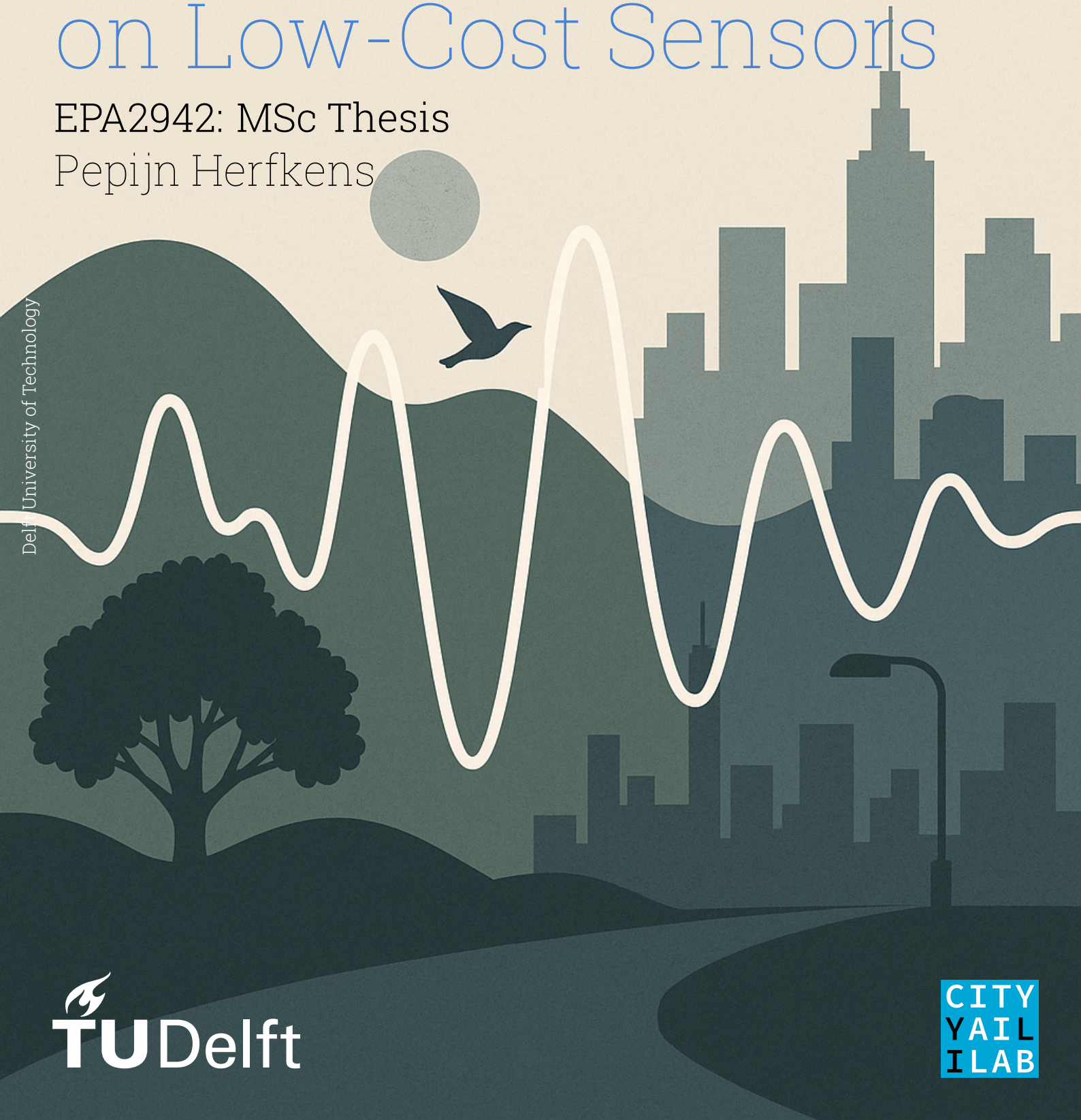
Edge AI for Urban Noise Monitoring: Perceptual Soundscape Prediction on Low-Cost Sensors

EPA2942: MSc Thesis

Pepijn Herfkens



Delft University of Technology

TU Delft

CITY YAIL ILAB

Edge AI for Urban Noise Monitoring: Perceptual Soundscape Prediction on Low-Cost Sensors

by

Pepijn Herfkens

in partial fulfilment of the requirements for the degree of

Master of Science
in Engineering & Policy Analysis

at the Delft University of Technology,
to be defended publicly on Monday 14 July 2025 at 09:00 AM.

Thesis committee:

First supervisor:	S. van Cranenburgh
Second supervisor:	S. Azimi Rashti
Advisor:	L. Cassens
Faculty:	Technology, Policy & Management | Delft
Project duration:	January 2025 – July 2025
Student number:	4962036

TU Delft

Preface

While I was in Bologna for my elective courses, I began thinking about how I wanted to approach my graduation project. When I came across Sander van Cranenburgh's website, I saw an opportunity to work on a project involving Machine Learning. Since I had thoroughly enjoyed his course *ML for Socio-Technical Systems* during my first master's year, this seemed like the perfect chance to deepen my knowledge in ML while also contributing to a broader societal issue: noise pollution - the subject of Lion Cassen's PhD research.

After a few video calls with Lion, I was able to start working on my research proposal. Once back in the Netherlands, I initially struggled to properly define the project scope. However, a single meeting with Lion was enough to clear things up. From then on, we met weekly on Wednesdays. These meetings were highly effective and always a pleasure. It's incredibly helpful to brainstorm with someone so deeply involved in the topic. In addition, Sander organised monthly workshops for all his thesis students. These were very inspiring sessions that kept me motivated and provided me with valuable feedback during presentations.

By May, it was time for my midterm presentation. Just before that, I had to revise my research direction because the original scope had already been covered in previous work. The midterm made it clear I had to pick up the pace to stick to the original timeline, and I managed to do so. By mid-June, I received the green light to finalise the thesis and prepare for the defence.

To my supervisors, Sander van Cranenburgh, Sepinoud Azimi Rashti, and Lion Cassens, thank you all for your guidance. Lion, a special thank-you to you: I truly appreciated our collaboration, and I wish you the best of luck with your PhD.

I also want to thank my parents, for always supporting me throughout my academic journey, and in everything I do. To my brother and sister: even though we no longer live together, I always knew I could count on your support. To my housemates: thank you for the coffee breaks at the library and the chilling together in the evenings. To my girlfriend: thank you for your kind words during the tough moments. To my uncle, who has always shown a strong interest in my academic journey and offered wise words and advice, the kind only an uncle can give. And finally, a special thanks to my grandparents. You've always been interested in the progress of my thesis, but more than that, you've supported me throughout my life. I'm grateful that you're still here to share this special moment of graduation with me.

I hope this research contributes meaningfully to tackling the issue of noise pollution. It would be amazing if it could one day be deployed on one of Lion's sensors!

Pepijn Herfkens
Delft, July 2025

Executive summary

Understanding how people perceive urban soundscapes is crucial for creating healthier, more liveable cities. Rather than focusing solely on reducing unwanted noise, modern urban planning increasingly emphasizes enhancing the overall auditory experience by incorporating the perceptual responses of individuals to their acoustic environment. This shift recognizes that not all sounds are undesirable. Natural sounds like birdsong or water features can positively contribute to the quality of public spaces. To support this perceptual approach, sensors offer a practical solution for in situ monitoring of soundscapes. When equipped with intelligent models capable of predicting how sounds are experienced, not just how loud they are, these sensors become powerful tools for informing urban policy. Low-cost, solar-powered sensors offer a scalable solution for monitoring soundscapes across diverse locations in real-time. Their affordability and ease of deployment make them well-suited for large-scale urban applications. However, these sensors also come with significant resource constraints, including limited memory, processing power, and energy availability. To effectively support perceptual soundscape analysis under these constraints, models must be compact, computationally efficient, and capable of running directly on the device, enabling privacy-preserving, edge-based processing without relying on centralized data transmission.

While a few models currently exist that can predict perceptual soundscape attributes, most of these architectures are too large and computationally intensive for deployment on the low-cost, resource-constrained sensors. To address this limitation, this study focuses on designing four parameter-reduced versions of the smallest available model capable of predicting the full set of perceptual attributes: AD_CNN. These adapted versions are designed to meet the specific requirements outlined in this thesis, including a maximum of 300K parameters, the ability to predict both sound source classes and all eight perceptual attributes, and competitive performance in terms of predictive accuracy and inference speed. The two best-performing models in terms of predictive accuracy are combined into an ensemble model to explore whether their complementary strengths can further enhance performance. Furthermore, it remains unclear how well predictive models in soundscape research generalise to unseen data when trained on a specific dataset. To address this, a generalisation study was conducted to evaluate the robustness and cross-context applicability of the developed models.

With the availability of the Affective Responses to Augmented Urban Soundscapes (ARAUS) dataset, it became possible to replicate the AD_CNN training pipeline for this thesis. Building on this foundation, the models developed in this study are designed to implement four targeted parameter reduction strategies: reducing the number of convolutional filters, shrinking the dense layers, lowering temporal resolution by increasing the spectrogram hop size, and applying more aggressive Max Pooling. Among the lighter variants, AD_CNN_dense_layer and AD_CNN_hop_length performed best, closely approaching the baseline AD_CNN in accuracy while significantly reducing parameter count. Building on these findings, an ensemble model, AD_CNN_ dense_hop_combined, was introduced, integrating both optimizations. Despite having only ~160K parameters, it outperformed the original AD_CNN in perceptual attribute prediction.

A generalisation study tested the lightweight ensemble model on a separate dataset of urban parks. While it performed well given its compact design, it struggled to fully capture human perceptions, particularly for pleasantness. In contrast, the larger benchmark model, SoundAQnet, generalised more effectively and aligned more closely with subjective ratings.

In conclusion, this thesis demonstrates that it is technically feasible to design lightweight neural network models capable of accurately predicting perceptual soundscape attributes and environmental sound sources, even under the strict computational and memory constraints of low-cost, resource-constrained sensors. These results offer a promising step towards enabling real-time, soundscape monitoring on sensors in urban environments. However, the generalisation study underscores a persistent challenge: models trained on one specific dataset, tend to struggle when applied to another,

unseen dataset. Even high-capacity models like SoundAQnet showed limited ability to generalise perceptual predictions across different soundscapes. This highlights a core limitation for policymakers: despite promising technical performance, predictive models are currently not yet reliable enough to serve as stand-alone tools for evaluating soundscapes across diverse urban settings. This is because perceptions of soundscapes are inherently subjective and shaped by a range of contextual factors, such as demographics, social activity, and the visual environment, that are not captured by the audio signal alone.

As a result, the predictive output of current models may not accurately reflect how a particular community or user group experiences their soundscapes. Therefore, fine-tuning models with localised perceptual data becomes essential. By collecting perceptual ratings from residents in specific locations and using them to recalibrate or adapt the model, predictions can be better aligned with local expectations and lived experiences. This context-aware fine-tuning not only improves predictive accuracy but also increases public acceptance and policy relevance. Without such adaptation, the risk remains that model outputs misrepresent how soundscapes are actually perceived, potentially leading to misguided interventions or inequitable policy outcomes.

Contents

List of Figures

List of Tables

1

Introduction

1.1. Background

Urban environments are important components of modern cities. Historically, focus was placed on visual elements like buildings and street layout, but auditory elements, such as bird sounds or water fountains, are now also recognized as vital to the urban environment (Botteldooren et al., 2008). The European Environment Agency (EEA) highlights the significant issue of environmental noise pollution in Europe, where approximately one in five people are exposed to harmful long-term noise levels (European Environment Agency, 2020). Traditionally, urban noise management has focused on limiting exposure to unwanted or harmful sounds. Its primary aim was to reduce traffic noise, with two straightforward approaches: reducing noise at the source, for example, by using quieter road surfaces, and blocking its propagation with barriers, such as sound walls along highways (Brown & Van Kamp, 2017). In open urban spaces like squares and parks, though, this approach does not always apply in the same way, as not all sounds in these settings are considered undesirable. Some sounds do fit in an environment. For instance, annoying traffic noise in open urban spaces may be masked or offset by natural sounds, such as birdsong or water features, which are generally perceived as pleasant and contextually appropriate (Chitra et al., 2020).

This shift in focus, from merely reducing unwanted sounds to enhancing the overall auditory experience, has led to the development of the *soundscape* concept. In densely populated urban areas, noise pollution remains a persistent issue. Soundscape analysis offers a valuable tool for policymakers (World Health Organization, 2018). By integrating soundscape research into urban planning, they can better design and manage acoustic environments. This enables more tailored and effective solutions. It also supports the creation of healthier, more pleasant, and liveable spaces. This is especially relevant for new residential and recreational developments, as well as for the revitalisation of older urban areas that were developed without consideration for acoustic quality or sustainability (Kang et al., 2019). Moreover, soundscape research influences well-being. It can impact health, cultural identity, and economic factors in urban settings (Kang & Schulte-Fortkamp, 2016).

Schafer's (1969) idea of a *soundscape*, as the auditory equivalent of landscape, provides a framework for interpreting urban noise more positively while also accounting for individual experiences (Raimbault & Dubois, 2005). Murray Schafer introduced the term "soundscape" in his seminal work *The Soundscape: Our Sonic Environment and the Tuning of the World* (Schafer, 1977). His work is considered as pioneering in soundscape (Kang, 2021). A soundscape, according to the International Organization for Standardization (ISO 12913-1), is defined as an "acoustic environment as perceived or experienced and/or understood by a person or people, in context" (ISO, 2014). Building on this, as stated by Kang (2023) a soundscape differs from an "acoustic environment" because it emphasizes how sounds are perceived and experienced rather than being purely about their physical properties. It takes a more holistic view, treating sounds as valuable "resources" rather than "wastes". The focus is on appreciating "wanted" sounds that people prefer, rather than solely addressing "unwanted" sounds that cause discomfort.

Therefore, understanding those individual experiences and how people perceive sounds can contribute to more effective urban planning and design (Brown & Muhar, 2004). For urban open public spaces, it is essential to study how users perceive sounds (Yang & Kang, 2004). Such research is called soundscape research. People's perceptions of a soundscape can be studied in situ, simulated or reproduced in a controlled indoor environment, or recalled from memory. Among these methods, in situ research offers the most realistic representation of the external world but comes with lower experimental validity (Aletta et al., 2016). Conducting research on-site also captures the visual aspects of a soundscape, an element absent in laboratory settings, which can influence evaluation results (Cadena et al., 2017). In contrast, laboratory-based simulations allow for better control over auditory and contextual variables, enhancing reproducibility and experimental rigour (Cadena et al., 2017). A memory-based method is a qualitative or mixed-method approach that explores how memories shape individuals' perception, evaluation, and emotional response to sound environments. However, this approach also introduces subjectivity and variability, which can limit the generalisability and reliability of the findings (Jo & Jeon, 2021).

1.2. Foundation of soundscape perception

1.2.1. Perceptual attributes and circumplex model

Nilsson et al. (2007) found that natural sounds are generally perceived as pleasant, human sounds as eventful, and technological sounds as unpleasant. Moreover, these perceptual categories were shown to be stronger predictors of overall soundscape quality than conventional acoustic metrics, such as the equivalent continuous sound pressure level (LAeq). Consequently, Axelsson et al. (2010) found the need for a model that would capture the main aspects of how people perceive soundscapes and which could help with measuring and improving soundscape quality. They propose a principal components model of soundscape perception, identifying three key dimensions: *Pleasantness*, *Eventfulness*, and *Familiarity*. The final model only includes *Pleasantness* and *Eventfulness*, as *Familiarity* was found to have limited variance and practical relevance. Based on Russell's circumplex model of affect, two primary dimensions are used to describe emotions: perceived pleasantness and the level of arousal or stimulation (Russell, 1980). Building on this model, Västfjäll et al. (2003) demonstrated that it is applicable for describing emotions evoked by interior aircraft sounds. Following these findings, Axelsson et al. (2010) proposed that a similar framework could be relevant for describing soundscape perception.

The resulting circumplex model utilizes a two-dimensional framework defined by attributes such as Pleasant, Chaotic, Exciting, Uneventful, Calm, Unpleasant (or Annoyance), Eventful, and Monotonous, as shown in Figure 1.1.

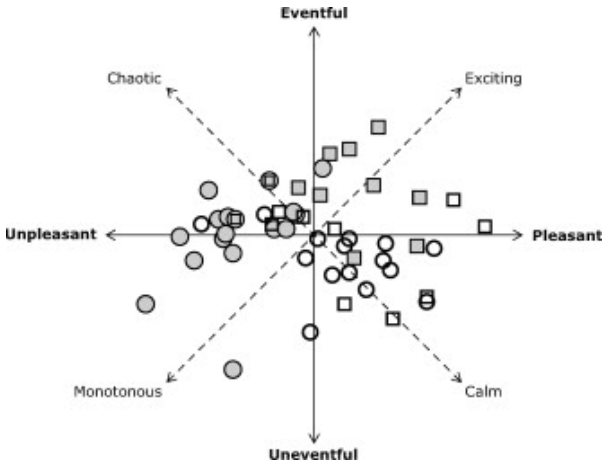


Figure 1.1: Circumplex model as developed by Axelsson et al. (2010)

1.2.2. ISO Methods for soundscape evaluation and representation

One commonly used procedure for soundscape assessments is "Method A" from ISO/TS 12913-2:2018 (ISO, 2018). It relies on Likert-scales and is suitable for large-scale, on-site surveys, allowing data to be gathered from potentially hundreds of public space users in a short time. To analyse the Likert-scale

responses for the perceptual attributes, they are encoded as ordinal variables, ranging from 1 (strongly disagree) to 5 (strongly agree). Following ISO/TS 12913-3:2019 (ISO, 2019), these eight attributes can be projected onto two primary dimensions, *Pleasantness* and *Eventfulness*, using a mathematical transformation, as shown below. This transformation exploits the 45-degree relationship between the diagonal axes and the pleasant and eventful axes.

$$\begin{aligned}\text{ISO Pleasantness} = [&(\text{pleasant} - \text{annoying}) + \cos 45^\circ \cdot (\text{calm} - \text{chaotic}) \\ &+ \cos 45^\circ \cdot (\text{vibrant} - \text{monotonous})] \cdot \frac{1}{4+\sqrt{32}}, \end{aligned} \tag{1.1}$$

$$\begin{aligned}\text{ISO Eventfulness} = [&(\text{eventful} - \text{uneventful}) + \cos 45^\circ \cdot (\text{chaotic} - \text{calm}) \\ &+ \cos 45^\circ \cdot (\text{vibrant} - \text{monotonous})] \cdot \frac{1}{4+\sqrt{32}}. \end{aligned} \tag{1.2}$$

In addition to the single-point summaries, Mitchell, Aletta, and Kang (2022a) propose a probabilistic approach to soundscape representation. Instead of reducing the collective perception of a soundscape to a single *Pleasantness-Eventfulness* coordinate, they advocate for visualising the entire distribution of perceptual responses within the circumplex. This method captures the variability and diversity of human perception, offering a more nuanced and realistic understanding of how different people experience urban soundscapes.

1.3. Existing models to predict perceptual attributes

Due to the growing importance of soundscape research, there is an increasing demand for practical tools, such as predictive models, to integrate the soundscape approach into urban planning and design (Aletta & Xiao, 2018). Several studies have focused on predicting individual perceptual soundscape attributes such as *Annoyance*, *Pleasantness*, or *Eventfulness* (Hou, Ren, et al., 2023; Mitchell et al., 2021, 2023b). In recent years, deep learning models have emerged as the dominant approach for such predictions. For instance, dual-branch CNN architectures combining Mel spectrogram and psychoacoustic features have achieved root mean square error (RMSE) as low as 1.05 in predicting annoyance ratings on the DeLTA dataset (Deep Learning Techniques for noise Annoyance detection), while also classifying sound sources with over 90% accuracy (Hou, Mitchell, et al., 2023). More advanced fusion models, such as the dual-branch convolutional neural network with cross-attention-based fusion (DCNN-CaF), have further enhanced both source classification and annoyance prediction performance (Hou, Ren, et al., 2023). These results highlight the advantages of deep learning representations compared to relying solely on traditional acoustic or psychoacoustic features.

However, limited research has addressed the prediction of the full set of eight perceptual attributes (*pleasant*, *vibrant*, *eventful*, *chaotic*, *annoying*, *monotonous*, *uneventful*, and *calm*). To date, the only study that comprehensively modelled human perception of soundscapes is presented by Hou et al. (2024). In this work, the authors introduced SoundAQnet, a deep learning model trained to predict these perceptual attributes directly from audio recordings. The model achieved a mean squared error (MSE) of 1.054, indicating strong predictive performance. In the same study, several additional models were benchmarked against SoundAQnet. These models represent the only existing approaches that jointly predict both sound sources and perceptual attributes, or *affective qualities*, as referred to in the original work.

Using such models enables a more comprehensive representation of a soundscape. By predicting all eight perceptual attributes, it becomes possible to derive standardized indicators such as *Pleasantness* and *Eventfulness*, as defined in ISO 12913-3 (ISO, 2019). These indicators provide valuable insights into the overall quality of a soundscape and support its practical integration into urban planning and design processes (Axelsson, 2015).

1.4. Edge AI and sensor-based approaches for in situ soundscape monitoring

Cities have been a key focus for in-situ soundscape monitoring, largely driven by the need to manage noise pollution and improve urban liveability. In dense urban areas, sensor networks are deployed to record sound levels and events, which contribute to the creation of noise maps, the identification of problematic noise sources, and the analysis of urban soundscape composition (Aletta & Kang, 2015; Hajnal & Kocsis, 2022; J. Liu et al., 2013).

The introduction of Internet of Things (IoT) technologies in smart cities has recently shifted approaches to environmental noise monitoring, leading to the development of wireless acoustic sensor networks (WASNs) (Alías & Alsina-Pagès, 2019). Several large projects in different cities have implemented such networks. For instance, the SONYC project in New York City uses its sensor network to automatically recognize sound events (e.g. traffic, honking, construction noise) and assist city agencies in noise mitigation (SONYC Project, n.d.). This work demonstrates how in situ sensor data, combined with machine learning, can assist city authorities in enforcement and provide insights into urban soundscapes (Bello et al., 2019).

In line with efforts like SONYC, Alsina-Pagès et al. (2020) proposed a sensor design featuring low-cost, reconfigurable hardware capable of real-time audio capture, on-device sound source classification via machine learning, and wireless data transmission. However, the estimated cost of around €139 per unit and the requirement for continuous connection to the power grid limit its scalability for large-scale deployment.

This highlights the need for more affordable sensors that can operate independently of the power grid by solar power (Cassens et al., 2024). In response, Cassens et al. (2024) developed a sensor based on the ESP32-S3 microcontroller, which features a dual-core 240 MHz processor, 8MB of RAM, and a similar amount of Flash memory. In comparison, the sensor from Alsina-Pagès et al. (2020) uses a Raspberry Pi 3 with a 1.2 GHz quad-core CPU.

1.5. Knowledge gap

As urban noise management increasingly adopts a comprehensive approach by incorporating the soundscape perspective into planning and policy-making (Mitchell, Aletta, & Kang, 2022a), sensor deployment has become a valuable method for measuring and researching soundscapes to inform such policies. Low-cost sensors make it feasible to build large-scale acoustic sensor networks. However, most of these sensors function merely as sound level meters, measuring only loudness (Picaut et al., 2020). Since soundscapes encompass more than loudness alone, and must also account for human perception, there remains a clear gap in low-cost sensors capable of predicting both sound sources and perceptual attributes.

Moreover, due to transmission limitations and, more importantly, privacy concerns, sensors deployed in urban settings must process audio locally - on the edge - without transmitting large segments of raw audio. This necessitates the development of models that are both lightweight in storage requirements and computationally efficient enough to perform real-time, on-device processing. By processing data locally, edge AI reduces energy consumption and enhances data privacy (Karges et al., 2022).

While some progress has been made in modelling individual perceptual attributes, limited research has addressed the prediction of the complete set of eight perceptual dimensions: *pleasant*, *vibrant*, *eventful*, *chaotic*, *annoying*, *monotonous*, *uneventful*, and *calm*. Currently, SoundAQnet is the only model that can predict both sound sources and all eight perceptual attributes. However, its relatively large size (approximately 10 MB) makes it unsuitable for deployment on low-cost, solar-powered sensors, which have restricted computational resources. For instance, the sensor developed by Cassens et al. (2024) provides only 8 MB of Flash memory, a portion of which is already allocated for system files and other essential data. This leaves even less available space for storing a machine learning model. Moreover, large models like SoundAQnet require substantial computational power and processing time, further limiting their feasibility for real-time, on-device inference on resource-constrained sensors.

