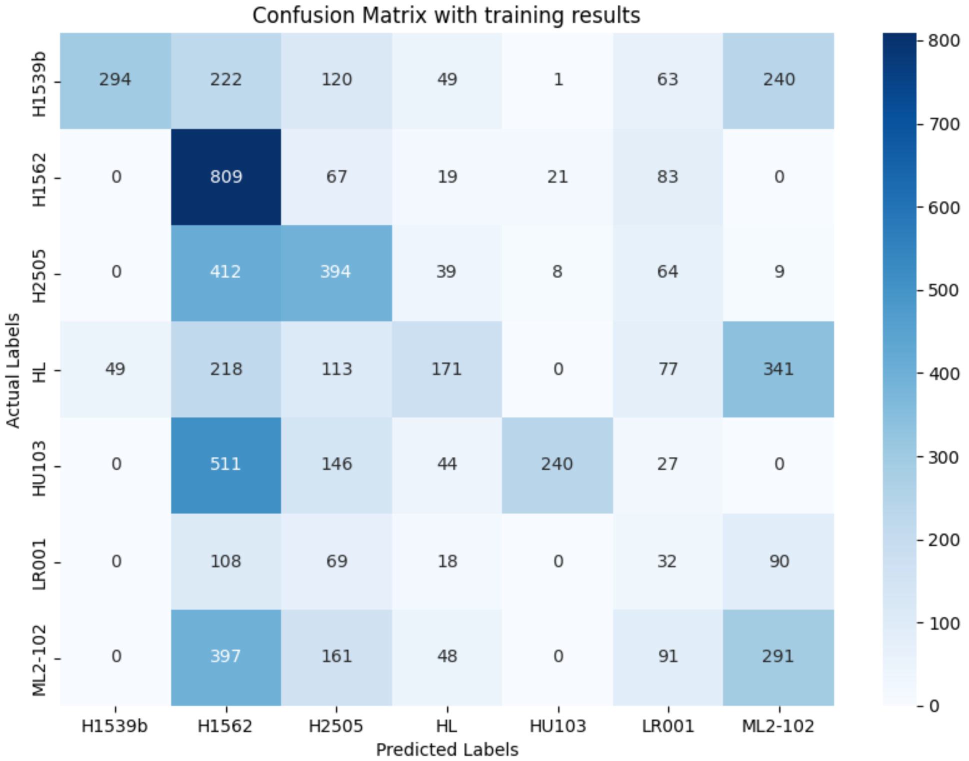
NOTE: All rooms have the same no. of inputs except LR001 which has 3 times less inputs

Figure 6.3: Confusion matrix for noise padded impulses

Fig. 6.3 shows the test results for noise padded impulses. As shown in Table 4.2 and mentioned before, the "impulses" path has less inputs than the "convolved audio" path so it needs to be kept in mind there are less total number of inputs tested for this path. For the noise padded impulses input:

- Accuracy: 0.98, Precision: 0.98, Recall: 0.98, and F1 score: 0.98

From Fig. 6.3 it is seen that the main inaccuracy in noise padded inputs is that the rooms H1539b and HL get mistaken for each other. Another trend is that the rooms LR001 and ML2-102 get mistaken for each other too. This relationship between the two pairs can also be seen in the findings of Treybig et al. [80]. They perform Linear Discriminant Analysis (LDA) which separates rooms on the basis of room acoustic parameters. This shows the similarity or dissimilarity between these rooms. Treybig et al. find that rooms H1539b and HL are mapped in the same region by the LDA analysis and overlap almost entirely. Rooms LR001 and ML2-102 are also mapped very closely with significant overlap. Thus it is seen that Fig. 6.3 lines up very well with the findings of Treybig et al.



