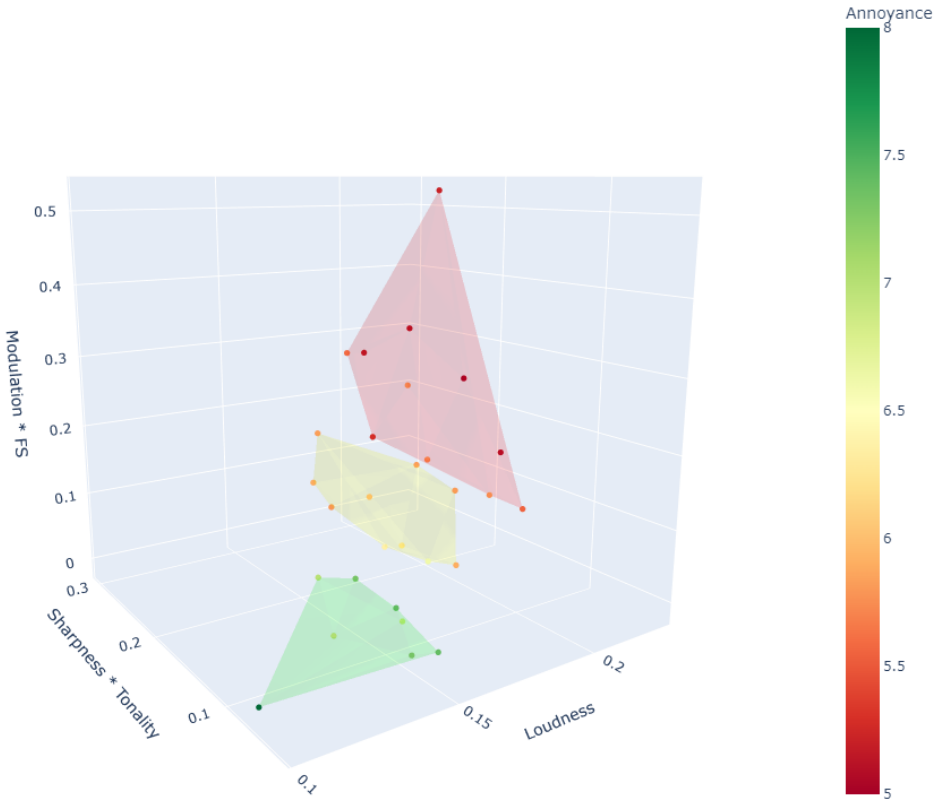
CHALMERS
UNIVERSITY OF TECHNOLOGY



Assessing Annoyance in Automotive Seat Adjustments: Perception and Prediction

Modeling of subjective annoyance responses using advanced approaches: a comparison of regression methods and neural networks

Master's thesis in Sound & Vibration

VALENTIN QUONIAM-BARRE

DEPARTMENT OF ARCHITECTURE & CIVIL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

Master's thesis 2025

Assessing Annoyance in Automotive Seat Adjustments: Perception and Prediction

Modeling of subjective annoyance responses using advanced approaches: a comparison of regression methods and neural networks

VALENTIN QUONIAM-BARRE

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Architecture & Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology

In collaboration with
Volvo Cars | *Wind Noise & Component NVH*

Gothenburg, Sweden 2025

Assessing Annoyance in Automotive Seat Adjustments: Perception and Prediction Modeling of subjective annoyance responses using advanced approaches: a comparison of regression methods and neural networks
VALENTIN QUONIAM-BARRE

Supervisors:	Jens Ahrens	Division of Applied Acoustics
	Karin Kvickström	Volvo Cars | Component NVH
Examiner:	Jens Ahrens	Division of Applied Acoustics

Master's Thesis 2025
Department of Architecture and Civil Engineering
Division of Applied Acoustics
In collaboration with Volvo Cars
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 2200

Cover: Clustering of sound stimuli in psychoacoustic feature space, colored by their predicted annoyance, on a scale from 5 (high annoyance) to 8 (low annoyance).