Furthermore, many models, including SoundAQnet, are typically trained and evaluated on a single dataset, which raises concerns about their generalisation ability. Without cross-dataset validation or

testing on diverse urban environments, these models risk overfitting to specific acoustic or contextual characteristics. As a result, their applicability in broader urban planning or policymaking remains limited. For models to inform soundscape-related policy effectively, they must demonstrate robust performance across varied geographical, cultural, and environmental contexts.

In summary, this highlights a clear research gap in the development of models that can predict perceptual attributes and sound sources in real-time while being lightweight enough for deployment on low-cost sensors, and that can generalise well to new and unseen data. This thesis addresses this gap by designing and evaluating neural networks that meet the computational and memory constraints of such devices, enabling real-time, on-device processing in urban environments.

1.6. Research objective and research questions

The objective of this research is to evaluate and compare a range of lightweight deep learning models for predicting perceptual soundscape attributes and sound sources from urban audio recordings. All models will be designed to meet the constraints of low-cost, solar-powered sensors, which have limited memory, processing power, and energy availability. In addition, this study will assess the generalisation performance of the models by evaluating them on a completely unseen urban audio dataset, thereby testing their robustness across different acoustic and contextual environments.

These low-cost sensors have the following characteristics and constraints, as stated by (Cassens et al., 2024):

- They are solar-powered, allowing for flexible placement without reliance on the power grid, but this also implies limited energy availability, so the sensor must run an energy-efficient model.
- They perform audio processing directly on the device (i.e., on the edge), which helps preserve user privacy and minimizes data transmission.
- They are constrained in memory and storage, typically offering no more than 8 MB of RAM, and 8 MB of Flash memory.

Given these constraints, the models developed in this thesis must adhere to the following requirements:

- **Model size:** Each model must be lightweight, with a maximum of 300,000 parameters. This upper limit reflects the known capacity of models that can reliably run on low-cost, resource-constrained sensors.
- **Functionality:** The model must be capable of predicting both sound source classes and the full set of eight perceptual attributes: *pleasant*, *vibrant*, *eventful*, *chaotic*, *annoying*, *monotonous*, *uneventful*, and *calm*.
- **Performance:** The model must demonstrate competitive performance relative to state-of-the-art models (such as SoundAQnet) in terms of predictive accuracy and inference speed.

This thesis formulates the following main research question: **How can lightweight neural network models be designed to accurately predict perceptual soundscape attributes and sound sources from urban audio recordings in real-time on low-cost, resource-constrained sensors, and to what extent can such models generalise across diverse soundscapes to support urban policy making?**

To collectively address the main research question, the following sub-questions have been formulated:

1. **Which existing deep learning models are suitable for predicting multiple perceptual soundscape attributes simultaneously, and how can they be adapted for low-cost, resource-constrained sensors?**
   This question will be explored through a detailed review of existing literature. The goal is to find and compare different deep learning models that work well for regression tasks. The review will also look at how suitable these models are for use in devices with limited computing power, such as low-cost, resource-constrained sensors in urban areas. The results of this review will help guide the choice and development of lightweight prediction models used in this study.

2. **How does the performance (accuracy, computational efficiency, and storage requirements) of the designed lightweight models compare to state-of-the-art model SoundAQnet?**
   This question compares the designed models with the current state-of-the-art model, SoundAQnet. It looks at key aspects like the models' accuracy at predicting attributes, inference speed, memory usage, and how well it could work in deployment.
3. **How well do predictive soundscape models generalise to different soundscapes, and what are the implications of their generalisation performance for urban policy making?**
   This question will be addressed through a generalisation study. The designed models will be evaluated in inference mode on a dataset that differs from the one used during training. By comparing the predictive performance on this unseen dataset to the performance achieved on the original test set, the study will assess each model's ability to generalise across varying urban acoustic environments. The findings will provide insight into the models' robustness and their practical relevance for informing urban soundscape policy across diverse contexts.

1.7. Thesis outline

This thesis is structured as follows. Chapter 2 reviews the evolution of soundscape research, from perceptual studies to data-driven approaches, with a focus on lightweight deep learning models. Chapter 3 details the methodology, including data preprocessing, model architecture design, training procedures, and the generalisation study. Chapter 4 describes the datasets used (ARAUS and ISD) and outlines dataset-specific adaptations. Chapter 5 presents experimental results, benchmarking the developed models against SoundAQnet and evaluating their generalisation. Chapter 6 discusses interpretability, limitations, and implications for urban policy. Chapter 7 concludes by answering the research questions, summarizing key findings, and proposing directions for future work.

2

Literature Review

This chapter presents an overview of machine learning and deep learning techniques applied in soundscape research. It reviews prominent models used for audio classification and the prediction of perceptual attributes. The chapter concludes by evaluating which models are best suited for deployment on low-cost, resource-constrained sensors, focusing on those with minimal storage requirements that can simultaneously predict both sound sources and perceptual attributes.

2.1. Methodologies for soundscape analysis, classification and prediction

Soundscape research marked a paradigm shift in environmental acoustics by initially focusing on people's perceptions before incorporating physical measurements (Brooks et al., 2014). The first significant initiative in this field was established in the early 1970s: the World Soundscape Project, which involved comprehensive studies of the 'sonic environment' (World Soundscape Project, n.d.). A common research methodology adopted was the *soundwalk*, wherein researchers and local participants walked through specific areas, attentively listening and subsequently describing the soundscape in their own words (Schafer, 1977).

Modern soundscape research employs a variety of methodological approaches. In the context of urban soundscapes, these approaches are typically categorised into subjective evaluations and objective measurements. Subjective evaluations entail collecting individuals' perceptions and experiences of soundscapes, typically through methods such as interviews, questionnaires, and field observations (Bild et al., 2018; F. Liu & Kang, 2016; Ma et al., 2021). Objective measurements, on the other hand, focus on quantifying the physical characteristics of soundscapes using acoustic indices and metrics (Herranz-Pascual et al., 2017).

2.1.1. Machine learning and deep learning in soundscape analysis

Analytical strategies in soundscape research vary depending on the nature of the data and research objectives. Linear statistical models are often used to identify associations between perceptual ratings and environmental or contextual variables (Jeon et al., 2010; J. Liu et al., 2014). Increasingly, data-driven approaches such as machine learning and deep learning are employed to model complex relationships and improve predictive performance. For instance, Support Vector Machines (SVM) have been used for soundscape classification (Torija et al., 2013), Convolutional Neural Networks (CNN) for detecting species-specific sounds such as birds and frogs (LeBien et al., 2020), and Artificial Neural Networks (ANN) for categorising urban soundscape types (Jeon & Hong, 2015).

Deep learning has significantly advanced the field of audio classification by enabling models to automatically learn complex features directly from raw or minimally processed audio data (Zaman et al., 2023). These approaches can achieve higher accuracy and adaptability across diverse audio classification tasks than traditional methods such as SVM and ANNs. Among the deep learning models developed for audio classification, the Pretrained Audio Neural Network (PANN) (Kong et al., 2020) and the Au-

dio Spectrogram Transformer (AST) (Gong et al., 2021) have emerged as state-of-the-art approaches. Early CNN-based models tend to give decent results in predicting audio tags, but would only capture a limited amount of sound classes (Choi et al., 2016). Since computer vision and natural language processing have greatly benefited from large-scale datasets, such as ImageNet for image classification (Deng et al., 2009) and Wikipedia for training language models (Devlin et al., 2019), Kong et al. (2020) were motivated to develop a system trained on a large-scale audio dataset. This became possible with the release of AudioSet (Gemmeke et al., 2017), which contains a substantial amount of hours of audio recordings labelled with 527 sound classes. By training their model on raw audio recordings rather than on embeddings from a pre-trained convolutional neural network (Hershey et al., 2017), the authors achieved a state-of-the-art mean average precision (mAP) of 0.439, significantly outperforming benchmarks such as Google's CNN-based baseline (mAP = 0.314) (Gemmeke et al., 2017).

2.1.2. Models for predicting soundscape attributes

Different from classification are predictive models which incorporate various factors to predict soundscape perception and quality. *Annoyance* is a common descriptor to describe subjective perceptions (Lionello et al., 2020) and has been well-researched (Mitchell, Erfanian, Soelistyo, et al., 2022; Mitchell et al., 2023b).

DCNN-CaF

One recent study applies artificial intelligence (AI) to analyse urban soundscapes by combining the recognition of sound sources with predictions of the annoyance expierenced by people (Hou, Ren, et al., 2023). The researchers introduced a model called the dual-branch convolutional neural network with cross-attention-based fusion (DCNN-CaF). This model can identify different sound sources while also predicting the level of annoyance these sources might cause. It uses two distinct audio features: Mel spectrograms and loudness-related root mean square values (RMS). A spectrogram visually represents how a signal's frequency content changes over time. The Mel scale provides a linear scale that matches human hearing, meaning each step in frequency on the Mel scale has the same distance from one another. As a result, Mel spectrograms more closely reflect human auditory perception (B. Zhang et al., 2019). The RMS is used to measure the loudness of a signal within a spectrogram (Mulimani & Koolagudi, 2018).

The DCNN-CaF model was trained on the DeLTA dataset, a collection of urban audio recordings annotated by human listeners (Mitchell, Erfanian, Soelitsyo, et al., 2022). It outperformed several popular deep learning models, including YAMNet, PANN, and AST, in the task of Sound Source Classification (SSC). Moreover, in the combined task of SSC and *annoyance* prediction, DCNN-CaF also surpassed the considerably larger CNN-Transformer model, which contains approximately 20 million parameters. The researchers attribute this increased performance to DCNN-CaF's relatively low parameter count and architectural depth, which helped prevent overfitting, a common issue when training large models on limited data. DCNN-CaF also outperformed smaller baseline models, such as a Deep Neural Network (DNN) with 0.38 million parameters and a Convolutional Neural Network (CNN) with 1.27 million parameters. To evaluate generalizability, the researchers tested the model on previously unseen sounds from external datasets. The DCNN-CaF maintained strong performance under these conditions, reinforcing its robustness. The final model comprised approximately 7.6 million parameters.

TinyCNN

Another research group explored predicting *annoyance* using deep learning models in a study conducted at The Alan Turing Institute (Mitchell et al., 2023a). The Data Study Group explored multiple models to predict *human-perceived noise annoyance* in urban environments, also using the DeLTA dataset. The team researched whether the inclusion of sound sources improves the prediction of accuracy. They developed deep learning models, including Convolutional Neural Networks (CNN), TinyCNNs, Temporal Convolutional Networks (TCN), Feedforward Neural Networks (FNN), and Long Short-Term Memory (LSTM). Among these, the pretrained audio neural network (PANN), as developed by Kong et al. (2020), demonstrated the best performance in predicting *annoyance* ratings. Interestingly, a much smaller model was found to perform nearly as well, suggesting that lightweight architectures may offer competitive alternatives with lower computational demands.

The study includes multiple sets of models. The W-models serve as classical machine learning baselines, without the use of deep learning. Models W.1 and W.2 are based on linear regression, while W.3

and W.4 use random forests. Whereas W.1 and W.3 rely solely on the presence of sound sources, W.2 and W.4 also incorporate *spectral features*, which can be interpreted as indicators of a sound's perceived brightness. A higher spectral centroid suggests that a sound is dominated by higher frequencies, making it sharper or brighter, while a lower centroid indicates dominance by lower frequencies, resulting in a deeper or duller auditory quality. Among the baseline models, W.4 (random forest with spectral features) performs best, achieving a test RMSE of 1.17. This comparison demonstrates that including even basic spectral information enhances predictive accuracy beyond using only sound source presence.

The study then introduces deep learning models to augment the prediction of *annoyance*. Models C.1 to C.4 are simple deep learning models trained on compressed Mel spectrograms of the recordings, without using any explicit sound source information. Each model tests a different type of neural network architecture: C.1 uses a Temporal Convolutional Network (TCN), C.2 uses a Feedforward Neural Network (FNN), C.3 applies a Convolutional Neural Network (CNN), and C.4 incorporates a Long Short-Term Memory network (LSTM). However, none of these models achieves strong performance, likely due to the use of compressed Mel spectrograms. In contrast, model Y.0, which is based on a simple CNN but receives a high-resolution Mel spectrogram as input, performs well.

Next, the R-models aim to improve prediction performance by explicitly incorporating human-labeled sound source information as additional inputs. Both R.1 and R.2 combine Mel spectrograms with sound source labels, but differ in architecture and label representation. R.1 uses a TCN and includes binary sound source indicators, while R.2 employs a TinyCNN and encodes source labels as probabilistic values, reflecting the degree of agreement among human annotators. Despite these enhancements, the R-models do not perform well, likely due to the added complexity of combining different input types, which leads to overfitting and reduced generalizability.

The Y.2 to Y.4 models explore whether pretrained sound source classification models can improve *annoyance* prediction. These models use features extracted from PANN to provide automated sound source information. In Y.2, PANN is applied directly to the Mel spectrograms to perform both sound source classification and *annoyance* prediction. Y.3 extends this approach by feeding the PANN-extracted features into a one-layer FNN to output an *annoyance* rating. Y.4 builds on this further by incorporating a more complex architecture that includes a CNN-Transformer following the PANN feature extraction. These models achieve strong performance, with Y.2 and Y.3 reaching test RMSE values around 1.08–1.10.

Finally, models Y.1 and Y.6 incorporate sound source information implicitly, using it as an additional prediction target to enhance *annoyance* prediction. Rather than using sound source labels as input features, these models are trained to jointly predict both *annoyance* and sound source presence, similar to Y.2 to Y.4, but with a key difference: Y.1 and Y.6 employ a simple TinyCNN trained from scratch, without pretrained weights from PANN. In Y.1, the model predicts binary sound source labels alongside *annoyance*, while Y.6 predicts probabilistic source labels, reflecting the level of agreement among annotators. Both models achieve strong performance, although Y.6 shows signs of overfitting.

The main takeaways from this paper, in the context of this thesis, are as follows:

- Including sound source labels explicitly as model inputs (R-models) often decreases performance due to added complexity. In contrast, implicitly incorporating this information by predicting sound sources jointly (Y-models) improved annoyance prediction.
- A lightweight TinyCNN trained on high-resolution spectrograms performed nearly as well as larger and more complex models, showing that simplicity paired with good input features can be highly effective.

This latter finding suggests that such models could be deployed on low-power devices, potentially even on remote monitoring sensors used in smart city applications, a recommendation also emphasized by the researchers. While the researchers did not publish their model implementations, an estimation can be made based on the model architectures described in the report. Models that use PANN likely contain over 80 million parameters, as PANN alone accounts for approximately 79.6 million. In contrast, the TinyCNN models are estimated to contain around 2.3 million parameters.

SoundAQnet

SoundAQnet is a deep learning model designed to predict perceptual soundscape attributes by jointly modelling acoustic scenes (AS), audio events (AE), and affective qualities (AQ). For example, an AS might be a ”park” or ”busy street”, an AE could be ”birds chirping” or ”car horn”, and AQ, referred to in this thesis as a *perceptual attribute*, describes how listeners perceive the soundscape, such as feeling “calm”, “chaotic”, or “annoying”. Developed by Hou et al. (2024), it represents a state-of-the-art approach for soundscape analysis and captioning. It also integrates the perceptual dimensions, such as *Pleasantness* and *Eventfulness*, as defined by ISO/TS 12913-3:2019 (ISO, 2019), and further discussed in Section 1.2.2.

The model architecture features two main branches: a Mel-spectrogram-based branch and a loudness-based branch, which extract temporal-frequency and psychoacoustic features respectively. The extracted features are fused using a gated graph convolutional network (GatedGCN), which allows the model to learn relationships across different timescales and acoustic dimensions. This fusion enables SoundAQnet to simultaneously output scene and event classifications, as well as regression estimates for all eight perceptual soundscape attributes, and the indices for *Pleasantness* and *Eventfulness*.

SoundAQnet was trained on the ARAUS dataset (Ooi et al., 2024), a large-scale synthetic soundscape dataset built from urban audio recordings enriched with human-annotated AQ scores. Although computationally intensive, the model achieves high accuracy across classification and regression tasks, making it a strong benchmark for evaluating other soundscape prediction models. The SoundAQnet itself was also benchmarked against different models, both smaller and larger models.

The smallest benchmark model is the AD_CNN model, which is based on the baseline CNN architecture developed by Ooi et al. (2024) for the ARAUS dataset, which was originally designed to predict ISO *Pleasantness* from log-mel spectrograms. While the baseline model focuses solely on ISO *Pleasantness*, Hou et al. (2024) extended it into the AD_CNN, enabling it to predict the full set of perceptual features used in SoundAQnet. Unlike SoundAQnet, however, the AD_CNN uses only mel spectrograms as input, without incorporating loudness features. The architecture comprises three convolutional layers, each followed by batch normalization, ReLU activation, and max pooling to extract hierarchical features from the input. The resulting feature maps are flattened and passed through fully connected layers to perform classification (scene and event labels) and regression (perceptual attributes). The model contains approximately 520K parameters, making it considerably larger than the baseline CNN for ARAUS, but still a lightweight architecture overall.

Secondly, the Baseline_CNN was introduced as a reference model to serve as a benchmark for evaluating more advanced architectures. It consists of four convolutional layers with progressively increasing kernel sizes and contains approximately 1.01 million parameters. The model uses only mel spectrograms as input and does not incorporate psychoacoustic features such as loudness. While it achieves satisfactory performance in both acoustic scene classification (ASC) and audio event classification (AEC), its predictive accuracy for perceptual attributes is relatively limited.

An extension of this model, the Hierarchical_CNN, was developed to incorporate the hierarchical relationship between audio events and acoustic scenes. This architecture retains the same convolutional architecture and number of parameters as the Baseline_CNN but modifies the output structure: predictions from the audio event classification branch are used as additional input to the scene classification branch. This hierarchical dependency improves performance in scene classification; however, it results in a slight decline in event classification accuracy, likely due to the added dependency between the two tasks.

In addition to the benchmark models described above, two established deep learning architectures, MobileNetV2 and YAMNet, were included. MobileNetV2 (Sandler et al., 2018) is a lightweight convolutional neural network originally developed for efficient image classification tasks. It employs depthwise separable convolutions and inverted residual blocks to significantly reduce the number of parameters and computational complexity. For the purpose of soundscape analysis, the architecture was adapted to operate on mel spectrogram inputs. With approximately 2.26 million parameters, MobileNetV2 achieves competitive performance across all tasks. Its strength lies in its ability to maintain a good balance between model complexity and predictive accuracy, particularly in classification tasks. However, like other general-purpose CNNs, it shows limited capability in capturing nuanced perceptual attributes,

which are better modelled by architectures specifically tailored for soundscape analysis.

YAMNet (Yu et al., 2020) is a convolutional neural network model developed by Google, trained on AudioSet for large-scale audio event classification. It is based on the MobileNetV1 architecture and utilizes pre-trained weights to recognize a wide range of sound events. In this benchmark, YAMNet was fine-tuned on the target tasks to ensure fair comparison. The model contains approximately 3.21 million parameters and operates exclusively on mel spectrogram inputs. While YAMNet performs robustly on audio event classification, its performance on acoustic scene classification and perceptual attribute regression is less consistent. This is likely due to its original design focus being on event-level tagging rather than contextual or affective interpretation.

Additionally, benchmark models, CNN-Transformer and PANNs, were included to evaluate the performance of SoundAQnet against more complex and expressive deep learning architectures, particularly those capable of modelling temporal dependencies in audio signals. In this architecture, the convolutional layers first extract time-frequency representations from mel spectrograms, after which a transformer encoder captures long-range dependencies across time. The model contains approximately 12.29 million parameters, making it significantly more complex than the previously described benchmarks. While the CNN-Transformer model achieves strong performance in both acoustic scene and audio event classification, its results in perceptual attribute regression are less consistent. This may be because the transformer focuses on modelling long-term temporal patterns, which helps with event detection but may be less effective for capturing the more subtle, localized features needed to predict perceptual attributes accurately.

Lastly, a variant of PANNs was fine-tuned for acoustic scene classification, audio event classification, and perceptual attribute regression. The model operates on mel spectrogram inputs and features a deep convolutional architecture with over 79 million parameters. While PANNs achieve strong performance on classification tasks, their ability to predict perceptual attributes is less consistent. Additionally, their high computational cost limits their practicality for lightweight or real-time applications.

All models are summarized in Table 2.1. The table specifies the input dimensions for the mel spectrogram in the format $(time\ frames, mel\ bins)$, and, where applicable, the input dimensions for loudness as $(time\ frames, 1)$, with each time frame corresponding to a single loudness value. It further indicates whether the model predicts sound sources and which perceptual attributes it targets. The number of model parameters is listed in millions, with an asterisk (*) denoting estimates where the exact size is unknown due to unavailability on GitHub. Finally, the table includes references to the developers of each model; in cases where a model builds on earlier work, the original sources are also cited.

Table 2.1: Comparison of models predicting sound sources and perceptual attributes.

Model	Mel Input	Loudness Input	Sound Sources	Perceptual Attributes	Parameters (M)	Source
DNN	(480, 64)	(482, 1)	yes	annoyance	0.38	(1)
CNN	(480, 64)	(482, 1)	yes	annoyance	1.27	(1)
CNN_Transformer	(480, 64)	(482, 1)	yes	annoyance	17.97	(1)
DCNN-CaF	(480, 64)	(482, 1)	yes	annoyance	7.61	(1)
Y.2 (PANN)	(480, 64)	-	yes	annoyance	>80.0*	(2)
Y.3 (PANN)	(480, 64)	-	yes	annoyance	>80.0*	(2)
Y.4 (PANN)	(480, 64)	-	yes	annoyance	>80.0*	(2)
Y.1 (TinyCNN)	(480, 64)	-	yes	annoyance	~ 2.36*	(2)
Y.6 (PANN)	(480, 64)	-	yes	annoyance	>80.0*	(2)
Y.0 (TinyCNN)	(480, 64)	-	no	annoyance	~ 2.36*	(2)
AD_CNN	(3001, 64)	-	yes	full circumplex	0.52	(3) & (4)
Baseline CNN	(3001, 64)	-	yes	full circumplex	1.01	(3)
Hierarchical_CNN	(3001, 64)	-	yes	full circumplex	1.01	(3)
MobileNetV2	(3001, 64)	-	yes	full circumplex	2.26	(3) & (5)
Yamnet	(3001, 64)	-	yes	full circumplex	3.21	(3) & (6)
CNN-Transformer	(3001, 64)	-	yes	full circumplex	12.29	(3)
PANNs	(3001, 64)	-	yes	full circumplex	79.73	(3) & (7)
SoundAQnet	(3001, 64)	(15000, 1)	yes	full circumplex	2.7	(3)

*Exact size is unknown as the model is not available on GitHub.

Sources:
(1) (Hou, Ren, et al., 2023)
(2) (Mitchell et al., 2023b)
(3) (Hou et al., 2024)
(4) (Ooi et al., 2024)
(5) (Sandler et al., 2018)
(6) (Yu et al., 2020)
(7) (Kong et al., 2020)

2.2. Conclusion of literature study

As shown in the table, **AD_CNN** has the lowest parameter count among the models while still predicting the full circumplex of perceptual attributes. Furthermore, its availability on GitHub and accessibility - confirmed via direct request to Hou et al. (2024) - makes it a strong baseline model, as this ensures reproducibility and reliable benchmarking. The first model, DNN, was also considered but is less suitable because it only predicts annoyance rather than the full circumplex.

In the context of this thesis, the original AD_CNN and SoundAQnet serve as benchmarks for comparison. AD_CNN is used as a foundation, as its architecture forms the basis for the models developed in this work. SoundAQnet, on the other hand, represents the current state-of-the-art in performance. However, its large size and high resource demands limit its suitability for real-time inference on low-cost, resource-constrained sensors. This highlights the need to explore lightweight architectures that maintain strong predictive performance while enabling deployment on low-cost, resource-constrained sensors.

Furthermore, loudness was not included in this thesis' research, as its contribution to the SoundAQnet model, while measurable, resulted in only a marginal performance improvement compared to using mel spectrograms alone.

3

Methodology

This chapter outlines the methodology used to address the research objectives of this thesis. It begins with a flow-diagram that illustrates the overall research process and explains how each sub-question is addressed. Next, the methodological approach for developing, training, and evaluating lightweight variants of the AD_CNN model, designed for perceptual soundscape analysis on low-cost, resource-constrained sensors, is presented. Finally, the generalisation study is introduced, detailing how the models will be evaluated for their ability to generalise to unseen data.