NOTE: All rooms have the same no. of inputs except LR001 which has 3 times less inputs

Figure 6.4: Confusion matrix for silence padded impulses

Fig. 6.4 shows the test results for silence padded inputs, and they have the following results:

- Accuracy: 0.34, Precision: 0.36, Recall: 0.36, and F1 score: 0.36

Out of all the different inputs used for classification, the silence padded impulse input is the only one that does not show good results. It has an accuracy of only 34% and it can be seen that there is no trend or reasoning behind the inaccurate decisions unlike in the other input types. The only exception is that it can be seen in Fig. 6.4 that room H1562 is almost always predicted correctly, while every other room has many inaccurate classifications. The difference between room H1562 and the other rooms is that room H1562 is the only room which has an impulse recording longer than the duration of 1.7s, which is taken as the audio length for the "impulses" path. As such, H1562 is the only room where the entirety of the input is purely impulse response without any silent segment. For every other room, only a part of the input contains information about the room acoustic characteristics while the rest is silence. That is clearly insufficient for the room classification task and as such silence padding is not a viable option in a situation where there is a large variance in the length of the RIR or BRIR.

6.3 Analysis

From Chapters 5 and 6 the results of the room classification task can be seen. The overall verdict is a success at classifying rooms based on their acoustic parameters. Training was performed on increasingly larger sets of inputs from the TU Ilmenau dataset as seen in Sections 5.1-5.2. This provides insight into the learning process of the CNN classifier. As seen most notably in the "convolved audio" path, having been trained with 1k, 5k and then the full dataset of 72k inputs, the training accuracy increases in a logarithmic growth. Further by comparing the accuracy shown in Tables 5.1 and 5.2, it can be seen that there is a difference in accuracy of 1% between the two audio lengths of 1.7s and 5s in the convolved audio path for 5k inputs and the full dataset, but a difference of 5% for 1k inputs. From this, it can be surmised that the length of audio has greater impact for a smaller dataset but as we get more inputs, the accuracy for the two converges. Further, a longer input would mean bigger size of input to process. As such, a short 1-2 second input provides the sweet spot for accuracy as well as lesser training time.

When all parameters such as no. of inputs and batch size are kept the same, Table 5.2 shows that the impulses input format provides the highest accuracy followed by the convolved audio input where a longer duration leads to even lesser accuracy. However, as seen from Table 5.3, that only applies to a smaller dataset. For a large dataset with tens of thousands of inputs, the accuracy is nearly identical for all input methods. So, the input format plays a bigger role with a limited dataset.

Section 6.2.2 shows that silence padding is not a viable option when there is a large variance in the length of the impulse response. On the other hand, noise padding provides promising results, as does the convolved audio input. Section 6.2 shows that music/speech signals and impulse recordings (with noise padding if required) are both viable input formats. The CNN classifier provides good results and the classification decisions line up with findings from other papers. Results for the convolved audio path, which emulates music and speech signals itself, show similarity to [36] [37] who worked on music signals. Treybig et al. [80] created and worked on the TU Ilmenau dataset of impulse recordings, and their LDA analysis shows the same connections between rooms as seen in the convolved audio tests in Fig. 6.1.

A lot of the insights into the decision making of the CNN architecture gained in this thesis came in the middle of the learning process for the CNN classifier and from the inaccurate predictions. A large part of the inaccurate predictions occurred in situations where human perception would also find similarities. The findings of this thesis can be summarized as follows:

- As is mostly the case, there is no blanket best method for classification of room acoustic parameters.
- Music and speech signals, and impulse recordings are both viable input formats for a large dataset. However, impulse recordings are the superior input format for limited data.
- A short duration of 1-2 seconds is the sweet spot for music and speech inputs.
- For impulse recordings, padding with white noise created from the input itself provides great results. Padding with silent segments is not a viable option.
- When using music or speech signals, there is a higher chance of inaccuracies if all the rooms are highly reverberant. Reverberation time is the acoustic parameter that has the most measurable effect on music. As such, if all rooms are highly reverberant and there is no difference to the reverberation time between the rooms, then using impulse recordings as input is the better option.
- Using impulse recordings for training provides highly accurate classification. The results for this input format line up closely to human perception.
- The acoustic parameter that makes the biggest impact to the classification is reverberation time.

- For impulse response inputs, no one acoustic parameters affects classification more than others. The inaccuracies occur mostly when rooms are acoustically too similar as a whole.
- For music and speech signal inputs however, reverberation time plays a big role. Rooms with similar reverberation can get mistaken for one another.