3.1. Research flow-diagram

The goal of this thesis is to develop parameter-reduced versions of the AD_CNN model, making them suitable for deployment on low-cost, resource-constrained sensors. In the previous section, it was shown that AD_CNN is the most parameter-efficient model capable of predicting both sound sources and perceptual attributes. To strategically reduce the parameter count in new model variants, and to determine how they can be effectively trained and evaluated, a clear methodology is required. This begins with identifying which components of the current AD_CNN architecture contribute most to its parameter count, and exploring how these can be optimized or simplified. After preprocessing the dataset, the parameter-reduced models will be trained, tested, and evaluated. The two best-performing models will then be combined into a hybrid architecture, which will also be evaluated. Together with benchmark models, this process addresses sub-question 2. Finally, the best-performing model will be evaluated, alongside benchmarks, on a completely different dataset with distinct audio recordings. This evaluation will provide insight into the generalisation capabilities of the models used in soundscape research, addressing sub-question 3.

The research flow-diagram is presented in Figure 3.1.

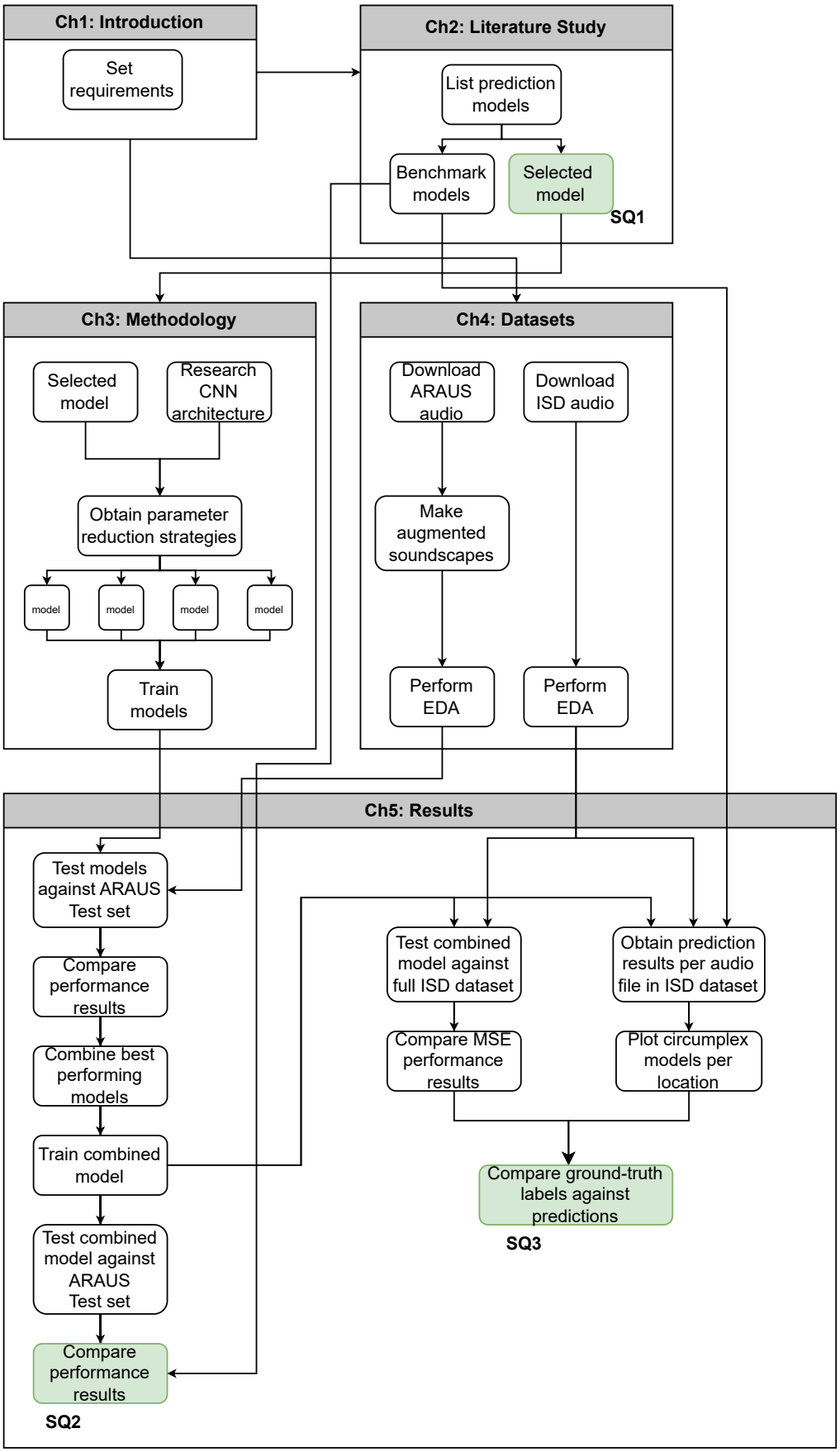


Figure 3.1: Research flow-diagram.

3.2. Model architecture

This thesis aims to design lightweight neural network models suitable for deployment on low-cost, resource-constrained sensors. As outlined in Chapter 1, the design process will adhere to the predefined requirements. In particular, the models developed in this thesis will comply with the parameter constraint, maintaining a maximum of approximately 300,000 parameters.

The AD_CNN model, which serves as a performance benchmark in the study of Hou et al. (2024), was originally developed by Ooi et al. (2024). While it achieves strong predictive performance, with an Area Under the Curve (AUC) of 0.84 on acoustic event classification (AEC), and an average Mean Squared Error (MSE) of 1.128 on perceptual attribute regression, the model contains approximately 520K parameters, making it too resource-intensive for direct deployment on low-cost, resource-constrained sensors.

With the AD_CNN model publicly available through the open-source GitHub repository SoundSCaper[1], it became feasible to adapt and compress the original architecture into a more efficient variant. To this end, four different architectural strategies are proposed and evaluated in this thesis. First, the baseline AD_CNN is introduced, with a focus on identifying the components that have the largest contribution to its parameter count. Subsequently, each of the four reduction approaches is presented and discussed in detail.

3.2.1. AD_CNN (baseline)

The AD_CNN model is a convolutional neural network designed to analyse acoustic data represented as mel spectrograms. Initially, the mel spectrogram input of shape $[32 \times 3001 \times 64]$ is reshaped to $[32 \times 1 \times 3001 \times 64]$ by adding a channel dimension, allowing compatibility with 2D convolutional layers. Following this, it processes the input through three convolutional layers, each followed by *BatchNorm2d* and a *ReLU* activation function. *BatchNorm2d* (Ioffe & Szegedy, 2015) normalizes the activations of each mini-batch to stabilize and accelerate training, while *ReLU* (Rectified Linear Unit) (Nair & Hinton, 2010) introduces non-linearity by setting all negative values to zero. After the second and third convolutional layers, *MaxPooling2D* and dropout layers are applied to reduce dimensionality and prevent the model from overfitting. The resulting features are then flattened into a single vector and passed through a fully connected layer, which extracts higher-level information from the data. Finally, this information is used simultaneously to predict multiple outputs: acoustic scene classes {*public square, park, street traffic*}, event classes {*silence - human sounds - wind - water - natural sounds - traffic - sounds of things - vehicle - bird - outside, rural or natural - environment and background - speech - music - noise - animal*}, ISO-attributes {*Pleasantness* & *Eventfulness*}, and eight perceptual attributes {*pleasant, eventful, chaotic, vibrant, uneventful, calm, annoying, monotonous*}.

This architecture is illustrated in Figure 3.2. The values in square brackets represent tensor shapes in the format $[batch\ size \times number\ of\ channels \times time\ frames \times mel\ bins]$. The values such as 7×7 denote the convolutional kernel size. The number of output feature maps (or channels) after each block is shown in the blue boxes. *MaxPooling* layers indicate their pooling size, while *Dropout* layers specify their dropout rates.

Parameters of AD_CNN

Understanding the parameter count of a convolutional neural network such as AD_CNN is crucial for evaluating its complexity, computational cost, and suitability for deployment on resource-constrained platforms. The total number of learnable parameters in a CNN is determined by summing the contributions from its trainable layers, typically convolutional layers, fully connected (dense) layers, and batch normalization layers.

[1] GitHub repository: https://github.com/Yuanbo2020/SoundSCaper

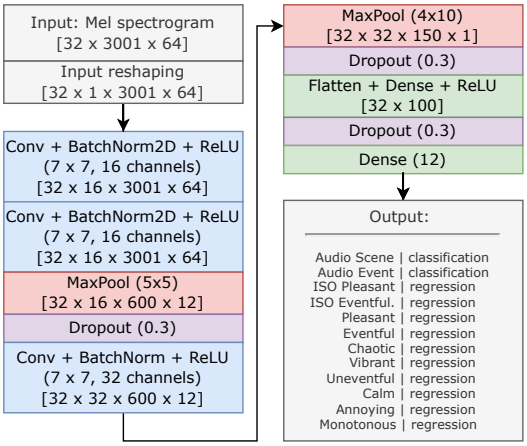


Figure 3.2: AD_CNN architecture by Hou et al. (2024), adapted from Ooi et al. (2024).

1. Convolutional Layers

Convolutional layers form the backbone of most CNN architectures. Each convolutional layer learns a set of filters that are applied across the spatial dimensions of the input. The number of learnable parameters in a 2D convolutional layer is given by:

$$Parameters_{\text{conv}} = (K_H \times K_W \times C_{in} + 1) \times C_{out} \tag{3.1}$$

Where:

- K_H: kernel height
- K_W: kernel width
- C_{in}: number of input channels
- C_{out}: number of output channels (filters)

The term $K_H \times K_W \times C_{\text{in}}$ corresponds to the number of weights in each filter, and the additional 1 accounts for the optional bias term. This bias term contributes only if bias is enabled in the layer (i.e., not set to `bias=False`). This formula applies to every convolutional layer in the network.

2. Fully Connected (Dense) Layers

Dense layers connect every input neuron to every output neuron. The number of parameters in such a layer is calculated as:

$$Parameters_{\text{dense}} = (N_{in} \times N_{out}) + N_{out} \tag{3.2}$$

where:

- N_{in}: number of input features
- N_{out}: number of output features

The term $N_{in} \times N_{out}$ represents the weight matrix, while N_{out} adds the biases.

3. Batch Normalization Layers

Batch normalization (BN) layers include two trainable parameters per channel: a scaling factor and a shift. Therefore, the total number of parameters is:

$$Parameters_{\text{BN}} = 2 \times C \tag{3.3}$$

where:

- C: number of channels

These parameters are learned to normalize and re-scale the feature maps, improving training stability and convergence.

4. Pooling Layers

Unlike the layers above, pooling layers (i.e. max-pooling) perform fixed mathematical operations without any trainable parameters. Therefore, they do not contribute to the total parameter count.

5. Total Parameter Count

The total number of parameters in AD_CNN is the sum of the parameters across all trainable layers $l = 1$ to L:

$$Total\ Parameters = \sum_{l=1}^{L} Parameters_l \tag{3.4}$$

where:

- L: number of trainable layers
- $Parameters_l$: number of parameters in layer l

Applying the formulas reveals that the AD_CNN model developed by Hou et al. (2024) contains a total of 521,472 parameters. For a detailed breakdown of the parameter calculation, refer to Appendix B, Table B.1. This indicates that there is significant potential to compress the model further to meet the constraints of a lightweight architecture suitable for deployment on low-cost, resource-constrained sensors.

3.3. Four parameter reduction strategies

The main bottleneck in terms of the number of parameters is clearly identifiable in the first dense layer, where the feature maps are flattened and passed through a fully connected layer. This layer alone accounts for approximately 480,100 parameters, making it the most significant contributor to the overall model size. Consequently, this dense layer presents the most promising opportunity for parameter reduction or compression to meet tighter resource constraints. Therefore, four different approaches will be explored, each contributing to a reduction in the dense layer's parameter count:

1. Decrease the number of convolutional filters
2. Decrease the fully connected layer
3. Decrease time frames via larger hop size
4. Increased MaxPooling kernel sizes

Furthermore, the scene classification and ISO (*Pleasantness* & *Eventfulness*) regression layers were removed, as they fall outside the scope of this thesis. The focus is limited to sound source classification and the prediction of the eight perceptual attributes. Other architectural components, such as kernel sizes, batch normalization, and dropout, were kept unchanged to isolate and clearly observe the impact of the aforementioned modifications.

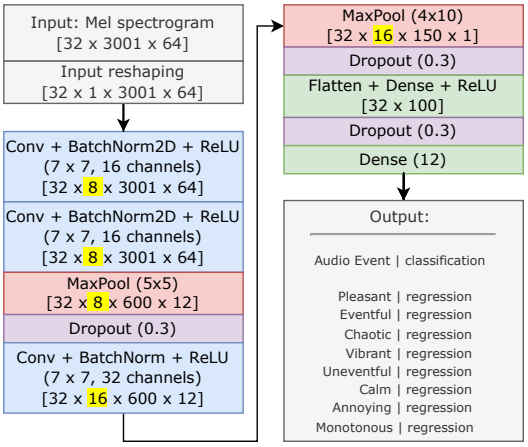


Figure 3.3: AD_CNN_decreased_filters architecture.

3.3.1. Reducing convolutional filters (AD_CNN_decreased_filters)

This model reduces complexity by decreasing the number of filters in each convolutional layer. Specifically, the output channels are reduced (e.g., from 16/32 to 8/16), resulting in a significantly smaller feature representation and a reduced fully connected input (2400 vs. 4800 in the original). The final dense layer still uses 100 units and continues to perform both event classification and perceptual attribute regression. The architecture is illustrated in Figure 3.3, with all modifications clearly highlighted for ease of identification.

The parameter calculations for this adapted version of the AD_CNN model are provided in Appendix B, Table B.2. Since the number of filters in the convolutional layers has been reduced by half, the output fed into the dense layer is also halved, which results in a flattened size of $32 \times 16 \times 150 = 2400$, compared to the previous 4800. As a result, the first dense layer now contains $240{,}100$ parameters. The total number of parameters in the modified model amounts to $252{,}287$.

3.3.2. Smaller dense layer (AD_CNN_dense_layer)

As stated in the overarching subsection, the goal was to reduce the size of the first dense layer, as it accounts for the majority of the model's parameters. In the original model, this layer contains 100 hidden units, resulting in a parameter count of $4800 \times 100 + 100 = 480{,}100$. The revised model's number of hidden units are reduced by half, which brings the total to $4800 \times 50 + 50 = 240{,}050$ parameters. A smaller hidden layer means all subsequent output heads receive fewer inputs, further reducing their parameter counts. While this design leads to a more compact model, it may also limit the expressiveness of the learned representations and potentially reduce prediction performance.

This modification only slightly alters the original model, but for completeness, the updated architecture is illustrated in Figure 3.4. The total number of parameters is now reduced to 279,767. A detailed overview of the parameter calculations can be found in Table B.3 in Appendix B.

3.3.3. Increased hop length (AD_CNN_hop_length)

To reduce the input size and consequently decrease the number of parameters in the dense layer, this approach focuses on reducing the temporal dimension of the mel spectrograms. The very first input to the model is the batch of mel spectrograms, which leads to an input shape of [$batch\ size$, $time\ frames$, $mel\ bins$]. While the number of mel bins is fixed at 64, consistent with the original model and widely used across audio-related research, the number of time frames can be adjusted. This provides an effective method to reduce input size and computational load without compromising frequency resolution, since this will result in smaller feature maps and fewer parameters in the subsequent dense layer.

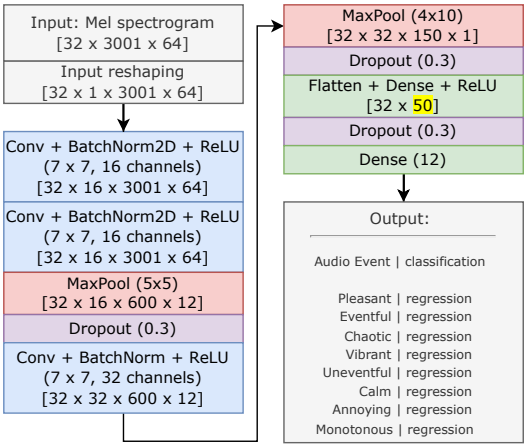


Figure 3.4: AD_CNN_dense_layer architecture.

The spectrograms are generated using the `Spectrogram` module from `torchlibrosa` (Kong et al., 2020), which internally performs the STFT using `librosa`'s implementation (McFee et al., 2021). This means the frame computation behaviour, including padding and windowing, follows the same logic as in `librosa.stft`. During STFT computation, time frames are created. Those represent the number of temporal segments, each frame corresponds to one column in the resulting spectrogram.

The number of time frames depends on the length of the audio in samples, which can be calculated as:

$$L = audio\ length\ (seconds) \times sample\ rate$$

If the signal is centred (`center=True`), it is padded with zeros to allow for symmetric framing. This padding ensures that each frame is centred around its corresponding time step. Specifically, the signal is padded with $n_{\text{fft}}//2$ samples at both the beginning and the end. This approach allows the first and last STFT windows to be fully aligned with the edges of the signal, avoiding truncation and maintaining symmetry in the time–frequency representation.

After padding, the total number of time frames in the spectrogram is given by:

$$n_{\text{time frames}} = 1 + \left\lfloor \frac{L + 2 \cdot pad - n_{fft}}{hop\ length} \right\rfloor$$

As this formula shows, increasing the hop length leads to fewer time frames, thereby reducing the input size. Specifically for the data used in this thesis, doubling the hop length from 160 to 320 results in halving the number of input time frames, from 3001 to 1501. When the original model architecture is kept unchanged, this reduction in temporal resolution leads to smaller intermediate feature maps throughout the convolutional layers, and consequently, a smaller input to the first dense (fully connected) layer, as illustrated in Figure 3.5. The final model comprises 280,967 parameters. By reducing the size of the first dense layer by half, the total number of parameters has been significantly decreased. A detailed breakdown of the parameter calculations is provided in Table B.4 in Appendix B.

3.3.4. Larger MaxPooling operations (AD_CNN_harder_max_pooling)

Unlike previous adaptations, this version retains the original convolutional layers and input dimensions but modifies only the MaxPooling operations, which play a critical role in determining the dimensionality of the feature maps passed to the dense layers. Specifically, the model now applies a 10×5 pooling operation after the second convolutional layer and a 16×12 pooling operation after the third. These more aggressive pooling steps substantially reduce the spatial dimensions of the feature maps to

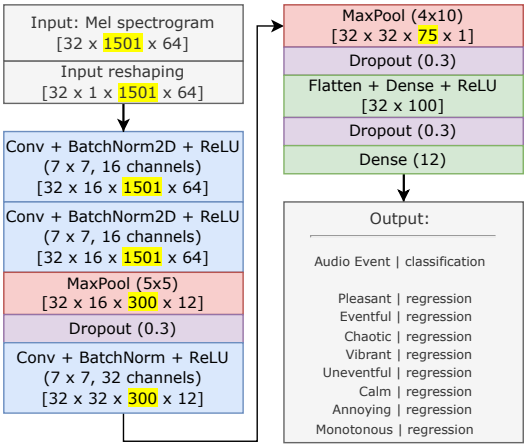


Figure 3.5: AD_CNN_hop_length architecture.

$[32, 32, 18, 1]$, resulting in a flattened input size of only 576 ($32 \times 18 \times 1$) for the fully connected layer, compared to 4800 in the original model. This structural change alone reduces the total number of trainable parameters to 98,567, making the architecture the most lightweight among all model variants explored. The architecture and its modifications are illustrated in Figure 3.6. Detailed parameter calculations can be found in Appendix B, Table B.5.

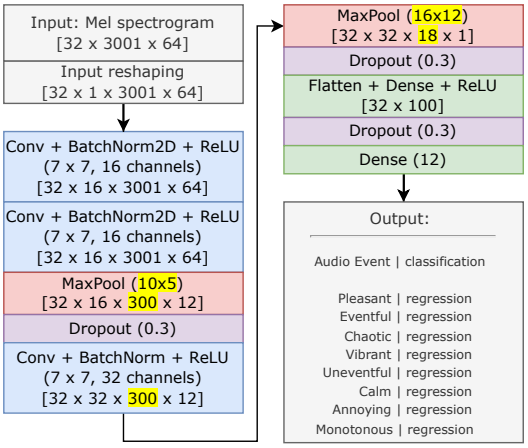


Figure 3.6: AD_CNN_harder_max_pooling.

3.4. Training pipeline of the models

Training is executed using the script provided in the SoundSCaper GitHub repository. The process begins by configuring the GPU environment and utilizing CUDA for accelerated computation; if CUDA is unavailable, training defaults to the CPU. The script loads all required training and validation data through the data-generator script and incorporates early stopping using. The models are trained on the ARAUS dataset (Ooi et al., 2024), which was also used to train AD_CNN and SoundAQnet. Chapter 4 provides a more detailed description of the dataset.

3.4.1. Audio preprocessing: WAV to mel spectrograms

Before model training can commence, the input features must be generated, typically in the form of spectrograms derived from audio files. All audio files are originally in WAV format and are converted to mono for consistency and model compatibility. This conversion is performed using a Python script from the SoundScaper GitHub repository. The script loads each file from the source directory, downmixes it to a single channel using the `librosa` library, and saves it as a WAV file with a sampling rate of 44.1 kHz. It also includes functionality to automatically create the output directory if it does not exist, ensuring a clean and organized processing pipeline. This step was crucial to avoid inconsistencies that could arise from multi-channel recordings during feature extraction and model training. The result of this process was a folder containing 25,440 mono-channel WAV files, each approximately 2.5 MB in size, precisely half the size of the original multi-channel files.

To prepare the audio data for model training and evaluation, the raw `.wav` files must be converted into log-mel spectrograms. The process, illustrated in Figure 3.7, begins by segmenting the audio signal $x(n)$ into overlapping frames using a sliding window of fixed length N_a and hop size L_a. Each frame is multiplied by a window function $w_a(n)$ to reduce spectral leakage. A *Short-Time Fourier Transform (STFT)* (Allen, 1977) is then applied to each windowed frame, producing a complex-valued spectrogram $X_a(n_a, k'_a)$, where n_a is the time frame index and k'_a the frequency bin index.

Next, the magnitude spectrogram is mapped onto 64 mel-frequency bins (k_a) using a mel filterbank spanning 50 Hz to 8000 Hz. This transformation mimics the human auditory system, which is more sensitive to changes in lower frequencies than in higher ones. Importantly, this sensitivity decreases logarithmically rather than linearly with increasing frequency. Mel-frequency bins are perceptually motivated frequency bands that reflect this nonlinear sensitivity to pitch, resulting in finer resolution at low frequencies and coarser resolution at high frequencies. The use of 64 mel bins is a common practice in sound analysis with machine learning, offering a good balance between spectral resolution and computational efficiency. It also matches the configuration used in baseline models for the ARAUS

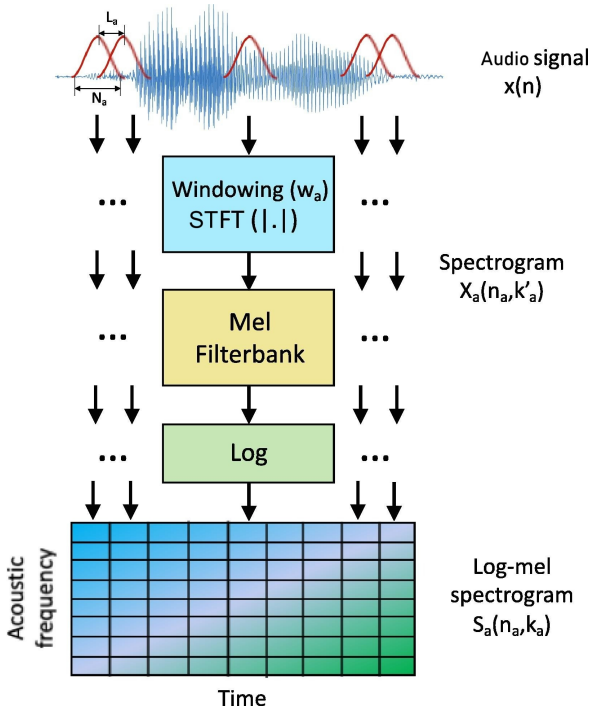


Figure 3.7: Block diagram of the log-mel spectrogram computation process, adapted from Gallardo-Antolín and Montero (2021), with DFT replaced by STFT, and speech signal replaced by audio signal.

dataset, ensuring compatibility and comparability. A logarithmic compression is then applied to obtain the final log-mel spectrogram $S_a(n_a, k_a)$, a compact time-frequency representation. An example of a mel spectrogram computed for one of the samples in the ARAUS dataset is given in Figure 3.8. The recording is characterized by a relatively constant background sound across most of the duration. A notable increase in low-frequency energy occurs between approximately 9 and 16 seconds, visible as a yellowish region in the lower frequency bands, indicative of higher decibel levels in that range. Listening to the corresponding audio file (available via this thesis' GitHub repository [2]) indicates that this section coincides with the onset of wind noise, confirming the visual observation. Additionally, distinct high-frequency components appear at seconds 18, 22, 26, and 29. These brief patterns correspond to bird vocalizations.

This thesis adopts the feature extraction configuration used by Hou et al. (2024), which follows the settings from Kong et al. (2020). Features are computed with a 32 ms window (N_a) and a 10 ms hop size (L_a), corresponding to 512 and 160 samples, respectively, at a 16 kHz sampling rate, using a Hann window (Harris, 1978) is applied during the STFT computation to reduce spectral leakage and smooth transitions between overlapping segments. This windowing function transitions smoothly to zero at both ends, providing better frequency analysis compared to rectangular windowing. The resulting log-mel spectrograms have a shape of (3001, 64) for each 30-second audio segment and are saved as NumPy array files (`.npy`).

3.4.2. Training pipeline modifications

Since all necessary files were available in the original GitHub repository, the training pipeline could be replicated. However, several targeted adaptations were required to align the codebase with the goals of this thesis. First, the configuration file was extended with a new argument to support variable hop lengths, allowing dynamic selection of the appropriate preprocessed pickle files. This change ensures that by adjusting a single argument, the model seamlessly loads spectrograms with different temporal resolutions, particularly relevant for evaluating models like AD_CNN_hop_length.

Corresponding modifications had to be made in the data loading process to respond to this argument

[2] GitHub repository: https://github.com/pherfkens/Thesis

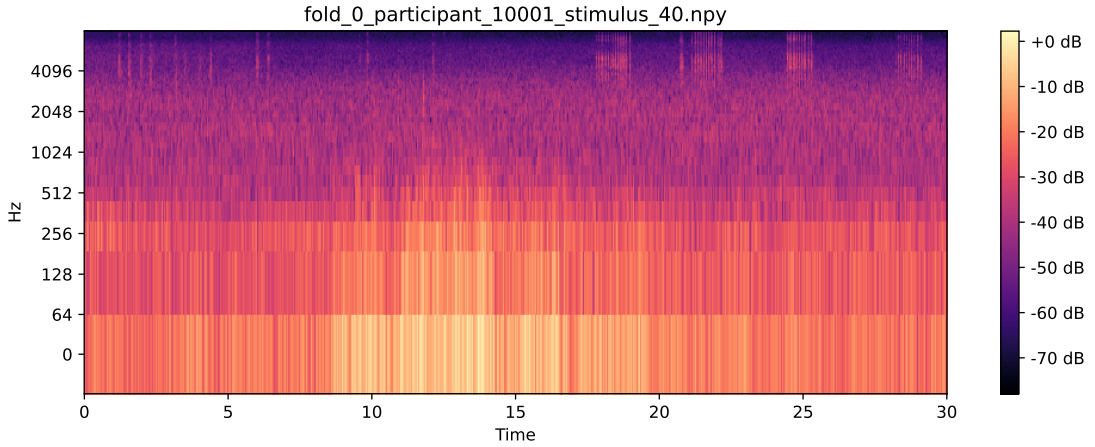


Figure 3.8: Random example of log-mel spectrogram.

and load the correct feature and normalization files accordingly. This loading process was further streamlined to yield only the necessary outputs for this study: acoustic events and perceptual attributes. Outputs irrelevant to this thesis, such as acoustic scene labels and ISO-defined metrics, were excluded from the generator functions.

All training and validation metrics, along with timing information, were logged to CSV files to support post hoc analysis and ensure reproducibility. This setup allowed for efficient experimentation across a range of models while maintaining consistency in data handling and evaluation. All modifications are clearly marked with comments in the version maintained on this thesis' GitHub repository.

3.4.3. Data handling and normalization

The data-generator, available from the original GitHub repository, was thus adapted to meet the specific objectives of this thesis. It efficiently manages the loading and preprocessing of datasets during the training, validation, and testing phases of model development. It is responsible for reading preprocessed input data stored in pickle files for each dataset subset and ensures consistent data formatting throughout the pipeline. Additionally, it manages normalization parameters by utilizing provided normalization files included in the GitHub repository, ensuring reproducibility of results by maintaining the same normalization strategy across different experimental runs.

In cases where models require modified spectrogram parameters, such as the AD_CNN_hop_length model, which employs a different hop length setting, distinct pickle files are generated. Consequently, separate normalization parameters are needed. This process is carried out as follows:

To normalize the input features, the script computes the mean and standard deviation of the training set features, along the time and sample axes. This is done using a z-score standardisation approach, where each feature value is rescaled according to $x_{norm} = (x - \mu)/\sigma$, with μ and σ representing the computed global mean and standard deviation, respectively. These statistics are calculated only once and saved as `.pickle` files in a designated normalization directory. If normalization files already exist, they are loaded to ensure consistency across training, validation, and testing. This normalization ensures that the input data fed to the model remains within a comparable scale across different runs and subsets, thereby improving numerical stability and convergence during training.

Finally, the data-generator yields methods for training, validation, and testing. These methods yield data batches that may include mel spectrogram features, acoustic event labels, and perceptual attributes. To enhance model robustness and mitigate overfitting, these generators incorporate randomization and shuffling mechanisms during data batching. Overall, this modular and automated data handling system facilitates a streamlined and reproducible workflow for model training, evaluation, and deployment.

3.4.4. Early stopping

The training loop includes early stopping, periodic validation, and automatic saving of the best-performing model based on a specified validation metric. The model is optimized using the Adam optimizer (Kingma & Ba, 2017), with a learning rate set to 1×10^{-4}, consistent with prior work on the AD_CNN architecture (Ooi et al., 2024).

Early stopping is implemented to prevent overfitting and unnecessary computation. This mechanism monitors the performance of the model on a specified validation metric. If no improvement is observed for 10 consecutive validations, training is halted. In this thesis, the chosen monitor is the perceptual attribute *pleasant*, represented as a regression task. This replaces the previously used ISO-defined *Pleasantness*, since the current model does not incorporate both ISO *Eventfulness* and *Pleasantness*. When the monitor is set to one of the perceptual attributes, performance is evaluated using mean squared error (MSE), where lower values indicate better performance. If instead the monitor is set to 'event', which tracks sound event classification, the evaluation metric switches to area under the ROC curve (AUC), where higher values are preferred.

During training, the loss values for both sound event classification and the eight perceptual attribute regressions are printed at each iteration. The number of iterations per epoch is determined by the dataset size and the batch size, which is set to 32 in this thesis. After each fully completed epoch, the model is evaluated on the validation set, and relevant metrics such as AUC (for sound event classification) and MSE (for perceptual attributes) are recorded. All models are trained for a maximum of 100 epochs unless the early stopping criterion is met.

3.5. Evaluation metrics and model testing

As mentioned in the previous section, the perceptual attributes are evaluated using the MSE. For each attribute, an individual MSE is computed. These values can subsequently be averaged to obtain an overall performance metric for the model. MSE is a widely used metric in regression tasks, as it quantifies the average squared difference between the predicted values and the corresponding ground truth values. Formally, it is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where n is the number of samples, y_i is the true value, and $\hat{y}_i$ is the predicted value for the i-th sample.

The main difference between the MSE and the Root Mean Squared Error (RMSE) lies in the final transformation: MSE does not take the square root of the squared error terms, which results in larger errors being penalized more heavily. A lower MSE indicates better model performance. This thesis adopts MSE as the primary evaluation metric for the perceptual attributes, as the work by Hou et al. (2024) also employed MSE to assess model performance. This consistency facilitates a more straightforward comparison between the results presented in this study and those in prior research.

In addition to MSE, this thesis also uses the Area Under the Curve (AUC) of the ROC as a performance metric to evaluate the classification performance of the model on acoustic events. AUC quantifies the model's ability to distinguish between classes and is especially useful in multi-label classification settings. For each sample, the true binary labels and predicted probabilities are extracted. If the sample contains at least one positive label, the AUC score is computed for that sample using the `roc_auc_score` function. The final AUC metric is obtained by averaging the individual scores across all valid samples. This approach ensures that AUC is only calculated where it is meaningful and avoids distortions due to empty ground truth vectors. A higher AUC indicates better classification performance, with a value of 1 representing perfect discrimination and 0.5 indicating random guessing.

To evaluate model performance, the model is applied to the test dataset to generate predictions for both acoustic events and perceptual attributes. Evaluation metrics include the area under the ROC curve (AUC) for sound event classification and mean squared error (MSE) for each of the eight perceptual attributes. The average MSE across these attributes serves as an overall indicator of regression performance.

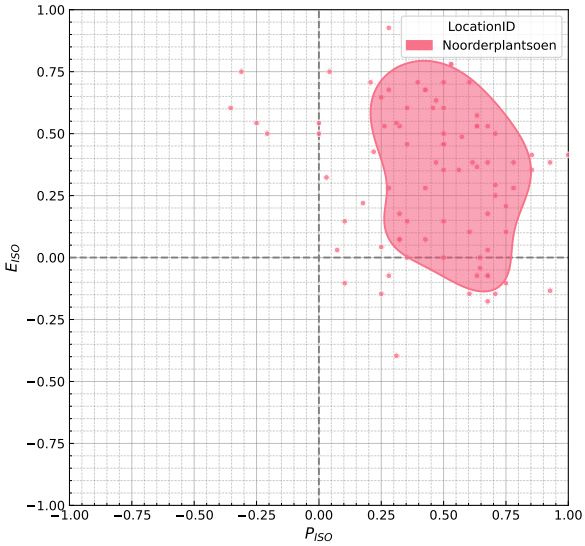


Figure 3.9: Questionnaire responses mapped in ISO-pleasantness/eventfulness space for Noorderplantsoen, Groningen.

3.6. Combining the two best-performing models

Based on these results, the two best-performing models, ranked by their MSE scores, are combined through ensemble learning to enhance predictive accuracy while maintaining computational efficiency. This model fusion approach is designed to capitalise on the complementary strengths of the individual models, particularly in predicting perceptual attributes with a reduced parameter footprint. The resulting ensemble model forms the basis for answering sub-question 2 and is also deployed in the generalisation study described in the next section.

3.7. Generalisation study

To evaluate how well the soundscape prediction models generalise to entirely unseen data from a different source, a generalisation study will be conducted. This study will include not only the ensemble model but also the baseline model, AD_CNN, and the benchmark model, SoundAQnet. The external dataset used for this evaluation is the International Soundscape Database (ISD). A more detailed description of this dataset is provided in Chapter 4.

Given that the ISD contains several locations with various responses per location, ISO *Pleasantness* and ISO *Eventfulness* can be computed accordingly for each response. These values can then be plotted on the two-dimensional circumplex model, with the *Soundscapy* package (Mitchell, Aletta, & Kang, 2022b). Figure 3.9 presents such a visualisation for *Noorderplantsoen, Groningen*.

In these circumplex plots, each dot represents an individual's perceptual assessment of the soundscape, transformed into the ISO coordinate system. The shaded area represents the contour of 50% of all responses, meaning that half of the responses fall inside it, and the other half outside. This approach highlights not only the central tendency but also the diversity of perception in a given location, offering insight into how consistently or variably a space is experienced by different people. In the results section, for each city, both the *ground-truth* and the *predicted* circumplex diagrams will be overlaid to facilitate comparison. This will help evaluate how well the models capture not just average perceptual qualities, but also the shape and spread of human soundscape perception. These plots will thus contribute to answering sub-question 3.

4

Datasets

This chapter introduces the datasets used in this thesis. The ARAUS dataset serves as the foundation for model training and validation, providing a large collection of soundscape recordings with perceptual attribute scores and audio event annotations. The International Soundscape Database (ISD) is used to evaluate model generalisation to new acoustic environments. Key preprocessing steps, exploratory analyses, and dataset comparisons are presented to highlight the characteristics and relevance of each dataset for predicting perceptual attributes.

4.1. ARAUS dataset

As the AD_CNN and SoundAQnet were both trained on the ARAUS dataset, using the same dataset ensures accurate benchmarking, as the entire training pipeline can be consistently replicated. The Affective Responses to Augmented Urban Soundscapes (ARAUS) dataset is a large-scale dataset that captures human responses to systematically modified urban sound environments (Ooi et al., 2024). According to Ooi et al. (2024) this was necessary because existing datasets lack subjective labels that reflect how listeners perceive the affective quality of the environments. This absence restricts their suitability for soundscape studies, as understanding human perception is essential for analysing and accurately modelling people's perception of a soundscape.

The ARAUS dataset includes over 25,000 human ratings collected through laboratory-based listening experiments using 30-second audiovisual stimuli. These stimuli were created by augmenting real-world urban recordings, base scenes from the Urban Soundscapes of the World (USotW) database (De Coensel et al., 2017), with additional sound elements, such as birdsong, water flow, construction noise, traffic, wind or silence, also called maskers. The augmented soundscapes were presented at different relative loudness levels to simulate valid environmental scenarios. The evaluation protocol follows ISO/TS 12913-2:2018 (ISO, 2018) guidelines and incorporates the Affective Response Questionnaire (ARQ), which measures responses across nine perceptual attributes: *pleasant, annoying, eventful, uneventful, vibrant, monotonous, chaotic, calm,* and *appropriate*. This includes the eight perceptual attributes that are the focus of this thesis. Furthermore, the dataset includes labels for scene categories, sound sources, and derived ISO metrics such as *Pleasantness* and *Eventfulness*. A total of 605 participants contributed to the study, and additional demographic and psychological data were also collected. These responses offer a rich foundation for perceptual modelling of soundscape quality. To support machine learning applications, the dataset is divided into five cross-validation folds and an independent test set.

4.1.1. Extending ARAUS for model training

To develop and train the models AD_CNN and SoundAQnet, Hou et al. (2024) utilized the ARAUS dataset. The models were designed to predict acoustic scenes (e.g., public square, park, street traffic), audio events (i.e., sound sources), and perceptual attributes such as *pleasant*, *eventful*, and *annoyance*. Although the ARAUS dataset was suitable for perceptual attribute prediction, it lacked scene labels and included only six annotated audio events. Given the complex acoustic environments in the Urban

Sound of the World (USotW) dataset, the limited event labels were insufficient for training a robust model.

To address this limitation, the authors leveraged the PANN (Pretrained Audio Neural Network) model introduced by Kong et al. (2020), which was trained on AudioSet and can recognize 527 distinct audio event categories. Each 30-second audio clip was segmented into one-second intervals, and PANN was applied to extract per-segment probabilities for each event class. The threshold was set at 0.5 to yield binary presence or absence labels for each event per segment. By aggregating the event detections across all segments, the researchers identified the most frequently occurring classes and selected the 15 most relevant audio event categories for their task. These categories were: bird - animal - wind - water - natural sounds - vehicle - traffic - sounds of things - environment and background - outside, rural or natural - speech - human sounds - music - noise - silence. For model training, only clip-level binary audio event labels were required. To obtain these, PANN was re-applied to the full 30-second clips, resulting in a single 527-dimensional probability vector per clip. From this vector, the probabilities associated with the 15 selected event categories were extracted and binarized using a lower threshold of 0.1, producing the final clip-level labels used for training SoundAQnet.

In the original ARAUS experiment (Ooi et al., 2024), the dataset was divided such that the validation set contained 5,040 samples, while the test set comprised only 48 samples. The big imbalance in the data and the limited size of the test set made it hard to accurately check how well the model would work on new data. To overcome this issue and make the evaluation more robust, the dataset was randomly mixed and then split again into new training, validation, and test sets.

The revised dataset configuration consists of 25,248 unique 30-second binaural audio clips, with strict enforcement of non-overlapping subsets. From this collection, 19,152 clips (approximately 75%) were allocated to the training set, 2,520 clips (10%) to the validation set, and the remaining 3,576 clips (15%) to the test set. This new split ensures a more balanced and statistically representative distribution of samples, thereby enabling a more reliable and meaningful assessment of model performance on unseen data.

This thesis follows the training pipeline of the AD_CNN and SoundAQnet models, using the same training, validation, and test sets. The corresponding `.txt` files were provided on the SoundSCaper GitHub repository. These include the following files for each of the training, validation, and testing splits:

- *..._set_audio_file_ids.txt* – provided on GitHub, contains audio file IDs
- *..._acoustic_scene_labels.txt* – provided on GitHub, contains scene label
- *..._set_audio_event_labels.txt* – provided on GitHub, contains event labels
- *..._set_PAQ_8D_AQs.txt* – provided on GitHub, contains labels for eight perceptual atttributes
- *..._set_masker.txt* – generated to indicate the corresponding masker for each audio file
- *..._set_soundscapes.txt* – generated to indicate the corresponding soundscape for each audio file
- *..._set_appropriate_leql_leqr.txt* – generated to specify the appropriate affective quality and corresponding sound level for each audio file

The audio files themselves had to be downloaded manually. The download pipeline provided in the ARAUS GitHub[3] repository was followed. Downloading and augmenting all audio resulted in a dataset totalling 132 GB of raw `.wav` files.

The pickle script and all `.txt` files can be found on this thesis' GitHub. An overview of the repository's homepage is provided in Appendix A.

4.1.2. Explanatory Data Analysis

To ensure a representative and balanced distribution of perceptual information across the training, validation, and test sets, an exploratory analysis was conducted on the eight perceptual attributes, or Perceived Affective Quality (PAQ) attributes: *pleasant*, *eventful*, *chaotic*, *vibrant*, *uneventful*, *calm*, *annoying*, and *monotonous*. By analysing and visualising their distributions across the dataset splits, it

[3] GitHub repository: https://github.com/ntudsp/araus-dataset-baseline-models

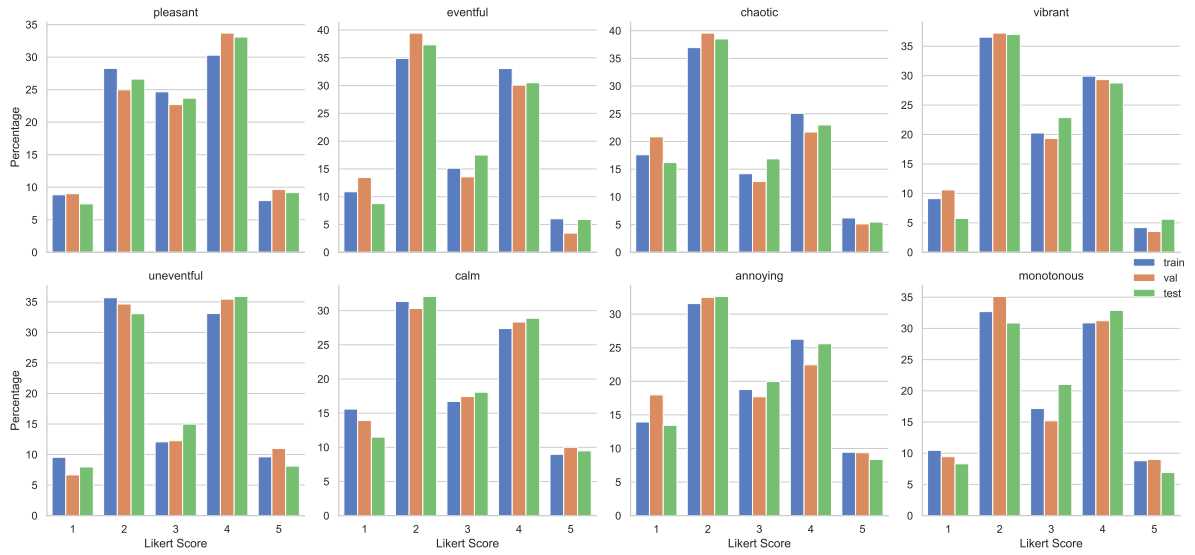


Figure 4.1: Bar plots showing the distribution of the eight perceptual attributes across the training, validation, and test sets.

is possible to assess the presence of imbalances or biases that could impact model generalisation. Furthermore, the sound source events are also briefly explored, as these will be among the target outputs of the parameter-reduced models. The dataset includes 15 distinct sound source categories, and analysing their distribution provides insight into the diversity and balance of the acoustic events that the model is expected to predict.

The original ARAUS paper also conducted an exploratory data analysis to verify data quality and empirical consistency with known literature. It found that the sound maskers used to augment the soundscapes aligned with established perceptual effects: for instance, bird and water sounds generally increased perceived pleasantness, while traffic, construction, and wind tended to decrease it. Interestingly, the negative effect of wind on pleasantness was somewhat unexpected but was partly attributed to the absence of physical sensations in laboratory listening conditions. Overall, the ARAUS dataset successfully spans the entire ISO *Pleasantness* and ISO *Eventfulness* space, covering all quadrants, making it well-suited for perceptual soundscape research.

Perceptual attributes

Figure 4.1 shows the distribution of the eight perceptual attributes across the training, validation, and test sets. To account for the larger size of the training set, all distributions have been normalized. The distributions are largely consistent across all three sets, suggesting that the data splitting process preserved the perceptual diversity of the dataset. Each attribute spans the full range of the 5-point scale, with notable density around the midpoints (2–4). Interestingly, the rating of 3 appears less frequently than 2 and 4. This may be because 3 represents a neutral response, which participants could have found less expressive or meaningful when describing their perception of a soundscape. The similarity in shape and central tendency across sets for each attribute implies that the model will encounter a comparable perceptual soundscape during training and evaluation, reducing the risk of performance bias due to distributional shifts.

Sound source events

Figure 4.2 presents the overall frequency of audio event categories across the full dataset. As can be seen, most sound sources occur more than 10,000 times, with *Animal*, *Sounds of things*, *Speech*, and *Human sounds* appearing most frequently. In contrast, events such as *Silence*, *Wind*, and *Outside, rural or natural* appear relatively infrequently, suggesting that certain ambient or less intrusive sounds are under-represented. As a result, these under-represented classes may be less likely to be accurately predicted by the models, particularly in multi-label classification tasks where class imbalance can significantly impact performance.

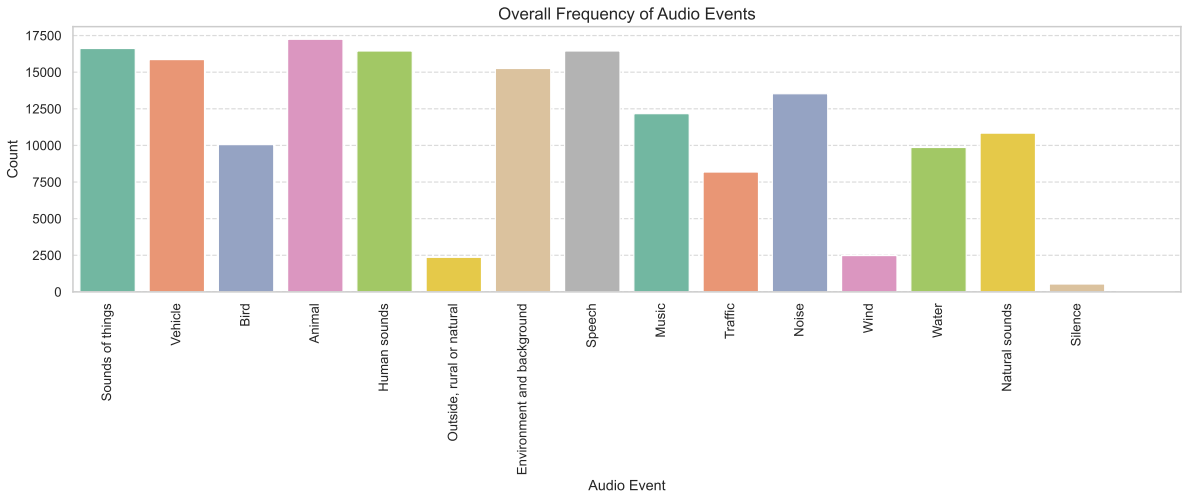


Figure 4.2: Audio events across entire dataset.

4.2. The International Soundscape Database

To assess the generalisability of models trained on the ARAUS dataset, their performances are evaluated on a different dataset. An appropriate candidate is the International Soundscape Database (ISD), which also consists of 30-second audio recordings in diverse urban settings across Europe and China, accompanied by subjective evaluations using Likert-scales (Mitchell et al., 2024).

4.2.1. Explanatory Data Analysis

Although the dataset includes questionnaire responses for Chinese cities, the corresponding `.wav` audio files are unavailable. Therefore, only recordings for which both the audio and complete perceptual attribute responses (i.e., no NaN values) are available are included in the analysis. Additionally, since multiple responses exist for each recording, the mean of these responses is computed per audio file. This results in a total of 812 usable audio files, each with corresponding averaged perceptual evaluations. The distribution of these attributes is shown in Figure 4.3.

Compared to the distribution presented for the ARAUS dataset (Section 4.1.2), a violin plot is now used, as the values are continuous (due to averaging) rather than strictly integer-based. As shown, the attributes *pleasant*, *eventful*, and *vibrant* tend to skew towards higher values, whereas *uneventful*, *annoying*, and *monotonous* generally exhibit lower scores. The only attributes that are relatively evenly distributed are *chaotic* and *calm*.