Chapter 7

Conclusion

Room acoustics is the study of the behaviour of sound in an enclosed or semi-enclosed space. By understanding the behaviour of sound, room design can be made efficient and more suitable to various tasks. A room's acoustic characteristics also play an important role in the audio perception of the listener. Such acoustic characteristics can be measured by the RIR or BRIR of the room, which are unique for different rooms as well as for the positions of speakers and receivers.

This thesis aims to use MLA and Deep learning to create a CNN architecture for classifying rooms based on prerecorded BRIR impulse recordings. It also aims to determine the acoustic parameters that play the biggest role in differentiating between the acoustic characteristics of various rooms. For this end, a dataset of impulse recordings from Technische Universität Ilmenau, created by Treybig et al. [80], is used as shown in Section 4.1. It contains BRIR recordings from seven rooms, made on the basis of sweep recordings in three different positions in the room totaling 20,520 audio files. Another aim of this thesis is to interpret and explain the "black box" decision making of the neural network.

Chapter 3 provides details about the inputs used for the CNN classifier. Impulse recordings from the TU Ilmenau dataset are convolved with dry audio tracks from the Lexicon Dry Tracks II dataset to create a new input format that emulates music and speech signals. Section 3.1 goes into detail about the data augmentation and feature extraction methods for the "convolved audio" path and Section 3.2 does the same for the "impulses" path. Melspectrogram features are used as input for the convolved audio path while FFT magnitude spectrogram features are used for the impulses path.

A CNN architecture is used with four convolutional layers, with ReLU activation and batch normalization at each layer as shown in Section 3.3. Section 3.4 provides detail into the implementation of Layer-wise Relevance Propagation (LRP), which is used to find the sections of the input that are relevant to the CNN classifier. Chapter 4 shows how the experiment was conducted.

Chapter 5 provide results of the classification in graphical form, and the results are tabulated in Sections 5.1.3 and 5.2.3 for the subsets and full dataset respectively. Training results show high accuracy for all input formats. The trained CNN model is then tested with the test dataset. Chapter 6 provides the test results and presents them in the form of Confusion Matrices for better visualisation. Section 6.3 analyses the training and test results shown in Chapters 5 and 6. It is seen that the classification results line up with findings of other published papers. The results from convolved audio path, which emulates music and speech signals, show similarity to [36] [37] who worked on music signals. Similarly the results from impulses path line up with the findings of Treybig et al. [80] who created and worked on the TU Ilmenau dataset.

While LRP did not provide a lot of visual aid in explaining the decision making of the CNN classifier, a lot of insight was gained into the decision making of the model from the test results and the few inaccurate classifications. The main findings of this thesis are that music and speech signals, and impulse recordings are both viable input formats when dealing with a large dataset, though impulse recordings are a superior input format when training on limited data. For music and speech signals, a short 1-2 second duration provides the sweet spot for accuracy and required training time. For impulse recordings it is seen that padding with white noise created from the input itself show great results while silence padding is not a viable option. Training performed on BRIR impulse recordings show higher accuracy when all other factors such as no. of inputs and batch size etc. are kept equal. Ultimately, it is seen that Reverberation Time is the acoustic parameter with the most visible influence on the room classification task. No individual acoustic parameter affects classification results more than others and inaccuracies are a result of rooms being acoustically similar as a whole. But in the case of music and speech signals, reverberation time plays a big role. Rooms with similar reverberation can get mistaken for one another.

7.1 Future Work

Future improvements will include training and testing on a more diverse input set to make the CNN model more robust and reduce the chance of overfitting. A lot of the scrapped ideas and challenges faced in this thesis can provide promising results and insights if fully explored. Training a neural network using separate room acoustic parameters such as reverberation time, EDT, IACC, C50, D50, centre time etc. can provide a more detailed insight into the influence of each acoustic parameter on the room classification task. Further, LRP can be adapted to be more suitable to audio classification and be more sensitive to the nature of audio signals.

The CNN architecture used in this thesis is a fairly simple model. It can be tweaked with dropout mechanism and the layers adjusted for a more robust network. Different pooling layers such as sum-pooling or average-pooling (such as the Adaptive average pooling layer used in this thesis) can be compared for a balance of model performance and explainability of decisions.