4.2.2. Model adjustments for generalisation study

To evaluate the models on the ISD dataset, several modifications must be made to accommodate its limited label availability. Audio clips need to be standardised to a fixed duration by mirroring shorter clips and truncating longer ones. Only files listed in the ISD metadata should be included. Since the dataset lacks labels for scenes, events, and ISO attributes, dummy variables must be inserted during data generation to maintain compatibility with existing model architectures. Loudness features, used in SoundAQnet, should also be replaced with placeholder values due to the absence of calibration data required for accurate estimation. During evaluation, only the regression outputs for the eight perceptual attributes are retained; all classification-based metrics must be disabled.

To obtain usable outputs from the models, inference should be performed on a per-audio-file basis rather than as a batch performance evaluation. The inference scripts must be adapted to output predictions directly for each file, bypassing any requirement for ground-truth labels. For consistency across models, only the acoustic event probabilities and eight perceptual attribute predictions should be considered. These outputs should be saved in a structured format to allow post-processing, including the calculation of *Pleasantness* and *Eventfulness* scores and their visualisation on a 2D circumplex model.

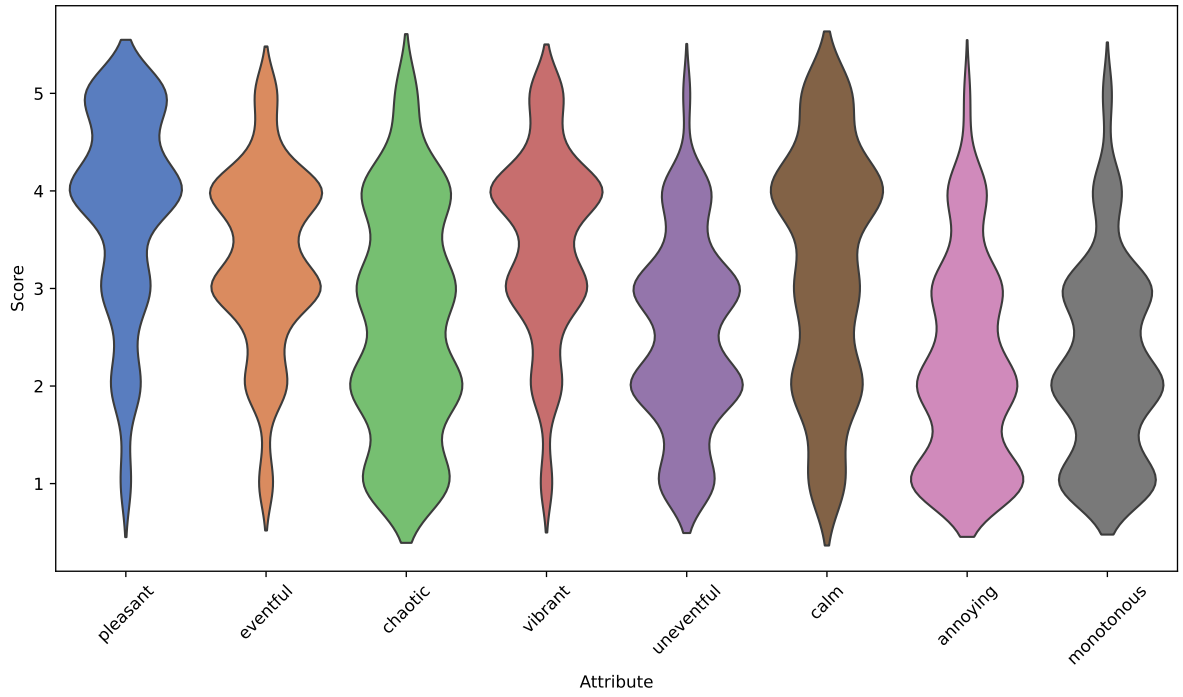


Figure 4.3: Distribution of Perceptual Attributes in International Soundscape Database.

4.3. Dataset comparison: ARAUS vs. ISD

To compare with the ARAUS training dataset used for model development, the average values of the perceptual attributes were analysed and are shown in Table 4.1. The ISD dataset shows higher average scores for *pleasant*, *eventful*, and *vibrant*, suggesting that it represents a more stimulating and positively perceived sound environment. On the other hand, the average values for *uneventful*, *annoying*, and *monotonous* are lower, indicating fewer negative perceptions.

This difference shows a clear shift in perception between the two datasets. The ARAUS dataset appears to be more neutral, while the ISD dataset is generally more positively rated. As a result, a model trained only on ARAUS data might underestimate positive attributes like *pleasant* and *vibrant*, and overestimate negative ones such as *annoyance* and *monotonous* when tested on ISD data.

Table 4.1: Mean perceptual attribute scores in ARAUS training set and ISD, and their differences.

Dataset	Pleas.	Eventf.	Chaot.	Vibra.	Uneventf.	Calm	Annoy.	Monot.
ARAUS Training set	3.00	2.88	2.65	2.84	2.98	2.83	2.86	2.95
ISD	3.78	3.30	2.65	3.47	2.52	3.20	2.20	2.24
Difference (ISD - ARAUS)	≈ 0.78	≈ 0.41	≈ 0.00	≈ 0.63	≈ -0.46	≈ 0.37	≈ -0.66	≈ -0.71

5

Results

This chapter presents the results of all experiments conducted in this thesis. It covers the training process, evaluation of four lightweight neural network models, performance benchmarking against existing baselines, and the development of a new ensemble model. A generalisation study is introduced to demonstrate generalisability.

5.1. Model benchmarking overview

To evaluate model performance, all models described in Chapter 3 were tested on a held-out test set consisting of 3,576 samples from the ARAUS dataset. For each model, both the AUC and MSE were computed. Furthermore, the mean of the MSE values was calculated to provide an overall indication of regression performance. Finally, inference time is measured by isolating the pure model forward pass computation. The computer used comprises an i7-8750H CPU @ 2.20GHz and a NVIDIA Quadro P1000 GPU with 4GB RAM. For each test sample, the mel spectrogram input is fed directly to the trained model, and the time is measured from the start of the forward pass until all outputs are generated. This process is repeated across 100 individual samples, and the mean inference time is reported.

The results are summarized in Table 5.1, which also includes inference time to indicate how efficiently each model processes a mel spectrogram to output. For reference, two benchmark models from the study by Hou et al. (2024) are included: the original AD_CNN, which has the lowest number of parameters, and SoundAQnet, which achieved the highest overall performance. As shown in the Table, several of the developed models approach the performance of the original AD_CNN baseline while substantially reducing model size. AD_CNN_dense_layer and AD_CNN_hop_length yield the best results among the variants, with MSE means of 1.155 and 1.159 respectively, only slightly higher than the baseline value of 1.128. In terms of classification performance, all variants perform comparably to AD_CNN, with AUC scores ranging from 0.82 to 0.85, indicating no substantial drop in event classification capability.

In terms of inference time, AD_CNN_hop_length is the most efficient among the developed models, requiring only 7.79 ms per sample. This improvement is primarily due to the reduced temporal resolution of its input spectrogram (1501×64), compared to the standard 3001×64 used by all other models. Despite this reduction in input size, its performance remains strong, with an AUC of 0.85 and an MSE mean nearly identical to that of AD_CNN_dense_layer.

Interestingly, AD_CNN_dense_layer shows the slowest inference among the variants (13.77 ms), despite containing roughly the same number of parameters as AD_CNN_hop_length. This suggests that parameter count alone does not determine inference speed. Notably, the AD_CNN_harder_max_pooling model, with the lowest parameter count (98,567), maintains comparable classification performance (AUC = 0.84), but at the cost of slightly higher MSE (1.228). Its inference time, however, remains on par with the original AD_CNN. The complete set of MSE calculations for the perceptual attributes is provided in Appendix C, Table C.1 & C.2.

Overall, the modified models represent effective trade-offs between predictive performance and computational efficiency. While SoundAQnet delivers the strongest results (AUC = 0.94, MSE = 1.052), the performance differences are relatively small considering its significantly larger parameter count. This highlights the potential of the adapted AD_CNN variants as practical and efficient alternatives.

Table 5.1: Comparison on developed models + benchmark models on the test set.

#	Model	Parameters	Event Classification	Perceptual Attributes	Inference Time
			AUC	*MSE Mean*	*ms*
1	AD_CNN_decreased_filters	252,287	0.84	1.488	8.46
2	AD_CNN_dense_layer	279,767	0.82	1.155	13.77
3	AD_CNN_hop_length	280,967	0.85	1.159	7.79
4	AD_CNN_harder_max_pooling	98,567	0.84	1.228	13.20
5	AD_CNN	520,967	0.84	1.128	13.79
6	SoundAQnet	2,701,812	0.94	1.052	114.32

5.2. Training process recap

To evaluate and compare the performance of lightweight neural networks in predicting perceptual soundscape attributes, a consistent training pipeline was implemented. The input data consisted of log-mel spectrograms, processed from 30-second raw audio clips and saved as `.npy` files. These spectrograms were normalized using dataset-wide statistics to ensure uniform scaling across training batches. All models shared identical preprocessing procedures and were trained with a batch size of 32 and a learning rate of 1×10^{-4}. Training was carried out for up to 100 epochs, with early stopping employed to prevent overfitting. The early stopping criterion monitored the validation loss for the perceptual attribute *pleasant*, with a patience of 10 epochs, and preserved the weights of the best-performing model. Each model was evaluated based on prediction performance, model complexity, and inference efficiency.

5.3. Results per model

This section presents the outcomes of the four lightweight neural network models evaluated in this thesis. Each model is a variation of the baseline AD_CNN architecture and aims to reduce parameter count while retaining predictive accuracy on perceptual soundscape attributes. Key differences in design are outlined per model, followed by an evaluation of their prediction performance and inference characteristics.

5.3.1. AD_CNN_decreased_filters

The model AD_CNN_decreased_filters reduces the number of convolutional filters by half in each layer compared to the baseline (e.g., 16→8, 32→16). This modification significantly decreases the dimensionality of the feature maps before reaching the dense layers. The input resolution remains unchanged at 3001 time frames and 64 mel bins, and the dense layer configuration remains unchanged, with a fully connected layer of 100 units. As a result, the total number of trainable parameters is reduced to 252,287.

Figure 5.1 shows that the model trained for 41 epochs, with a steady decrease in training loss and stable validation metrics. The Event AUC on the validation set initially improves and then stabilizes, while the regression MSEs for *pleasant* and the mean of the 8D PAQ attributes show modest variability. This indicates that the model converges, but may not have sufficient capacity to fully learn the complex patterns associated with perceptual attributes. This behaviour, combined with the reduced model capacity, could be a sign of underfitting, where the model lacks representational power.

5.3.2. AD_CNN_dense_layer

The model AD_CNN_dense_layer retains the original convolutional architecture of the baseline AD_CNN model but reduces the size of the dense layer from 100 to 50 units. This change significantly reduces the number of trainable parameters in the final part of the network while preserving the representational capacity of the convolutional feature extractor. The input resolution is maintained at 3001 time frames and 64 mel bins, and the rest of the model configuration, including the number of filters and

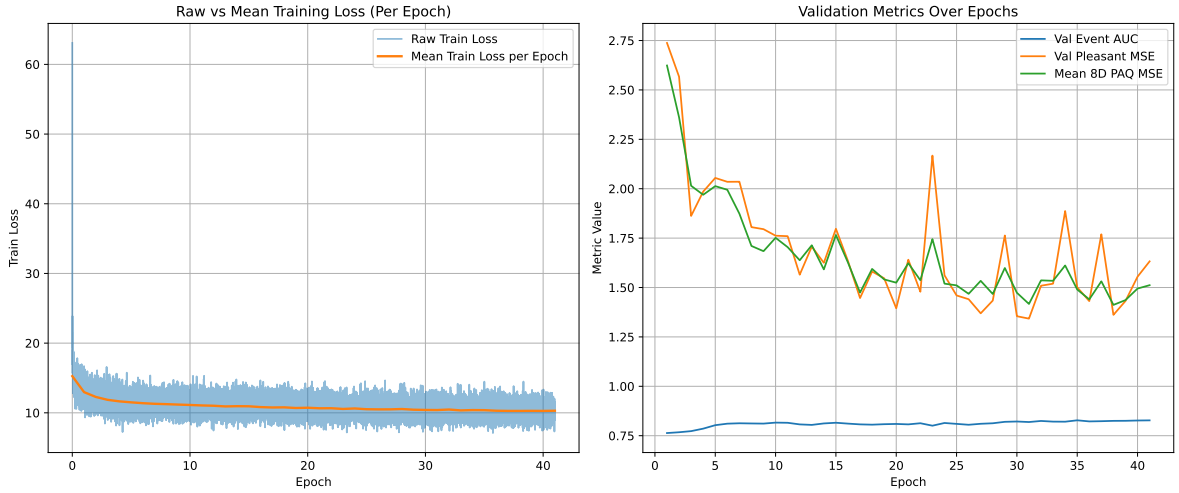


Figure 5.1: Training and Validation performances of AD_CNN_decreased_filters model.

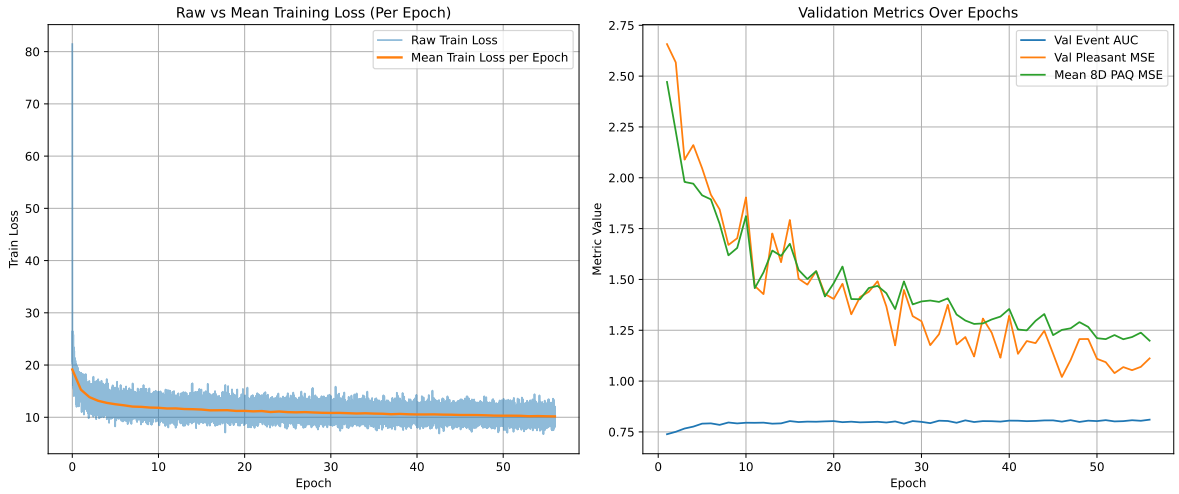


Figure 5.2: Training and Validation performances of AD_CNN_dense_layer model.

pooling strategy, remains unchanged. As a result, the total number of trainable parameters is reduced to 279,767.

Figure 5.2 shows that the model trained for 56 epochs, with a steady decrease in training loss and relatively stable validation performance. The Event AUC on the validation set shows a consistent improvement before reaching a plateau, suggesting reasonable discriminative capacity for audio events. The MSEs for *pleasant* and the mean of the 8D PAQ attributes continue to decrease, although only marginally beyond epoch 46, where the validation set achieved its lowest *pleasant* MSE. This suggests that the model's validation performance is still improving, albeit at a slower rate, while the training loss shows a steady decline. Therefore, there is no indication of overfitting, and the model appears to generalise well.

5.3.3. AD_CNN_hop_length

The model AD_CNN_hop_length increases the hop length used during mel spectrogram extraction, resulting in a lower temporal resolution and a shorter sequence length (fewer time frames) compared to the original configuration. This adjustment reduces the computational load by decreasing the number of time steps while preserving the overall time–frequency structure of the input. The rest of the model architecture remains unchanged.

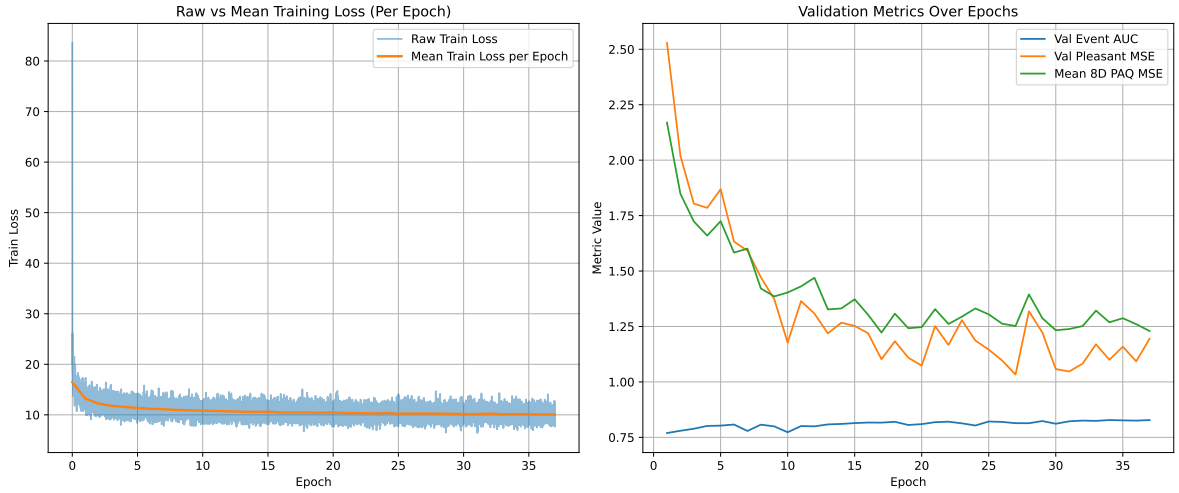


Figure 5.3: Training and Validation performances of AD_CNN_hop_length model.

Figure 5.3 shows that the model trained for 37 epochs, with a consistent decrease in training loss that begins to plateau around epoch 25. The validation metrics, Event AUC, *pleasant* MSE, and the mean 8D PAQ MSE, also improve steadily but show signs of early stabilisation. The *pleasant* MSE reaches its lowest value at epoch 27, after which performance levels off. This indicates that the model is capable of learning the relevant perceptual patterns in the data. However, the early plateau in both training and validation metrics may suggest limited model capacity, pointing to mild underfitting rather than overfitting.

5.3.4. AD_CNN_harder_max_pooling

The model AD_CNN_harder_max_pooling modifies the baseline architecture by applying more aggressive max pooling operations after the convolutional layers. This adjustment increases the degree of spatial downsampling, thereby reducing the dimensionality of intermediate feature maps and lowering the computational load. The overall structure of the model, including the convolutional blocks and dense layers, remains unchanged, resulting in a parameter count of 98,567.

Figure 5.4 shows that the model trained for 44 epochs, with a steady decline in training loss throughout, indicating effective convergence. The validation metrics reveal that audio event classification achieves high AUC values, while the MSEs for *pleasant* and the averaged 8D PAQ perceptual attributes remain relatively stable over epochs. This suggests that, despite the reduced model size, the network maintains adequate generalisation capability. However, the early flattening of the validation metrics could point to a mild underfitting effect due to overly compressed feature representations.

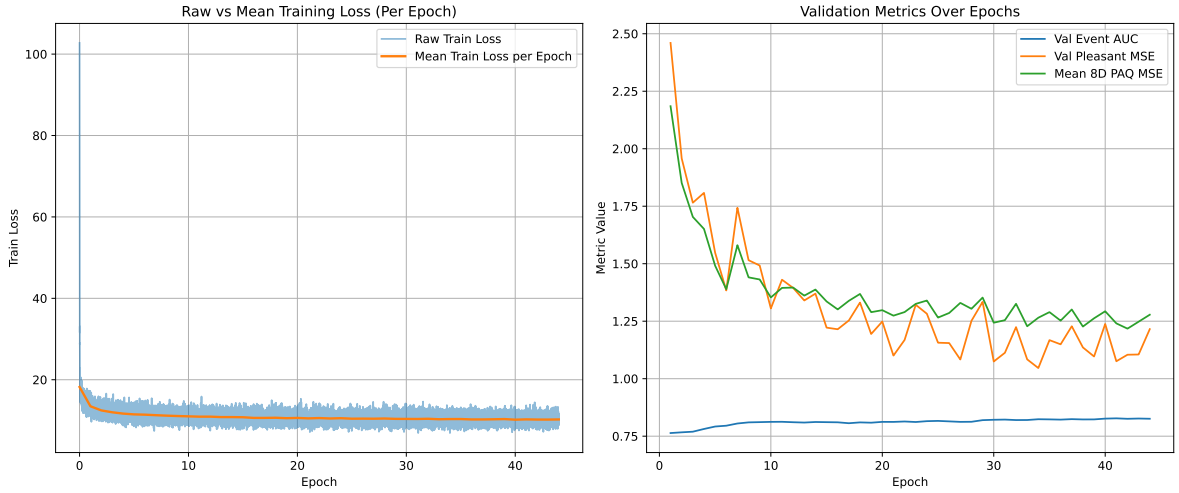


Figure 5.4: Training and Validation performances of AD_CNN_harder_max_pooling model.

5.4. The combined model

The results in Table 5.1 indicate that two models, namely AD_CNN_hop_length and AD_CNN_dense_layer, demonstrate the most promising performance. While AD_CNN_hop_length offers the fastest inference time (7.79 ms) due to its reduced temporal resolution input, AD_CNN_dense_layer achieves the lowest MSE among the modified models, with only a slight increase compared to the original AD_CNN baseline.

AD_CNN_dense_hop_combined

To explore whether further efficiency gains could be achieved without significantly sacrificing accuracy, a new hybrid model, called *AD_CNN_dense_hop_combined* was designed. This model combines the coarser temporal resolution of AD_CNN_hop_length with the smaller dense layer of AD_CNN_dense_layer, effectively integrating the main simplifications from both. The aim was to reduce the overall parameter count and improve inference speed while maintaining comparable prediction quality.

Figure 5.5 shows the training and validation metrics across epochs. The training stopped after 62 epochs, with the best validation MSE for *pleasant* achieved at epoch 52. The training loss steadily decreased over time, indicating consistent learning without signs of overfitting. The validation AUC remained stable throughout the training process, suggesting reliable event classification performance. Additionally, the mean MSE for the 8D PAQ attributes gradually declined and began to stabilize around epoch 50.

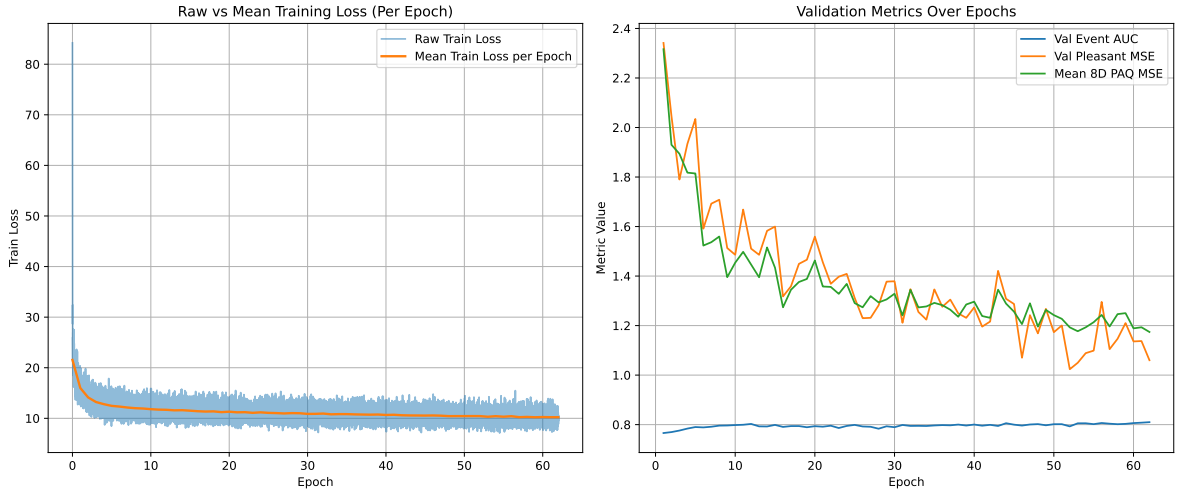


Figure 5.5: Training and Validation performances of AD_CNN_dense_hop_combined model.

To assess the effectiveness of the newly developed hybrid architecture, *AD_CNN_dense_hop_combined* (#5), the model was evaluated on the same held-out ARAUS test set used for previous comparisons. The results are summarized in Table 5.2. For reference, the partial models (#1 and #2) are also included, alongside the benchmark models (#3 and #4).

The hybrid model achieves an AUC of 0.83 for event classification, which is comparable to the other AD_CNN variants. For perceptual attribute regression, the model achieves a mean MSE of 1.114, the lowest among all parameter-reduced variants and even slightly outperforming the original AD_CNN (#3) baseline with approximately 1.2 percentage points. This indicates that the integration of reduced temporal resolution and a smaller dense layer does not compromise prediction accuracy and may even improve generalisation.

Table 5.2: Performance of the AD_CNN_dense_hop_combined model on the test set.

#	Model	Parameters	Event Classification	Perceptual Attributes	Inference Time
			AUC	*MSE Mean*	*ms*
1	AD_CNN_dense_layer	279,767	0.82	1.155	13.77
2	AD_CNN_hop_length	280,967	0.85	1.159	7.79
3	AD_CNN (baseline)	520,967	0.84	1.128	13.79
4	SoundAQnet	2,701,182	0.94	1.052	114.32
5	AD_CNN_dense_hop_combined	159,767	0.83	1.114	7.23

5.5. Generalisation performance of the models on ISD

To further assess the generalisability of the developed models, a generalisation study was conducted using the International Soundscape Database (ISD). While Chapter 3 detailed the ISD dataset's structure and the rationale for its selection, this section presents the results of evaluation.

The goal of this study is to examine how well models trained exclusively on the ARAUS dataset perform when applied to audio files from a different source. The ISD dataset offers a distinct set of urban parks, accompanied by subjective evaluations, enabling an external validation of model robustness.

Both SoundAQnet, the best performing model developed by Hou et al. (2024), and the ensemble model AD_CNN_dense_hop_combined, introduced in the previous section, are included in the evaluation. For completeness, the original baseline AD_CNN has also been included. Model predictions are compared against the ground-truth perceptual ratings, obtained by averaging human annotations from the ISD dataset. Performance is analysed via regression metrics, and by using perceptual mapping in the ISO 12913-2 circumplex model. These analyses provide insights into whether the models can accurately capture perceptual soundscape attributes beyond the conditions they were trained on.

5.5.1. Generalisation results on the ISD

Running the inference script for all three models produces MSE scores for each perceptual attribute individually, as well as the average MSE across all attributes. These results are presented in Table 5.3.

Table 5.3: MSE of Perceptual Attribute Predictions on the ISD.

Model	MSE								
	Pleas.	Eventf.	Chaot.	Vibra.	Uneventf.	Calm	Annoy.	Monot.	**Mean**
AD_CNN_ dense_hop_ combined	3.859	1.056	1.534	1.701	1.033	3.244	2.327	1.144	**1.987**
AD_CNN (baseline)	3.809	1.093	1.384	1.763	1.104	3.260	1.814	1.128	**1.919**
SoundAQnet	2.006	1.005	1.392	1.447	1.169	1.680	1.980	1.686	**1.545**

As showed above, the AD_CNN_dense_hop_combined model achieves a mean MSE of 1.987 across perceptual attributes, demonstrating solid performance despite its lightweight architecture. Its score is comparable to that of its baseline, the original AD_CNN, which attains a mean MSE of 1.919. Specifically, although the lightweight variant has nearly three times fewer parameters, it performs only slightly worse, by approximately 3.5 percentage points. Both models show higher prediction errors on the attributes *pleasant*, *calm*, and *annoying*. In contrast, SoundAQnet achieves the lowest mean MSE (1.545), consistently outperforming both AD_CNN and AD_CNN_dense_hop_combined across nearly all perceptual attributes. This aligns with its performance on the ARAUS test set and highlights its strong generalisation capability. Notably, it predicts *pleasant* with considerably lower error than the other models.

5.5.2. Perceptual mapping and model interpretability

Since the original AD_CNN performed comparably to the AD_CNN_dense_hop_combined, it is excluded from further evaluation in this section. The remainder of the analysis will therefore focus on the SoundAQnet and the AD_CNN_dense_hop_combined.

After running the inference scripts, a results notebook was developed to process the outputs. It computes the ISO *Pleasantness* and ISO *Eventfulness* scores for each audio sample. These scores, along with the ground-truth values, are then plotted on a two-dimensional circumplex model, providing a visual interpretation of the perceptual soundscape dimensions.

The ISD dataset comprises recordings from 19 urban soundscapes, specifically 19 parks located across four cities: Granada, Groningen, London, and Venice. However, for clarity and conciseness, this section focuses on a subset of four representative parks. The corresponding results are visualised in Figure 5.6. Visualisations for the remaining cities are included in Appendix C for reference.

In *Russell Square (London)*, ground-truth responses cluster moderately in the pleasant and slightly eventful quadrant, indicative of a generally positive and moderately stimulating soundscape. The AD_CNN_dense_hop_combined model substantially underestimates pleasantness, with many of its predictions falling into the unpleasant quadrant. This reflects limited accuracy in capturing perceptual attributes contributing to ISO *Pleasantness*. It is also noteworthy that the blue dots (AD_CNN_ dense_hop_combined predictions) are closely grouped together, indicating low variability. SoundAQnet performs better than the AD_CNN_dense_hop_combined model, though it still underestimates both pleasantness and eventfulness.

The *Noorderplantsoen (Groningen)* plot shows ground-truth ratings highly concentrated in the pleasant and moderately eventful quadrant, denoting a positively perceived and moderately lively soundscape. In this case, both AD_CNN_dense_hop_combined and SoundAQnet predict approximately the same region in the circumplex. They perform reasonably well on the *Pleasantness* dimension but tend to underestimate the *Eventfulness* of the park.

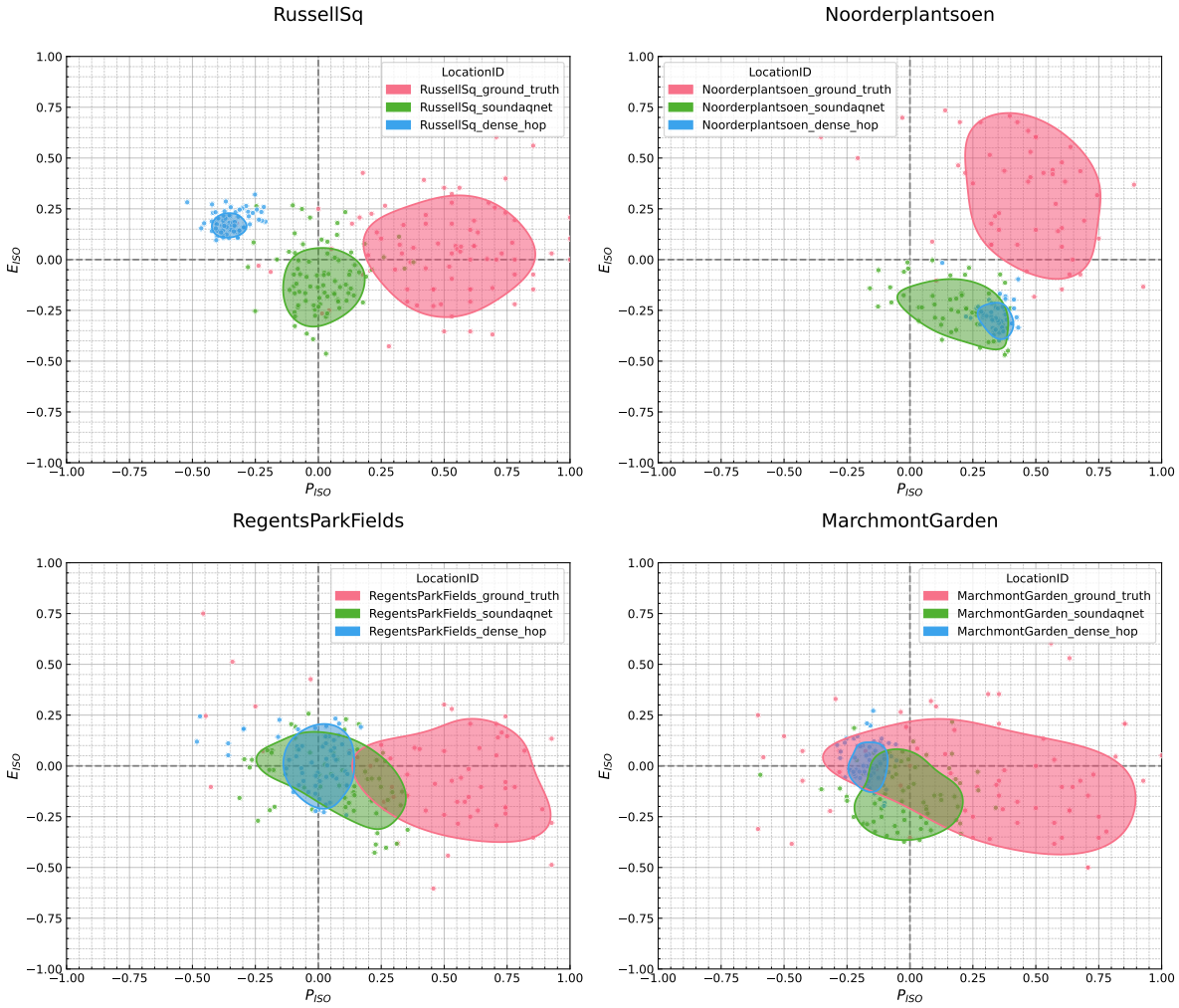


Figure 5.6: Two-dimensional circumplex models of perceptual attributes for four urban soundscapes, derived from the ISD dataset.

In *Regents Park Fields (London)*, ground-truth ratings clearly indicate high pleasantness and moderate eventfulness, reflective of a positively stimulating acoustic setting. The AD_CNN_dense_hop_combined model predicts values near the centre of the circumplex, but with a noticeably small spread of predictions. SoundAQnet closely matches the perceptual characteristics of this soundscape, slightly underestimating *Pleasantness* while accurately estimating *Eventfulness*.

Lastly, *Marchmont Garden (London)* exhibits moderate pleasantness and relatively low eventfulness in the ground-truth data, suggesting a calm and agreeable acoustic atmosphere. The relatively large spread of ground-truth responses implies varied perceptions among participants. The AD_CNN_dense_hop_combined model places its predictions near the neutral point, with a tendency toward higher *Eventfulness*, thus failing to accurately capture the pleasant nature of the soundscape. Once again, its predictions are tightly clustered, indicating low variance. SoundAQnet predictions are broadly consistent with the ground-truth but show slight underestimation of *Pleasantness*.

Overall, the circumplex analyses highlight the limitations of the AD_CNN_dense_hop_combined model in capturing the perceptual diversity of urban soundscapes. While it occasionally produces accurate predictions, its outputs tend to cluster within a narrow region of the circumplex, failing to reflect the broader range of perceptual responses observed in the ground-truth data. These ground-truth responses often span multiple quadrants, suggesting that individual soundscapes evoke highly variable and subjective experiences. This perceptual complexity poses a significant challenge for lightweight models like AD_CNN_dense_hop_combined, which may struggle to generalise due to their constrained number of parameters. In contrast, SoundAQnet demonstrates more robust generalisation, better capturing the diversity present across urban environments, albeit still with notable limitations.

Additionally, the AD_CNN_dense_hop_combined model appears to underperform particularly on the ISO *Pleasantness* dimension. This can be explained by the fact that the perceptual attributes *pleasant* and *annoying* are the primary contributors to the ISO *Pleasantness* score, as defined by the formulas introduced in Section 1.2.2. Notably, these are also the attributes for which the model showed the highest prediction errors, as reported in Table 5.3 in the previous section. This alignment further confirms that the model struggles to accurately capture ISO *Pleasantness*, highlighting a key limitation in its generalisation capability for affective soundscape dimensions.

6

Discussion

This chapter critically evaluates the implications and significance of the research findings presented in Chapter 5. It begins by interpreting the performance of the developed models, offering explanations for observed outcomes, discussing potential reasons for their generalisation behaviour, and considering their practical applicability in real-world urban settings. Subsequently, the limitations of the datasets, models, and experimental procedures are discussed, highlighting areas that could benefit from further investigation. Finally, the chapter addresses policy implications arising from the successful deployment of lightweight predictive models in urban soundscapes, along with associated ethical considerations.

6.1. Interpretation of results

This study evaluated four lightweight neural network architectures for the task of predicting perceptual soundscape attributes from urban audio recordings. The objective was to identify models suitable for real-time inference on low-cost, resource-constrained sensors, so limited to a maximum of 300,000 trainable parameters, while maintaining acceptable predictive accuracy and acceptable inference speed.

Among the developed models, AD_CNN_dense_layer and AD_CNN_hop_length demonstrated the strongest performance on the test set, achieving a mean MSE of 1.155 and 1.159, respectively. In the AD_CNN_dense_layer model, the fully connected layer was reduced from 100 to 50 units, resulting in fewer input and output heads. In the AD_CNN_hop_length model, the temporal resolution of the input was reduced during preprocessing: a larger hop length yielded spectrograms with less detailed information.

These two strategies were subsequently combined into a single ensemble model, named AD_CNN_dense_hop_combined, to assess their complementary effect. Somewhat unexpectedly, this combined model achieved a further reduction in mean MSE to 1.114 on the test set, outperforming the baseline AD_CNN benchmark proposed by Hou et al. (2024), based on the original architecture by Ooi et al. (2024), which reached a mean MSE of 1.128.

This outcome was initially surprising, as both parameter reduction and temporal coarsening are generally expected to reduce model capacity and potentially limit predictive accuracy. However, in line with the principle of Occam's Razor, simpler models are often preferable when they achieve comparable or better results. Reducing the size of fully connected layers in CNN architectures has previously been shown to be an effective simplification strategy that preserves accuracy (Sharma, 2022).

Moreover, while reducing temporal resolution limits the granularity of frequency content over time, it can be beneficial for tasks like sound event detection, where identifying the presence of events is more important than their exact timing (Leiber et al., 2023). This observation supports the finding that coarse temporal features, combined with smaller dense layers, can still capture sufficient perceptual information to make accurate predictions.

In the generalisation study, both the parameter-reduced ensemble model and the SoundAQnet model,

identified as the best-performing benchmark, were evaluated in inference mode on an unseen dataset: the International Soundscape Database (ISD). While both models demonstrated satisfactory performance on individual attributes such as *pleasant*, *annoying*, and *calm* when tested on the ARAUS dataset, the same dataset they were trained on, their performance on these same attributes declined noticeably when evaluated on the ISD. This highlights a drop in generalisability when the models are applied to data from a different distribution. A common challenge in deep learning, when models trained on one dataset struggle to predict the same target but then from a different dataset (X. Zhang et al., 2024).

One possible explanation could be the differences between the datasets themselves. For instance, the mean score for *pleasant* is 3.78 in the ISD dataset, compared to just 3.00 in ARAUS. However, if this argument is made, it would also need to be applied to other attributes such as *monotonous*, which shows a comparable difference of 0.71 between the datasets. Furthermore, the distribution of the scores does not serve as a sufficient justification either. While attributes like *pleasant* and *annoying* indeed show skewed distributions, this is also true for *monotonous* and *vibrant*. Similarly, the distribution of *calm* closely resembles that of *uneventful*, albeit in the opposite direction.

Valid reasons for the performance differences can be traced back to differences in the audio files themselves, particularly in how they were created. The ARAUS dataset consists of laboratory-generated audio recordings, whereas the ISD contains real-world in-situ recordings. In-situ soundscapes are typically more acoustically complex and less controlled with respect to which sound sources are present or dominant. In such environments, certain sounds may be overwhelmed by others, making them harder to perceive. By contrast, in ARAUS, maskers were deliberately added at varying loudness levels during the augmentation process, which could make specific sounds more prominent and easier to detect, both for human listeners and predictive models. For example, a bird sound is an important factor for perceived pleasantness; if the model fails to recognize it due to background masking in real-world recordings, it might underestimate the pleasantness of the soundscape.

6.2. Limitations

6.2.1. Limitations of the ARAUS dataset

The ARAUS dataset carries several known limitations, as identified by Ooi et al. (2024), including a demographically skewed participant pool, a limited set of masker types, and the artificial nature of the audio-visual stimuli. These limitations also apply to this study, as no corrective measures or alternative datasets were used.

6.2.2. Limitations of the models

The training was conducted using a fixed learning rate, batch size, and optimizer, in accordance with the training pipeline proposed by Hou et al. (2024). As a result, no systematic hyper-parameter optimization was performed, which may have limited the models from reaching their optimal performance. Although alternative batch sizes and learning rates were briefly explored, future work could incorporate automated hyper-parameter tuning methods, such as grid search or Bayesian optimization, to more effectively explore the parameter space and improve model performance.

This study did not explore modifications to components such as Batch Normalization, Dropout, or ReLU activations, which could influence training stability, model regularization, and generalisation performance. These elements were left unchanged to maintain a controlled comparison between different parameter-reduction strategies. However, their omission may have limited the models' ability to generalise. To mitigate this, future work could investigate lightweight forms of normalization and regularization that are compatible with low-cost, resource-constrained sensors. Incorporating techniques regarding Batch Normalization or different dropout rates could improve robustness without significantly increasing the model size, as those techniques add very few parameters to the model (Ioffe & Szegedy, 2015).

Furthermore, training could have been repeated multiple times with different random seeds to strengthen the robustness of the results. This would help account for variability due to weight initialization and randomness during training (Menart, 2020). However, each run required at least eight hours, depending on when early stopping was triggered, making it impractical to train each model, for example, ten times

to compute reliable average performance metrics. This limitation could be mitigated in future work by utilizing more powerful computational resources to enable extensive experimentation within a reasonable time-frame.

In the original AD_CNN and SoundAQnet training procedure, ISO *Pleasantness* was used as the validation monitor for early stopping. Since this thesis excludes the ISO metrics as prediction targets, the individual perceptual attribute *pleasant* was used instead to guide early stopping when training the parameter-reduced models. However, ISO *Pleasantness* is a composite construct derived from multiple attributes, while *pleasant* represents only a single component of this broader metric. As a result, using *pleasant* alone as the stopping criterion may not provide a sufficiently comprehensive reflection of overall model performance. Although this choice likely had a limited effect on training outcomes, a more balanced approach would have been to monitor the average validation error across all eight perceptual attributes. Such an aggregated metric would offer a more holistic view of model generalisation and help avoid overfitting to any one attribute, particularly since no single perceptual attribute can be considered inherently more important than the others.

Another limitation of this thesis is the relatively arbitrary approach taken to reduce the parameters of the original AD_CNN model. Specifically, parameter reduction was pursued by increasing the hop length, decreasing the number of convolutional layers, halving the size of the dense layer, and adjusting the max pooling configurations. These modifications were chosen somewhat straightforwardly, performed incrementally by a single step, and not based on any systematic exploration or optimization method. Consequently, the final model architectures might not represent the most efficient or optimal solutions for minimizing the parameter count while preserving predictive accuracy. For example, it remains unclear how predictive performance would be affected by further increases in hop length, or by additional reductions in the size of convolutional and dense layers. Therefore, a more systematic and methodologically rigorous exploration of parameter reduction strategies could potentially yield more optimized lightweight models, possibly even achieving improved performance.

6.3. Limitations of generalisation study

A key limitation of both the SoundAQnet model and the lightweight models developed in this study is their reliance on fixed-length input. In contrast, the audio recordings from the ISD dataset vary in duration, necessitating preprocessing steps to standardize their lengths. To achieve uniform input dimensions suitable for batch processing and model compatibility, shorter clips were extended using mirror padding, while longer clips were truncated. Although this approach enables consistent model input, it may introduce unintended consequences. Truncation risks omitting important acoustic information, particularly if salient events occur later in a recording. Mirror padding, while preferable to zero-padding, can artificially repeat or reinforce existing patterns, potentially biasing the model toward features that do not reflect the true soundscape. Additionally, this fixed-length preprocessing restricts the model's capacity to handle the natural variability of urban audio environments. Future research could explore architectures that accept variable-length inputs or apply masking techniques to reduce the influence of padded segments during training.

Another notable limitation is the omission of the loudness feature during inference with SoundAQnet. Although the original model incorporates loudness as an additional input alongside Mel spectrograms, it was excluded in this study because preprocessing relied on a specific calibration file. Using an incorrect or mismatched calibration file resulted in inaccurate loudness values, which negatively affected model predictions. As a result, loudness was deliberately omitted to preserve inference reliability. Nonetheless, loudness can still provide complementary information, and its inclusion may enhance predictive performance. In the case of SoundAQnet, previous results have shown that including loudness improved predicting performance (Hou et al., 2024). It is therefore plausible that the parameter-reduced models developed in this thesis could also benefit from its inclusion. However, leveraging this feature effectively requires a correct calibration file; otherwise, the model risks being misled by wrong input data. Future work could revisit the use of loudness, provided that a reliable and correct calibration file is used.

6.4. Policy implementations and implications

The integration of lightweight models for perceptual attribute prediction into urban policy frameworks represents a promising advance towards more human-centric and sustainable urban design. By translating complex acoustic environments into interpretable perceptual dimensions, such as *Pleasantness* and *Eventfulness*, these models provide policymakers with actionable insights that go beyond traditional noise metrics. However, the implementation of such models also brings several practical and ethical considerations, which are discussed below.

6.4.1. Implementing lightweight perceptual attributes predicting models in practice

As demonstrated in this chapter, it is technically feasible to develop a perceptual attribute prediction model that is both compact in size and capable of delivering strong predictive performance. This opens the door to deployment on low-cost, resource-constrained sensors, as introduced by Cassens et al. (2024), enabling scalable, city-wide deployment.

All four designed lightweight models, each based on the AD_CNN baseline architecture and constrained to fewer than 300,000 parameters, were capable of performing multi-label classification of environmental sound sources as well as regression-based prediction of the eight perceptual attributes. Among these, two models, the variant with a reduced dense layer and the one using a spectrogram input with a larger hop length, achieved performance comparable to the baseline. These two were subsequently combined into an ensemble model, which yielded the best overall results. This hybrid model contains only 159,767 parameters and predicts the eight perceptual attributes - pleasant, eventful, chaotic, vibrant, uneventful, calm, annoying, and monotonous - on a 5-point Likert scale, in accordance with the ISO/TS 12913-3 standard for perceptual soundscape assessment (ISO, 2019).

Of particular relevance to urban policymakers is the derivation of the ISO-based indices *Pleasantness* and *Eventfulness*, which can be visualised in real time using circumplex models on live dashboards. As described in Section 5.5, the transformation from attribute scores to ISO metrics allows for intuitive mapping of soundscapes in a two-dimensional perceptual space, with *Pleasantness* on the x-axis and *Eventfulness* on the y-axis, following the model developed by Axelsson et al. (2010). Practically, sensors could transmit the predicted perceptual attribute values to a computer, where the corresponding ISO indicators are calculated and visualised.

These real-time visualisations provide a powerful tool for managing urban soundscapes. Policymakers can use the circumplex dashboard to monitor how different areas of a city are acoustically perceived over time. Even a single urban location can be analysed in detail. For instance, Zaffaroni-Caorsi et al. (2025) investigated a square on their university campus and found that the paved section was perceived as less pleasant and less eventful compared to the grassy side. Such findings can inform urban planning decisions, potentially leading to policy changes, such as increasing greenery in specific areas, to enhance the acoustic experience of public spaces.

6.4.2. Privacy considerations in real-time monitoring

Privacy does not present a significant issue in real-time monitoring, as long as key safeguards are properly implemented. The current model uses 30-second audio clips for prediction, which could theoretically capture private conversations. However, this concern is effectively neutralized when all audio is processed locally on the sensor itself. With no raw audio being stored or transmitted externally, privacy is preserved by design. Furthermore, edge-based processing is significantly more energy-efficient than transmitting raw audio to the cloud, making it better suited for battery-powered or solar-powered deployments (Karges et al., 2022). Additionally, reducing the input length to 5 or 10 seconds can further minimize exposure to sensitive content, though it may impact the accuracy of perceptual attribute predictions that rely on longer contextual cues. Most importantly, the system must comply with the General Data Protection Regulation (GDPR), as outlined by European Union (2025).

6.4.3. Model bias and interpretability

While predictive models provide valuable insights, they are inherently shaped by the data on which they are trained. The SoundAQnet and AD_CNN_dense_hop_combined models, for example, are trained

on a large-scale dataset in which many individuals reviewed augmented soundscapes. This process tends to moderate and generalise perceptual responses, reducing the influence of personal, subjective emotions. As such, it is important to recognize that perceptions are always subjective, and contextual factors play a critical role. According to Tarlao et al. (2021), the three main facets are: personal factors (demographics) , situational factors (social activity), and environmental factors (visual environment). Therefore, policymakers and practitioners should view the model's predictions as probabilistic estimates rather than objective truths. One way to increase validity is by collecting localized perceptual ratings to retrain or fine-tune the model, thereby aligning predictions with community expectations and lived experience. However, this approach may also introduce new biases, as participants in perceptual experiments may subconsciously adjust their ratings toward what they believe the experimenter expects (Lindborg & Friberg, 2016).