Bibliography

[1] D. Hortvik, “Basic principles of room acoustics.” Online, accessed: 24.06.2022.

[2] J. Fröhlich and I. Wachsmuth, “Acoustically enriched virtual worlds with minimum effort,” in *2012 IEEE Virtual Reality Workshops (VRW)*, pp. 147–148, 2012.

[3] G. Montavon, W. Samek, and K.-R. Müller, “Methods for interpreting and understanding deep neural networks,” *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[4] K. Fabi, “Layer-wise relevance propagation for pytorch.” `https://github.com/KaiFabi/PyTorchRelevancePropagation`, 2021.

[5] D. Masovic, “Room acoustics (lecture notes),” 2021.

[6] D. Dziwis, S. Zimmermann, T. Lübeck, J. M. Arend, D. Bau, and C. Pörschmann, “Machine learning-based room classification for selecting binaural room impulse responses in augmented reality applications,” in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–8, 2021.

[7] D. Ward, E. Lehmann, and R. Williamson, “Particle filtering algorithms for tracking an acoustic source in a reverberant environment,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, 2003.

[8] R. Zelinski, “A microphone array with adaptive post-filtering for noise reduction in reverberant rooms,” in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pp. 2578–2581 vol.5, 1988.

[9] Y. Kaneda and J. Ohga, “Adaptive microphone-array system for noise reduction,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 6, pp. 1391–1400, 1986.

[10] N. Peters, H. Lei, and G. Friedland, “Name that room: Room identification using acoustic features in a recording,” pp. 841–844, 10 2012.

[11] A. H. Moore, P. A. Naylor, and M. Brookes, “Room identification using frequency dependence of spectral decay statistics,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6902–6906, 2018.

[12] A. H. Moore, M. Brookes, and P. A. Naylor, “Roomprints for forensic audio applications,” in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4, 2013.

[13] C. Papayiannis, C. Evers, and P. A. Naylor, “Discriminative feature domains for reverberant acoustic environments,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 756–760, 2017.

[14] T. Cox, F. Li, and P. Darlington, “Extracting room reverberation time from speech using artificial neural networks,” *Journal of the Audio Engineering Society*, vol. 49, pp. 219–230, 04 2001.

[15] F. Xiong, S. Goetze, B. Kollmeier, and B. T. Meyer, “Joint estimation of reverberation time and early-to-late reverberation ratio from single-channel speech signals,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 255–267, 2019.

[16] C. Papayiannis, C. Evers, and P. A. Naylor, “End-to-end classification of reverberant rooms using dnns,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3010–3017, 2020.

[17] Z. Tang, N. J. Bryan, D. Li, T. R. Langlois, and D. Manocha, “Scene-aware audio rendering via deep acoustic analysis,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 5, pp. 1991–2001, 2020.

[18] N. J. Bryan, “Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2020.

[19] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, “A summary on acoustic room divergence and its effect on externalization of auditory events,” in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2016.

[20] J. Huopaniemi, L. Savioja, T. Lokki, and R. Väänänen, “Virtual acoustics — applications and technology trends,” in *2000 10th European Signal Processing Conference*, pp. 1–8, 2000.

[21] R. T. Beyer, *Sounds of our times, two hundred years of acoustics.* Springer New York, NY, 1999.

[22] W. Von Kempelen, *Der Mechanismus der menschlichen Sprache. / The Mechanism of Human Speech.* 1791.

[23] S. Foster, E. Wenzel, and R. Tayior, “Real time synthesis of complex acoustic environments,” in *Final Program and Paper Summaries 1991 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 47–48, 1991.

[24] E. M. Wenzel, F. L. Wightman, and S. H. Foster, “A virtual display system for conveying three-dimensional acoustic information,” *Proceedings of the Human Factors Society Annual Meeting*, vol. 32, no. 2, pp. 86–90, 1988.

[25] J. Grosse and S. van de Par, “Perceptually accurate reproduction of recorded sound fields in a reverberant room using spatially distributed loudspeakers,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 867–880, 2015.

[26] A. J. Berkhout, “a holographic approach to acoustic control,” *journal of the audio engineering society*, vol. 36, pp. 977–995, december 1988.