6.4.4. Fine-tuning for specific usage

In the generalisation study, both developed models demonstrated a conservative tendency when predicting the *pleasant* and *eventful* attributes. This behaviour often resulted in poor estimation of the ISO metric *Pleasantness*, with predictions frequently defaulting to neutral values. This underlines the importance of local recalibration, especially when deploying models in real-world scenarios. Notably, model performance on the ARAUS dataset suggests higher potential than was realised in the generalisation study, highlighting that generalisation remains a key challenge. Local fine-tuning or recalibration not only enhances predictive accuracy but may also increase acceptance among stakeholders and the public, as it demonstrates that the model has been specifically adapted to their local context (Wu et al., 2024).

7

Conclusion

This chapter provides a comprehensive synthesis of the research conducted in this thesis. It begins by addressing each of the sub-questions that guided the investigation, followed by an answer to the overarching research question. The chapter then outlines the practical and scientific contributions of the work, highlighting the relevance of the developed models for real-world applications. Finally, it reflects on potential directions for future research, offering suggestions for continued development and validation of lightweight models for perceptual soundscape prediction.

7.1. Answers to research questions

7.1.1. Sub-question 1

Which existing deep learning models are suitable for predicting multiple perceptual soundscape attributes simultaneously, and how can they be adapted for low-cost, resource-constrained sensors?

This thesis identifies AD_CNN as the most suitable existing deep learning architecture for predicting multiple perceptual soundscape attributes at low-cost, resource-constrained sensors. The model contains only 520,967 parameters and uses mel spectrograms as its sole input. It achieves an AUC of 0.84 for audio event classification and a MSE of 1.128 for perceptual attribute prediction, representing strong performance for such a lightweight model. In comparison, more complex architectures such as SoundAQnet achieve higher overall accuracy, but are too computationally demanding for real-time deployment on sensors.

7.1.2. Sub-question 2

How does the performance (accuracy, computational efficiency, and storage requirements) of the designed lightweight models compare to state-of-the-art model SoundAQnet?

Initially, four models were developed, with two demonstrating particularly strong predictive performance for perceptual attributes. The first model, characterized by a smaller dense layer, achieved an MSE of 1.155, while the second model, utilizing a reduced temporal resolution spectrogram, scored an MSE of 1.159, both comparable to the original AD_CNN's performance (1.128). These two parameter-reduced models were subsequently integrated into an ensemble model, named AD_CNN_dense_hop_combined.

This design aimed to lower model complexity while retaining strong predictive capabilities. When compared to the state-of-the-art SoundAQnet, the lightweight model demonstrates a favourable balance between accuracy, inference time, and storage requirements. In terms of accuracy, it achieves a mean MSE of 1.114 and an event classification AUC of 0.83 on the ARAUS test set, closely approaching SoundAQnet's performance of 1.052 MSE and 0.94 AUC. Although SoundAQnet remains the most accurate model, the performance gap in regression is relatively small. In contrast, the difference in computational efficiency is substantial: AD_CNN_dense_hop_combined requires only 7.23 milliseconds per sample for inference, making it more than fifteen times faster than SoundAQnet, which averages

114.32 milliseconds. Furthermore, the lightweight model contains just 159,767 parameters, resulting in a properly reduced memory footprint. Overall, AD_CNN_dense_hop_combined achieves a compelling trade-off, offering strong prediction performance alongside major improvements in efficiency and deployability.

7.1.3. Sub-question 3

How well do predictive soundscape models generalise to different soundscapes, and what are the implications of their generalisation performance for urban policy making?

As demonstrated in the generalisation study, the lightweight ensemble model, AD_CNN_dense_hop_combined, struggled to generalise to the unseen International Soundscape Database (ISD) dataset. This model exhibited a narrow output range and consistently underestimated perceptual attributes, which led to a poor derivation of the ISO-based soundscape metrics. Interestingly, even the high-capacity benchmark model, SoundAQnet, showed limitations in this regard, suggesting that generalisation is a broader challenge in perceptual soundscape prediction. This reinforces the value of local recalibration: by fine-tuning on perceptual data collected in the target environment, model predictions can better align with the lived experiences and expectations of the local population. Doing so not only improves predictive performance but also enhances public acceptance and practical relevance.

From a policy perspective, the value of lightweight models lies in their suitability for deployment on low-cost, resource-constrained sensors, enabling large-scale, real-time monitoring of urban soundscapes. This thesis demonstrates that such models can efficiently predict perceptual attributes in a computationally efficient manner. These predictions can be used to derive ISO-based metrics *Pleasantness* and *Eventfulness*, offering policymakers intuitive, human-centric indicators for assessing and managing the soundscape quality of urban environments.

However, the limitations in generalisability underscore the risks of relying on pre-trained models without contextual adaptation. To ensure reliable and context-sensitive insights, local fine-tuning may be essential. This helps align model outputs with the specific cultural, environmental, and acoustic realities of each urban environment, thereby supporting more informed, responsive, and equitable urban policy making.

7.1.4. Main research question

After answering the three sub-questions, the main research question - *How can lightweight neural network models be designed to accurately predict perceptual soundscape attributes and sound sources from urban audio recordings in real-time on low-cost, resource-constrained sensors, and to what extent can such models generalise across diverse soundscapes to support urban policy making?* - can now be addressed. This research has demonstrated that lightweight neural network models are capable of accurately predicting perceptual soundscape attributes and sound sources. While their performance approaches that of more complex models, their ability to generalise across diverse soundscapes remains limited. This highlights the importance of local fine-tuning to ensure effective deployment in support of urban policy making.

7.2. Contributions

7.2.1. Practical contributions

Prior to this research, no existing model with fewer than 300K parameters was capable of jointly predicting sound sources and perceptual attributes. This parameter limit was motivated by the use case presented in Cassens et al. (2024), as described in Section 1.6. This thesis demonstrates that even highly lightweight models can effectively predict both perceptual attributes and sound sources without sacrificing predictive performance. It provides a foundation for future development toward deployment on low-cost, resource-constrained sensors, enabling human-centric monitoring of urban soundscapes.

7.2.2. Scientific contributions

This thesis introduces the AD_CNN_dense_hop_combined model. It combines two parameter-reduction strategies: a smaller dense layer and a larger hop length during spectrogram extraction, which reduces temporal resolution and computational cost. Despite these reductions, the model outperforms the larger AD_CNN baseline in regression accuracy (MSE = 1.114 vs. 1.128), demonstrating that strategic sim-

plification can enhance model performance.

A generalisation study using the ISD demonstrates that even advanced models like SoundAQnet face difficulties generalising beyond the controlled conditions of the training data. Models trained on augmented datasets such as ARAUS tend to perform worse when applied to real-world soundscape recordings, underscoring the need for context-aware fine-tuning. This reveals a broader challenge in adapting predictive models to diverse soundscapes.

7.3. Future work

Although this thesis deliberately excluded loudness features to maintain a lower parameter count, and due to the absence of a reliable calibration file that would have negatively impacted inference performance, loudness may nonetheless enhance model predictions, as demonstrated by Hou et al. (2024). It is also very plausible that loudness contributes meaningfully to perceptual attributes such as *eventful*, *annoying*, and *chaotic*. Therefore, incorporating loudness is a promising direction for future research, particularly if it leads to improved predictive accuracy. While its inclusion would increase the number of model parameters, this opens an additional area for exploration: identifying new strategies to offset the added complexity. Crucially, when implementing the loudness feature, as used by SoundAQnet, it is essential to have access to the correct calibration files to ensure reliable and meaningful input values.

In addition, future research should focus on systematically exploring other strategies to reduce model complexity. The strong performance of both the increased hop length and reduced dense layer models suggests that further changes to these components, such as increasing the hop size even more or trying different dense layer setups, may yield even more efficient architectures. Rather than applying single-step modifications, future work could use a grid search or automated search method to explore a wider range of design options. This would help find the best configurations that balance model size and predictive performance, and might even lead to better results than those presented in this thesis.

Lastly, to fully validate the practical viability of lightweight models for perceptual soundscape prediction, real-world deployment on sensors is essential. While this thesis focused on developing and benchmarking models in a controlled environment, actual sensor-based deployment would introduce new challenges such as hardware limitations, variable environmental conditions, and real-time processing constraints. Future work should focus on deploying one of the designed parameter-reduced models, for example AD_CNN_dense_hop_combined, on a real low-cost, resource-constrained sensor, as proposed by Cassens et al. (2024).

References

Aletta, F., & Kang, J. (2015). *Noise Mapping*, *2*(1). https://doi.org/doi:10.1515/noise-2015-0001

Aletta, F., Kang, J., & Axelsson, Ö. (2016). Soundscape descriptors and a conceptual framework for developing predictive soundscape models. *Landscape and Urban Planning*, *149*, 65–74. https://doi.org/10.1016/j.landurbplan.2016.02.001

Aletta, F., & Xiao, J. (2018). What are the Current Priorities and Challenges for (Urban) Soundscape Research? *Challenges*, *9*(1), 16. https://doi.org/10.3390/challe9010016

Alías, F., & Alsina-Pagès, R. M. (2019). Review of wireless acoustic sensor networks for environmental noise monitoring in smart cities. *Journal of Sensors*, *2019*(1), 7634860. https://doi.org/https://doi.org/10.1155/2019/7634860

Allen, J. (1977). Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *25*(3), 235–238. https://doi.org/10.1109/TASSP.1977.1162950

Alsina-Pagès, R. M., Hervás, M., Duboc, L., & Carbassa, J. (2020). Design of a low-cost configurable acoustic sensor for the rapid development of sound recognition applications. *Electronics (Switzerland)*, *9*(7), 1–20. https://doi.org/10.3390/electronics9071155

Axelsson, Ö. (2015). How to measure soundscape quality. *Proceedings of the Euronoise 2015 conference*, 1477–1481.

Axelsson, Ö., Nilsson, M. E., & Berglund, B. (2010). A principal components model of soundscape perception. *The Journal of the Acoustical Society of America*, *128*(5), 2836–2846. https://doi.org/10.1121/1.3493436

Bello, J. P., Silva, C., Nov, O., Dubois, R. L., Arora, A., Salamon, J., Mydlarz, C., & Doraiswamy, H. (2019). Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution. *Commun. ACM*, *62*(2), 68–77. https://doi.org/10.1145/3224204

Bild, E., Pfeffer, K., Coler, M., Rubin, O., & Bertolini, L. (2018). Public space users' soundscape evaluations in relation to their activities. an amsterdam-based study. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01593

Botteldooren, D., De Coensel, B., Renterghem, T., Dekoninck, L., & Gillis, D. (2008). The urban soundscape: A different perspective. *Duurzame Mobiliteit Vlaanderen de Leefbare Stad*, 177–204.

Brooks, B., Schulte-Fortkamp, B., Voigt, K., & Case, A. (2014). Exploring our sonic environment through soundscape research & theory. *Acoustics Today*, *10*, 30–40. https://doi.org/10.1121/1.4870174

Brown, A. L., & Muhar, A. (2004). An approach to the acoustic design of outdoor space. *Journal of Environmental Planning and Management*, *47*(6), 827–842. https://doi.org/10.1080/0964056042000284857

Brown, A. L., & Van Kamp, I. (2017). Who environmental noise guidelines for the european region: A systematic review of transport noise interventions and their impacts on health. *International Journal of Environmental Research and Public Health*, *14*(8). https://doi.org/10.3390/ijerph14080873

Cadena, L. F. H., Soares, A. C. L., Pavón, I., & Coelho, L. B. (2017). *Noise Mapping*, *4*(1), 57–66. https://doi.org/doi:10.1515/noise-2017-0004

Cassens, L., Kroesen, M., Calvert, S., & Van Cranenburgh, S. (2024). *Beyond loudness: development of a holistic solar-powered urban soundscape sensor* (tech. rep.). https://github.com/cityai-soundscapes

Chitra, B., Jain, M., & Chundeli, F. (2020). Understanding the soundscape environment of an urban park through landscape elements. *Environmental Technology & Innovation*, *19*, 100998. https://doi.org/10.1016/j.eti.2020.100998

Choi, K., Fazekas, G., & Sandler, M. (2016). Automatic tagging using deep convolutional neural networks. https://arxiv.org/abs/1606.00298

De Coensel, B., Sun, K., & Botteldooren, D. (2017). Urban soundscapes of the world : selection and reproduction of urban acoustic environments with soundscape in mind. *Proceedings of the 46th International Congress and Exposition on Noise Control Engineering*, 3647–3653.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

European Environment Agency. (2020). *Environmental noise in europe — 2020* (Report No. 22/2019). Publications Office of the European Union. https://www.eea.europa.eu/publications/environmental-noise-in-europe

European Union. (2025). Data protection under GDPR [Accessed: June 11, 2025]. *Your Europe*. https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_en.htm

Gallardo-Antolín, A., & Montero, J. M. (2021). On combining acoustic and modulation spectrograms in an attention lstm-based system for speech intelligibility level classification. *Neurocomputing*, *456*, 49–60. https://doi.org/https://doi.org/10.1016/j.neucom.2021.05.065

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., Ritter, M., & Google, I. (2017). *Audio Set: An Ontology And Human-Labeled Dataset For Audio Events* (tech. rep.). IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://static.googleusercontent.com/media/research.google.com/nl//pubs/archive/45857.pdf

Gong, Y., Chung, Y.-A., & Glass, J. (2021). AST: Audio Spectrogram Transformer. https://doi.org/https://doi.org/10.48550/arxiv.2104.01778

Hajnal, P., & Kocsis, D. (2022). Interpretation of road traffic noise changes with noise mapping. *Műszaki Tudományos Közlemények*, *16*, 24–30. https://api.semanticscholar.org/CorpusID:250625166

Harris, F. (1978). On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, *66*(1), 51–83. https://doi.org/10.1109/PROC.1978.10837

Herranz-Pascual, K., García, I., Diez, I., Santander, A., & Aspuru, I. (2017). Analysis of field data to describe the effect of context (acoustic and non-acoustic factors) on urban soundscapes. *Applied Sciences*, *7*(2). https://doi.org/10.3390/app7020173

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. (2017). Cnn architectures for large-scale audio classification. https://arxiv.org/abs/1609.09430

Hou, Y., Mitchell, A., Ren, Q., Aletta, F., Kang, J., & Botteldooren, D. (2023). Exploring annoyance in a soundscape context by joint prediction of sound source and annoyance. *10th Convention of the European Acoustics Association (EAA)*. https://doi.org/10.61782/fa.2023.0713

Hou, Y., Ren, Q., Mitchell, A., Wang, W., Kang, J., Belpaeme, T., & Botteldooren, D. (2024). Soundscape Captioning using Sound Affective Quality Network and Large Language Model. http://arxiv.org/abs/2406.05914

Hou, Y., Ren, Q., Zhang, H., Mitchell, A., Aletta, F., Kang, J., & Botteldooren, D. (2023). AI-based soundscape analysis: Jointly identifying sound sources and predicting annoyance. *The Journal of the Acoustical Society of America*, *154*(5), 3145–3157. https://doi.org/10.1121/10.0022408

Ioffe, S., & Szegedy, C. (2015, July). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning* (pp. 448–456, Vol. 37). PMLR. https://proceedings.mlr.press/v37/ioffe15.html

ISO. (2014, September). *Acoustics — Soundscape — Part 1: Definition and conceptual framework*. https://standards.iteh.ai/catalog/standards/sist/b7814739-18bd-445c-b381-ef0aadfbaf4c/iso-12913-1-2014

ISO. (2018). ISO/TS 12913-2:2018 acoustics “soundscape” part 2: Data collection and reporting requirements [Geneva: Standard, International Organization for Standardization].

ISO. (2019). *ISO/TS 12913-3:2019; Acoustics—Soundscape—Part 3: Data Analysis* (tech. rep.). ISO. Geneva, Switzerland.

Jeon, J. Y., & Hong, J. Y. (2015). Classification of urban park soundscapes through perceptions of the acoustical environments. *Landscape and Urban Planning*, *141*, 100–111. https://doi.org/10.1016/j.landurbplan.2015.05.005

Jeon, J. Y., Lee, P. J., You, J., & Kang, J. (2010). Perceptual assessment of quality of urban soundscapes with combined noise sources and water sounds. *The Journal of the Acoustical Society of America*, *127*, 1357–1366. https://doi.org/10.1121/1.3298437

Jo, H. I., & Jeon, J. Y. (2021). Compatibility of quantitative and qualitative data-collection protocols for urban soundscape evaluation. *Sustainable Cities and Society*, *74*, 103259. https://doi.org/10.1016/j.scs.2021.103259

Kang, J. (2021). Soundscape: Progress in the past 50 years and challenges in the next 50 years. *NOISE-CON proceedings*, *263*(6), 132–139. https://doi.org/10.3397/in-2021-1302

Kang, J. (2023). Soundscape in city and built environment: current developments and design potentials. *City and Built Environment*, *1*(1). https://doi.org/10.1007/s44213-022-00005-6

Kang, J., Aletta, F., Oberman, T., Erfanian, M., Kachlicka, M., Lionello, M., & Mitchell, A. (2019). Towards soundscape indices. *X*, 2488–. https://doi.org/10.18154/rwth-conv-239249

Kang, J., & Schulte-Fortkamp, B. (2016). *Soundscape and the built environment* (J. Kang & B. Schulte-Fortkamp, Eds.; 1st). CRC Press. https://doi.org/10.1201/b19145

Karges, N., Staab, J., Rauh, J., Wegmann, M., & Taubenböck, H. (2022). *Soundscapes on edge-The real-time machine learning approach for measuring Soundscapes on resource-constrained devices* (tech. rep.).

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. https://arxiv.org/abs/1412.6980

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2020). PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *28*, 2880–2894. https://doi.org/10.1109/TASLP.2020.3030497

LeBien, J., Zhong, M., Campos-Cerqueira, M., Velev, J. P., Dodhia, R., Ferres, J. L., & Aide, T. M. (2020). A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecological Informatics*, *59*, 101113. https://doi.org/10.1016/j.ecoinf.2020.101113

Leiber, M., Marnissi, Y., Barrau, A., & El Badaoui, M. (2023). *Differentiable short-time Fourier transform with respect to the hop length accepted for IEEE SSP workshop 2023* (tech. rep.).

Lindborg, P., & Friberg, A. (2016). Personality traits bias the perceived quality of sonic environments. *Applied Sciences*, *6*(12). https://doi.org/10.3390/app6120405

Lionello, M., Aletta, F., & Kang, J. (2020). A systematic review of prediction models for the experience of urban soundscapes. *Applied Acoustics*, *170*, 107479. https://doi.org/10.1016/j.apacoust.2020.107479

Liu, F., & Kang, J. (2016). A grounded theory approach to the subjective understanding of urban soundscape in sheffield. *Cities*, *50*, 28–39. https://doi.org/https://doi.org/10.1016/j.cities.2015.08.002

Liu, J., Kang, J., Behm, H., & Luo, T. (2014). Effects of landscape on soundscape perception: Soundwalks in city parks. *Landscape and Urban Planning*, *123*, 30–40. https://doi.org/https://doi.org/10.1016/j.landurbplan.2013.12.003

Liu, J., Kang, J., Luo, T., Behm, H., & Coppack, T. (2013). Spatiotemporal variability of soundscapes in a multiple functional urban area. *Landscape and Urban Planning*, *115*, 1–9. https://doi.org/https://doi.org/10.1016/j.landurbplan.2013.03.008

Ma, K. W., Mak, C. M., & Wong, H. M. (2021). Effects of environmental sound quality on soundscape preference in a public urban space. *Applied Acoustics*, *171*, 107570. https://doi.org/https://doi.org/10.1016/j.apacoust.2020.107570

McFee, B., Metsai, A., McVicar, M., Balke, S., Thomé, C., Raffel, C., Zalkow, F., Malek, A., Dana, Lee, K., Nieto, O., Ellis, D., Mason, J., Battenberg, E., Seyfarth, S., Yamamoto, R., viktoran-dreevichmorozov, Choi, K., Moore, J., … Thassilo. (2021, May). *Librosa/librosa: 0.8.1rc2* (Version 0.8.1rc2). Zenodo. https://doi.org/10.5281/zenodo.4792298

Menart, C. (2020). Evaluating the variance in convolutional neural network behavior stemming from randomness. In R. I. Hammoud, T. L. Overman, & A. Mahalanobis (Eds.), *Automatic target recognition xxx* (p. 1139410, Vol. 11394). SPIE. https://doi.org/10.1117/12.2558227

Mitchell, A., Aletta, F., & Kang, J. (2022a). How to analyse and represent quantitative soundscape data. *JASA Express Letters*, *2*(3). https://doi.org/10.1121/10.0009794

Mitchell, A., Aletta, F., & Kang, J. (2022b, March). How to analyse and represent quantitative soundscape data. https://doi.org/10.1121/10.0009794

Mitchell, A., Brown, E., Deo, R., Hou, Y., Kirton-Wingate, J., Liang, J., Sheinkman, A., Soelistyo, C., Sood, H., Wongprommoon, A., Xing, K., Yip, W., & Aletta, F. (2023a, November). Data study group final report: Iede acoustics group, university college london deep learning techniques for noise annoyance detection (delta). https://doi.org/10.5281/zenodo.10090651

Mitchell, A., Brown, E., Deo, R., Hou, Y., Kirton-Wingate, J., Liang, J., Sheinkman, A., Soelistyo, C., Sood, H., Wongprommoon, A., Xing, K., Yip, W., & Aletta, F. (2023b). Deep learning techniques for noise annoyance detection: Results from an intensive workshop at the Alan Turing Institute. *The Journal of the Acoustical Society of America*, *153*($3_{supplement}$), A262. https://doi.org/10.1121/10.0018787

Mitchell, A., Erfanian, M., Soelistyo, C., Oberman, T., Kang, J., Aldridge, R., Xue, J.-H., & Aletta, F. (2022). Effects of Soundscape Complexity on Urban Noise Annoyance Ratings: A Large-Scale Online Listening Experiment. *International Journal of Environmental Research and Public Health*, *19*(22), 14872. https://doi.org/10.3390/ijerph192214872

Mitchell, A., Erfanian, M., Soelitsyo, C., Oberman, T., & Aletta, F. (2022, October). *Delta (deep learning techniques for noise annoyance detection) dataset* (Version 1.0). Zenodo. https://doi.org/10.5281/zenodo.7158057

Mitchell, A., Oberman, T., Aletta, F., Erfanian, M., Kachlicka, M., Lionello, M., & Kang, J. (2024). The International Soundscape Database: An integrated multimedia database of urban soundscape surveys – questionnaires with acoustical and contextual information (1.0.1-alpha.1). https://doi.org/10.5281/zenodo.10672568

Mitchell, A., Oberman, T., Aletta, F., Kachlicka, M., Lionello, M., Erfanian, M., & Kang, J. (2021). Investigating urban soundscapes of the covid-19 lockdown: A predictive soundscape modeling approacha). *The Journal of the Acoustical Society of America*, *150*(6), 4474–4488. https://doi.org/10.1121/10.0008928

Mulimani, M., & Koolagudi, S. (2018, October). Acoustic event classification using spectrogram features. https://doi.org/10.1109/TENCON.2018.8650444

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 807–814.

Nilsson, M., Botteldooren, D., & De Coensel, B. (2007). Acoustic indicators of soundscape quality and noise annoyance in outdoor urban areas. *Proceedings of the 19th International Congress on Acoustics*.

Ooi, K., Ong, Z.-T., Watcharasupat, K. N., Lam, B., Hong, J. Y., & Gan, W.-S. (2024). Araus: A large-scale dataset and baseline models of affective responses to augmented urban soundscapes. *IEEE Transactions on Affective Computing*, *15*(1), 105–120. https://doi.org/10.1109/taffc.2023.3247914

Picaut, J., Can, A., Fortin, N., Ardouin, J., & Lagrange, M. (2020, April). Low-cost sensors for urban noise monitoring networks—A literature review. https://doi.org/10.3390/s20082256

Raimbault, M., & Dubois, D. (2005). Urban soundscapes: Experiences and knowledge. *Cities*, *22*(5), 339–350. https://doi.org/https://doi.org/10.1016/j.cities.2005.05.003

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

Schafer, R. M. (1977). *The tuning of the world: Toward a theory of soundscape design*. University of Pennsylvania Press.

Schafer, R. M. (1969). *The new soundscape*. BMI Canada Limited Don Mills.