[27] M. A. Gerzon, “periphony: with-height sound reproduction,” *journal of the audio engineering society*, vol. 21, pp. 2–10, february 1973.

[28] O. Young, “Creating a convolution reverberation effect from impulse responses in physical spaces,” 2020-06-24.

[29] M. R. Schroeder, “statistical parameters of the frequency response curves of large rooms,” *journal of the audio engineering society*, vol. 35, pp. 299–306, may 1987.

[30] C. Hak and R. Wenmaekers, “The impact of sound control room acoustics on the perceived acoustics of a diffuse field recording,” *WSEAS Transactions on Signal Processing*, vol. 6, 10 2010.

[31] S. Spors, H. Buchner, and R. Rabenstien, “A novel approach to active listening room compensation for wave field synthesis using wave-domain adaptive filtering,” in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. iv–iv, 2004.

[32] J. O. Jungmann, R. Mazur, M. Kallinger, and A. Mertins, “Robust combined crosstalk cancellation and listening-room compensation,” in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 9–12, 2011.

[33] D. de Vries and E. M. Hulsebos, “Auralization of room acoustics by wave field synthesis based on array measurements of impulse responses,” in *2004 12th European Signal Processing Conference*, pp. 1377–1380, 2004.

[34] B. Katz, D. Murphy, and A. Farina, *The Past Has Ears (PHE): XR Explorations of Acoustic Spaces as Cultural Heritage*, pp. 91–98. 08 2020.

[35] N. Eley, S. Mullins, P. Stitt, and B. F. Katz, “Virtual notre-dame: Preliminary results of real-time auralization with choir members,” in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–6, 2021.

[36] K. Kato, K. Ueno, and K. Kawai, “Musicians’ adjustment of performance to room acoustics, part iii: Understanding the variations in musical expressions,” *Journal of the Acoustical Society of America*, vol. 123, pp. 3610–3610, 2008.

[37] K. Kato, K. Ueno, and K. Kawai, “Effect of room acoustics on musicians’ performance. part ii: Audio analysis of the variations in performed sound signals,” *Acta Acustica united with Acustica*, vol. 101, 08 2015.

[38] S. LABS, “Understanding deep learning: Dnn, rnn, lstm, cnn and r-cnn.” Online, accessed: 12.07.2022.

[39] J. Brownlee, “A gentle introduction to the imagenet challenge (ilsvrc).” Online, accessed: 12.07.2022.

[40] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014.

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[42] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[43] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising,” *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[44] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “Image denoising by sparse 3-d transform-domain collaborative filtering,” *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[45] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[46] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

[47] Y. Goldberg, *Neural Network Methods in Natural Language Processing.* Morgan Claypool Publishers, 4 2017.

[48] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," 2014.

[49] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," 2015.

[50] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014.

[51] R. Nallapati, B. Zhou, C. N. d. santos, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence rnns and beyond," 2016.

[52] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition," in *Interspeech 2013*, ISCA, August 2013.

[53] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," 2013.

[54] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[55] M. E. P. Davies and M. D. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1009–1020, 2007.

[56] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[57] "Research areas - machine learning." Online, accessed: 14.07.2022.

[58] K. Doshi, "Audio deep learning made simple (part 1): State-of-the-art techniques." Online, accessed: 14.07.2022.

[59] 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. Wilson, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135, 2017.

[60] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 504–516, 2002.

[61] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[62] C. M. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[63] J. Terra, "Keras vs tensorflow vs pytorch: Key differences among the deep learning framework." Online, accessed: 15.07.2022.

[64] A. Systems, "Tensorflow vs keras vs pytorch: Which framework is the best?." Online, accessed: 15.07.2022.

[65] "Keras - simple. flexible. powerful.." Online, accessed: 15.07.2022.

[66] "Tensorflow - an end-to-end open source machine learning platform." Online, accessed: 15.07.2022.

[67] "Pytorch - from research to production." Online, accessed: 15.07.2022.

[68] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nature Reviews Genetics*, vol. 11, no. 10, pp. 733–739, 2010.

[69] M. Bojarski, P. Yeres, A. Choromanska, K. Choromanski, B. Firner, L. Jackel, and U. Muller, "Explaining how a deep neural network trained with end-to-end learning steers a car," *arXiv preprint arXiv:1704.07911*, 2017.

[70] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730, 2015.

[71] B. J. Taylor, *Methods and procedures for the verification and validation of artificial neural networks.* Springer Science & Business Media, 2006.

[72] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[73] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, *et al.*, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature medicine*, vol. 7, no. 6, pp. 673–679, 2001.

[74] J. Zurada, A. Malinowski, and I. Cloete, "Sensitivity analysis for minimization of input data dimension for feedforward neural network," in *Proceedings of IEEE International Symposium on Circuits and Systems - ISCAS '94*, vol. 6, pp. 447–450 vol.6, 1994.

[75] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, *Layer-Wise Relevance Propagation: An Overview*, pp. 193–209. Cham: Springer International Publishing, 2019.

[76] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, "Interpretable deep neural networks for single-trial eeg classification," *Journal of neuroscience methods*, vol. 274, pp. 141–145, 2016.

[77] F. Arbabzadah, G. Montavon, K.-R. Müller, and W. Samek, "Identifying individual facial expressions by deconstructing a neural network," in *German Conference on Pattern Recognition*, pp. 344–354, Springer, 2016.

[78] L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek, "What is relevant in a text document?": An interpretable machine learning approach," *PloS one*, vol. 12, no. 8, p. e0181142, 2017.

[79] S. S. Stevens, J. E. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America*, vol. 8, pp. 185–190, 1937.

[80] L. Treybig, S. Werner, U. Sloma, and G. Stolz, "Measure - analyze - auralize : From room impulse response to room classification and binaural reproduction," *DAGA 2022*, 2022.

[81] Y.-Y. Yang, M. Hira, Z. Ni, A. Chourdia, A. Astafurov, C. Chen, C.-F. Yeh, C. Puhrsch, D. Pollack, D. Genzel, D. Greenberg, E. Z. Yang, J. Lian, J. Mahadeokar, J. Hwang, J. Chen, P. Goldsborough, P. Roy, S. Narenthiran, S. Watanabe, S. Chintala, V. Quenneville-Bélair, and Y. Shi, "Torchaudio: Building blocks for audio and speech processing," *arXiv preprint arXiv:2110.15018*, 2021.

Acronyms

AI Artificial Intelligence.

AIR Acoustic Impulse Response.

API Application Programming Interface.

AR Augmented Reality.

BRIR Binaural Room Impulse Response.

CNN Convolutional Neural Network.

CRNN Convolutional Recurrent Neural Network.

dB Decibels.

DH Dummy Head.

DNN Deep Neural Network.

DSP Digital Signal Processing.

EDT Early Decay Time.

FDRT Frequency-Dependant Reverberation Times.

FFNN Feed Forward Neural Network.

FFT Fast Fourier Transform.

GMM Gaussian Mixture Model.

IACC Inter Aural Cross Correlation.

ILSVRC ImageNet Large Scale Visual Recognition Challenge.

LDA Linear Discriminant Analysis.

LRP Layer-wise Relevance Propagation.

LSTM Long Short Term Memory.

MFCC Mel-frequency Cepstral Coefficient.

MIR Music Information Retrieval.

MLA Machine Learning Algorithm.

MTB Motion Tracked Binaural.

NBC Naïve Bayes Classifier.

NLP Natural Language Processing.

NSV Negative-Side Variance.

PHE Past Has Ears.

ReLU Rectified Linear Units.

RIR Room Impulse Response.

RNN Recurrent Neural Network.

SDM Spatial Decomposition Method.

SOTA State Of The Art.

T60 Reverberation Time.

UBM Universal Background Model.

VA Virtual Acoustics.

WFS Wave Field Synthesis.

Declaration of Originality

I declare that the work has been conducted and written in its entirety by myself.
I certify that all references, direct and indirect, are indicated as such and have been cited accordingly. This thesis has not been used in the same or similar forms in parts or total in other examinations.

Sign:

Sayantan Gope

Date: August 1, 2022