Sharma, S. (2022). Analyzing Effect on Residual Learning by Gradual Narrowing Fully-Connected Layer Width and Implementing Inception Block in Convolution Layer. *Journal of Computer Science*, *18*(5), 339–349. https://doi.org/10.3844/jcssp.2022.339.349

SONYC Project. (n.d.). *Sonyc: Sounds of new york city*. Retrieved May 7, 2025, from https://wp.nyu.edu/sonyc/

Tarlao, C., Steffens, J., & Guastavino, C. (2021). Investigating contextual influences on urban soundscape evaluations with structural equation modeling. *Building and Environment*, *188*, 107490. https://doi.org/https://doi.org/10.1016/j.buildenv.2020.107490

Torija, A. J., Ruiz, D. P., & Ramos-Ridao, Á. F. (2013). A tool for urban soundscape evaluation applying Support Vector Machines for developing a soundscape classification model. *The Science of The Total Environment*, *482-483*, 440–451. https://doi.org/10.1016/j.scitotenv.2013.07.108

Västfjäll, D., Kleiner, M., & Gärling, T. (2003). Affective reactions to interior aircraft sounds. *Acta Acustica united with Acustica*, *89*(4), 693–701.

World Health Organization. (2018). *Environmental noise guidelines for the european region* (tech. rep.). WHO Regional Office for Europe. Copenhagen. https://www.who.int/europe/publications/i/item/9789289053563

World Soundscape Project. (n.d.). World Soundscape Project. https://www.sfu.ca/~truax/wsp.html

Wu, P., Zhang, Z., Peng, X., & Wang, R. (2024). Deep learning solutions for smart city challenges in urban development. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-55928-3

Yang, W., & Kang, J. (2004). Acoustic comfort evaluation in urban open public spaces. *Applied Acoustics*, *66*(2), 211–229. https://doi.org/10.1016/j.apacoust.2004.07.011

Yu, H., Chen, C., Du, X., Li, Y., Rashwan, A., Hou, L., Jin, P., Yang, F., Liu, F., Kim, J., & Li, J. (2020). TensorFlow Model Garden.

Zaffaroni-Caorsi, V., Azzimonti, O., Potenza, A., Angelini, F., Grecchi, I., Brambilla, G., Guagliumi, G., Daconto, L., Benocci, R., & Zambon, G. (2025). Exploring the soundscape in a university campus: Students' perceptions and eco-acoustic indices. *Sustainability*, *17*(8). https://doi.org/10.3390/su17083526

Zaman, K., Sah, M., Direkoglu, C., & Unoki, M. (2023). A survey of audio classification using deep learning. *IEEE Access*, *11*, 106620–106649. https://doi.org/10.1109/ACCESS.2023.3318015

Zhang, B., Leitner, J., & Thornton, S. (2019). Audio recognition using mel spectrograms and convolution neural networks. https://api.semanticscholar.org/CorpusID:237274283

Zhang, X., Huang, H., Zhang, D., Zhuang, S., Han, S., Lai, P., & Liu, H. (2024, October). *Cross-Dataset Generalization in Deep Learning* (tech. rep.). https://doi.org/https://doi.org/10.48550/arXiv.2410.11207

A

GitHub

All model architectures, training and inference scripts, and data processing notebooks used in this thesis are made available through the following GitHub repository:

- **Thesis GitHub Repository:** https://github.com/pherfkens/Thesis

This repository includes:

- `/GENERALISATION_STUDY/` – contains the data required to execute the generalisation study, including scripts for model inference and pickle generation, and evaluation notebooks for generating circumplex maps.
- `/Dataset_all_ARAUS/` – contains information about the complete ARAUS dataset, comprising a total of 25,248 files.
- `/Dataset_training_validation_test/` – contains information about the training, validation, and test sets.
- `/Feature_log_mel/` – contains a script for converting a FLAC file to a WAV file, followed by a script for converting the WAV file into a log mel spectrogram, both suitable for use with the deep learning models.
- `/MY_CNN/` – contains the models developed for this thesis, including both training and inference scripts.
- `/Other_AD_CNN/` – contains the original AD_CNN model, used as a baseline for this thesis. It includes the full pipeline for inference (testing). Originally developed by Hou et al. (2024) and slightly adapted for this thesis.
- `/SoundAQnet/` – contains the original SoundAQnet model. It includes the full pipeline for inference (testing). Originally developed by Hou et al. (2024) and slightly adapted for this thesis.
- `/pickles/` – created as part of this thesis to replicate the training pipeline. Contains both pickle generators that link audio filenames with spectrogram arrays and audio filenames with questionnaire responses.
- `README.md` – Setup instructions, dependency installation, and usage documentation.

All training, validation, and test splits were based on the splits provided in the SoundSCaper GitHub repository: https://github.com/Yuanbo2020/SoundSCaper.

To generate raw audio dataset follow the steps in the ARAUS dataset repository: https://github.com/ntudsp/araus-dataset-baseline-models.

B

Parameter calculations of CNN architectures

The calculations of the original AD_CNN as developed by Hou et al. (2024).

Table B.1: Calculations for the original model AD_CNN

Layer	Output Dimension	Calculation	Parameters
Convolution 1	$16 \times 3001 \times 64$	$(7 \times 7 \times 1) \times 16$	784
BatchNorm 1	$16 \times 3001 \times 64$	2×16	32
Convolution 2	$16 \times 3001 \times 64$	$(7 \times 7 \times 16) \times 16$	$12,544$
BatchNorm 2	$16 \times 3001 \times 64$	2×16	32
Max-pool	$16 \times 600 \times 12$	–	–
Convolution 3	$32 \times 600 \times 12$	$(7 \times 7 \times 16) \times 32$	$25,088$
BatchNorm 3	$32 \times 600 \times 12$	2×32	64
Max-pool	$32 \times 150 \times 1$	–	–
Dense 1	100	$(32 \times 150 \times 1) \times 100 + 100$	$480,100$
Event classifier	15	$100 \times 15 + 15$	$1,515$
Scene classifier	3	$100 \times 3 + 3$	303
ISO (2 outputs)	2	$2 \times (100 \times 1 + 1)$	202
PAQ (8 outputs)	8	$8 \times (100 \times 1 + 1)$	808
Total			$521,472$

The calculation for the model AD_CNN_decreased_filters.

Table B.2: Calculation for the model AD_CNN_decreased_filters

Layer	Output Dimension	Calculation	Parameters
Convolution 1	$8 \times 3001 \times 64$	$(7 \times 7 \times 1) \times 8$	392
BatchNorm 1	$8 \times 3001 \times 64$	2×8	16
Convolution 2	$8 \times 3001 \times 64$	$(7 \times 7 \times 8) \times 8$	$3,136$
BatchNorm 2	$8 \times 3001 \times 64$	2×8	16
Max-pool	$8 \times 600 \times 12$	–	–
Convolution 3	$16 \times 600 \times 12$	$(7 \times 7 \times 8) \times 16$	$6,272$
BatchNorm 3	$16 \times 600 \times 12$	2×16	32
Max-pool	$16 \times 150 \times 1$	–	–
Dense 1	100	$(16 \times 150 \times 1) \times 100 + 100$	$240,100$
Event classifier	15	$100 \times 15 + 15$	$1,515$
PAQ (8 outputs)	8	$8 \times (100 \times 1 + 1)$	808
Total			$252,287$

The calculation for the model AD_CNN_dense_layer.

Table B.3: Calculations for the model AD_CNN_dense_layer

Layer	Output Dimension	Calculation	Parameters
Convolution 1	$16 \times 3001 \times 64$	$(7 \times 7 \times 1) \times 16$	784
BatchNorm 1	$16 \times 3001 \times 64$	2×16	32
Convolution 2	$16 \times 3001 \times 64$	$(7 \times 7 \times 16) \times 16$	$12,544$
BatchNorm 2	$16 \times 3001 \times 64$	2×16	32
Max-pool	$16 \times 600 \times 12$	–	–
Convolution 3	$32 \times 600 \times 12$	$(7 \times 7 \times 16) \times 32$	$25,088$
BatchNorm 3	$32 \times 600 \times 12$	2×32	64
Max-pool	$32 \times 150 \times 1$	–	–
Dense 1	50	$(32 \times 150 \times 1) \times 50 + 50$	$240,050$
Event classifier	15	$50 \times 15 + 15$	765
PAQ (8 outputs)	8	$8 \times (50 \times 1 + 1)$	408
Total			$279,767$

The calculation for the model AD_CNN_hop_length.

Table B.4: Calculation for the model AD_CNN_hop_length

Layer	Output Dimension	Calculation	Parameters
Convolution 1	$16 \times 1501 \times 64$	$(7 \times 7 \times 1) \times 16$	784
BatchNorm 1	$16 \times 1501 \times 64$	2×16	32
Convolution 2	$16 \times 1501 \times 64$	$(7 \times 7 \times 16) \times 16$	$12,544$
BatchNorm 2	$16 \times 1501 \times 64$	2×16	32
Max-pool	$16 \times 300 \times 12$	–	–
Convolution 3	$32 \times 300 \times 12$	$(7 \times 7 \times 16) \times 32$	$25,088$
BatchNorm 3	$32 \times 300 \times 12$	2×32	64
Max-pool	$32 \times 75 \times 1$	–	–
Dense 1	100	$(32 \times 75 \times 1) \times 100 + 100$	$240,100$
Event classifier	15	$100 \times 15 + 15$	$1,515$
PAQ (8 outputs)	8	$8 \times (100 \times 1 + 1)$	808
Total			$280,967$

The calculation for the model AD_CNN_harder_max_pooling.

Table B.5: Calculations for the model AD_CNN_harder_max_pooling

Layer	Output Dimension	Calculation	Parameters
Convolution 1	$16 \times 3001 \times 64$	$(7 \times 7 \times 1) \times 16$	784
BatchNorm 1	$16 \times 3001 \times 64$	2×16	32
Convolution 2	$16 \times 3001 \times 64$	$(7 \times 7 \times 16) \times 16$	$12,544$
BatchNorm 2	$16 \times 3001 \times 64$	2×16	32
Max-pool	$16 \times 300 \times 12$	–	–
Convolution 3	$32 \times 300 \times 12$	$(7 \times 7 \times 16) \times 32$	$25,088$
BatchNorm 3	$32 \times 300 \times 12$	2×32	64
Max-pool	$32 \times 18 \times 1$	–	–
Dense 1	100	$(32 \times 18 \times 1) \times 100 + 100$	$57,700$
Event classifier	15	$100 \times 15 + 15$	$1,515$
PAQ (8 outputs)	8	$8 \times (100 \times 1 + 1)$	808
Total			$98,567$

The calculation for the model AD_CNN_dense_hop_combined.

Table B.6: Calculation for the model AD_CNN_dense_hop_combined

Layer	Output Dimension	Calculation	Parameters
Convolution 1	$16 \times 1501 \times 64$	$(7 \times 7 \times 1) \times 16$	784
BatchNorm 1	$16 \times 1501 \times 64$	2×16	32
Convolution 2	$16 \times 1501 \times 64$	$(7 \times 7 \times 16) \times 16$	$12,544$
BatchNorm 2	$16 \times 1501 \times 64$	2×16	32
Max-pool	$16 \times 300 \times 12$	–	–
Convolution 3	$32 \times 300 \times 12$	$(7 \times 7 \times 16) \times 32$	$25,088$
BatchNorm 3	$32 \times 300 \times 12$	2×32	64
Max-pool	$32 \times 75 \times 1$	–	–
Dense 1	50	$(32 \times 75 \times 1) \times 50 + 50$	$120,050$
Event classifier	15	$50 \times 15 + 15$	765
PAQ (8 outputs)	8	$8 \times (50 \times 1 + 1)$	408
Total			$159,767$

C

Results Appendix

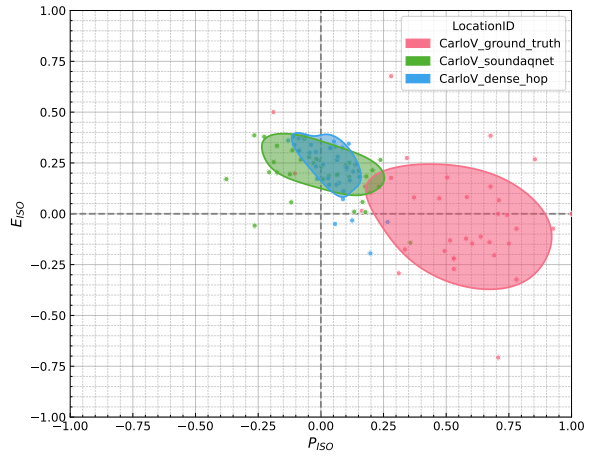
Table C.1: Perceptual Attribute MSE Comparison (Pleasant, Eventful, Chaotic, Vibrant)

#	Model	Pleasant	Eventful	Chaotic	Vibrant
1	AD_CNN_decreased_filters	1.582	1.358	1.280	1.386
2	AD_CNN_dense_layer	1.076	1.161	1.140	1.185
3	AD_CNN_hop_length	1.125	1.117	1.081	1.127
4	AD_CNN_harder_max_pooling	1.154	1.214	1.193	1.225
5	AD_CNN (baseline)	1.015	1.157	1.125	1.130
6	SoundAQnet	0.890	1.047	1.052	0.974

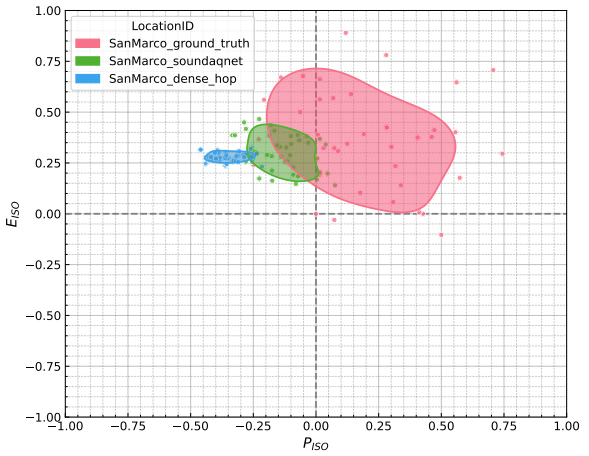
Table C.2: Perceptual Attribute MSE Comparison (Uneventful, Calm, Annoying, Monotonous)

#	Model	Uneventful	Calm	Annoying	Monotonous
1	AD_CNN_decreased_filters	1.647	1.648	1.411	1.595
2	AD_CNN_dense_layer	1.208	1.123	1.152	1.195
3	AD_CNN_hop_length	1.256	1.194	1.100	1.273
4	AD_CNN_harder_max_pooling	1.261	1.224	1.250	1.300
5	AD_CNN (baseline)	1.184	1.069	1.160	1.196
6	SoundAQnet	1.152	0.995	1.065	1.151

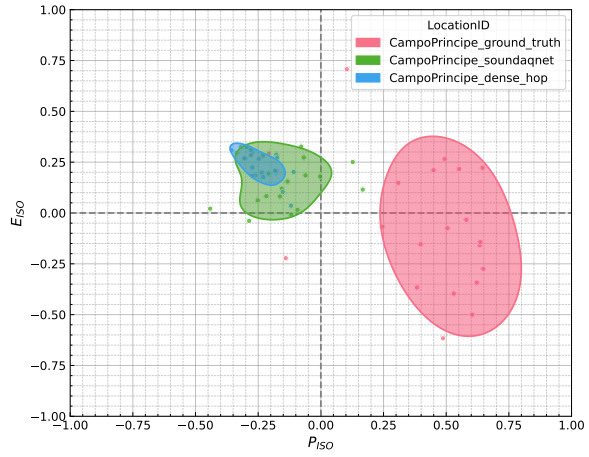
CarloV



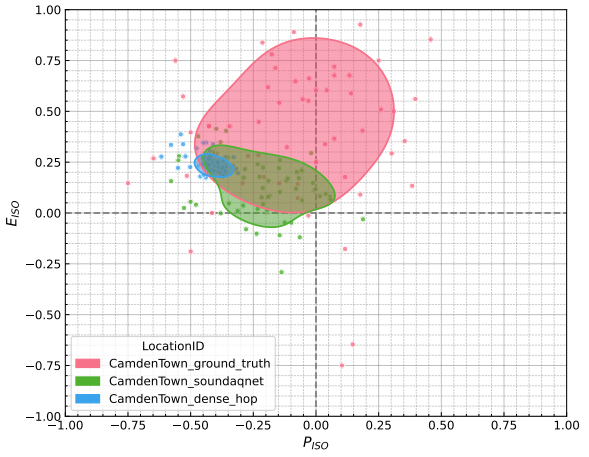
SanMarco



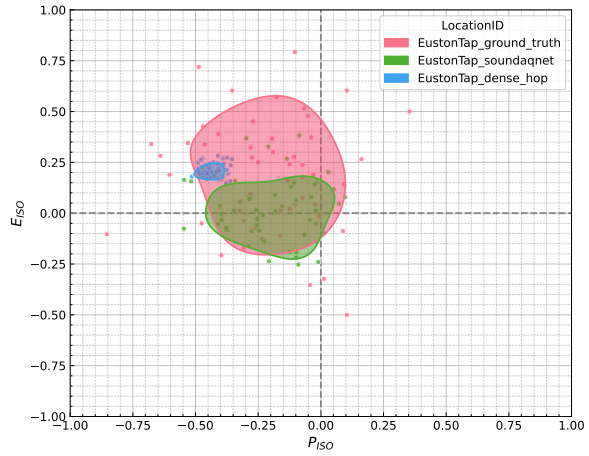
CampoPrincipe



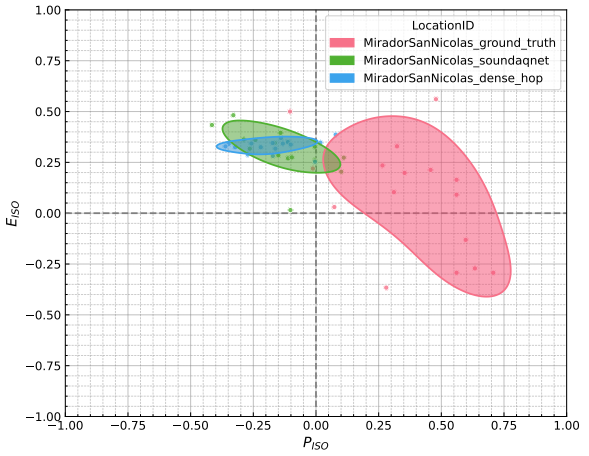
CamdenTown



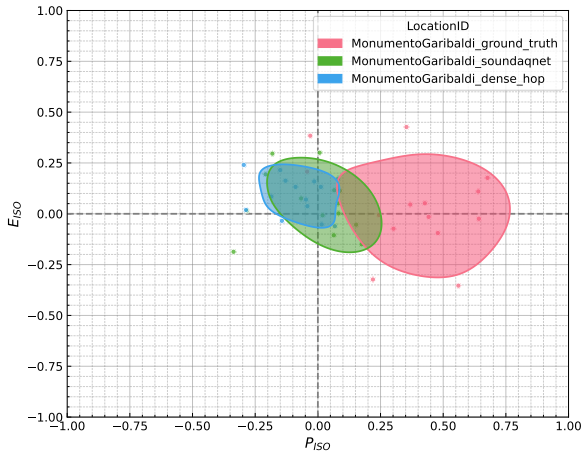
EustonTap



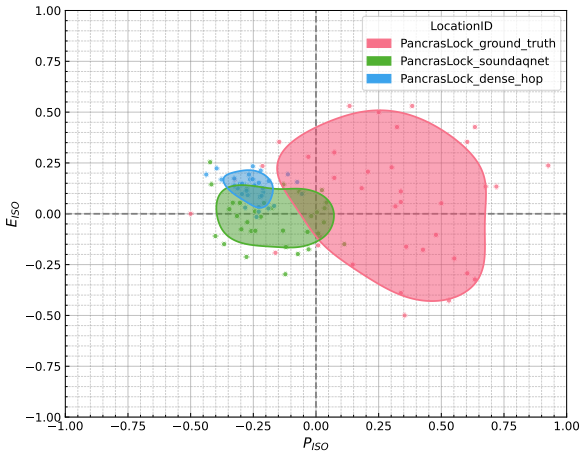
MiradorSanNicolas



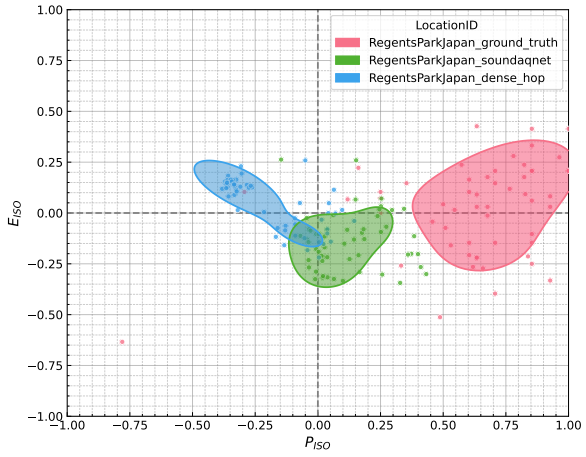
MonumentoGaribaldi



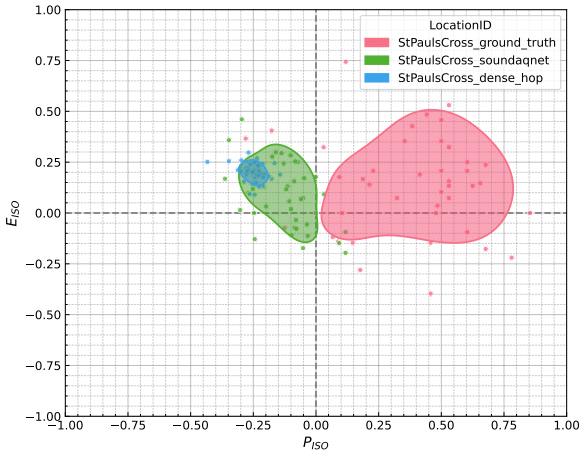
PancrasLock



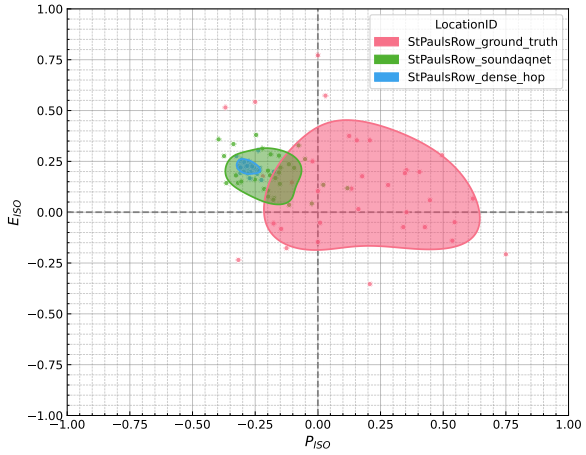
RegentsParkJapan



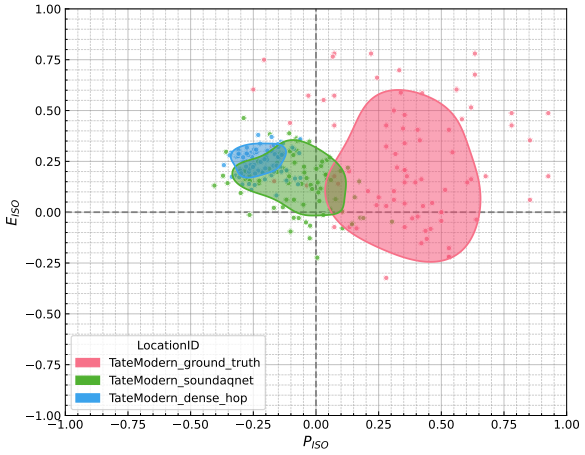
StPaulsCross



StPaulsRow



TateModern



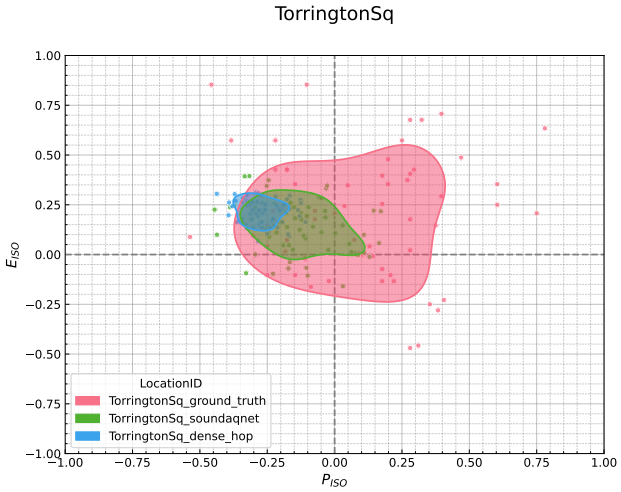


Figure C.1: Two-dimensional circumplex models of perceptual attributes for remaining soundscapes, derived from the ISD dataset