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2025

Assessing Annoyance in Automotive Seat Adjustments: Perception and Prediction
Benefits of a large-scale listening test for advanced modeling approaches: comparing regression methods and neural networks
VALENTIN QUONIAM-BARRE
Department of Applied Acoustics
Chalmers University of Technology

Abstract

In the automotive industry, acoustic comfort is a key aspect of perceived quality, especially in high-end vehicles where even subtle noise such as squeaks, buzzes, or rattles can negatively impact user satisfaction. Sound Quality (SQ) assessments help ensure a premium experience but typically rely on subjective jury testing, which is time-consuming and depends on expert judgment that may not reflect everyday user expectations.

This thesis aims to develop a predictive model for annoyance ratings based on objective acoustic parameters. This is a challenging task, as annoyance is inherently subjective and influenced by various perceptual factors. While traditional methods such as linear regression can estimate simple perceptual attributes like loudness, they fall short when modeling more complex, non-linear characteristics. To overcome these limitations, machine learning approaches, including neural networks and random forests, are investigated and compared to linear and polynomial regression models.

The study focuses on seat adjustment mechanisms as a use case. These sounds are relatively easy to isolate and analyze, show noticeable variation across vehicle brands, and are less affected by external noise sources. This makes them a suitable candidate for controlled testing and model development. Using those sounds, a large scale listening test is made to assess annoyance and be able to train the models. Results demonstrate that machine learning models can successfully predict perceived annoyance based on objective metrics, offering a promising alternative to traditional jury testing. Such models could significantly improve the efficiency and scalability of SQ evaluations in the automotive industry.

Keywords: Psychoacoustics, Machine Learning, Jury Testing, Sound Quality, Regression, Neural Networks, Prediction

Acknowledgements

I would like to express my gratitude to my supervisor, Jens. His insightful guidance and genuine enthusiasm have been invaluable throughout the course of this project.

I'm also deeply thankful to everyone at Volvo who helped me feel welcome and supported. In particular, I'd like to thank the Component NVH team, the Wind Noise team, and the ASX group for their help. Deep thanks to Karin, Yidan, Piyush, Berker, and Melina for their generous assistance and encouragement.

Finally, I am sincerely grateful to all the friends and staff in the Applied Acoustics division. Over the past two years, this inspiring environment has significantly shaped both my academic journey and personal growth.

Valentin Quoniam-Barre, Gothenburg 2025

Disclosure - Usage of AI Tools

AI tools were employed to support certain aspects of this thesis work, such as:

- Plot formatting in Matlab
- LaTeX formatting, especially tables and visuals
- Code Optimization

However, AI was not used to write the main report or to generate any of the core code behind the thesis work.

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

ASX	Active Sound Experience
DSP	Digital Signal Processing
MA	Modulation Analysis
NN	Neural Networks
NVH	Noise, Vibration, Harshness
OEM	Original Equipment Manufacturer
OSQ	Objective Sound Quality
SPL	Sound Pressure Level
SQ	Sound Quality
RF	Random Forest

Nomenclature

Below is the nomenclature of indices, sets, parameters, and variables that have been used throughout this thesis.

Parameters

A	Annoyance
AMS	Average Modulation Spectrum
F	Fluctuation Strength
ρ	Pearson correlation coefficient
$RMSE$	Root Mean Square Error
R	Roughness
S	Sharpness
SC_d	Singular Concordance
SC_s	Singular Consistency
SSE	Sum of Squared Errors
T	Tonality
N	Total Loudness

Variables

res_n	n^{th} residual of a prediction in comparison to the original data
ϵ_K	Noise added to the parameter K
c_i	i^{th} coefficient of a regression model
$N'(z)$	Specific Loudness in the critical band z

Contents

List of Acronyms **IX**

Nomenclature **XI**

List of Figures **XVII**

List of Tables **XXI**

1 Introduction **1**
1.1 Historical overview . . . 1
1.2 Background & motivation . . . 2
1.3 Aim . . . 3
1.3.1 Research questions . . . 3
1.3.2 Objectives . . . 3
1.4 Scope and limitations . . . 3
1.4.1 Simplification of the annoyance . . . 3
1.4.2 Electric seat adjustment mechanisms . . . 4
1.4.3 Dataset size . . . 4
1.5 Previous work . . . 4
1.6 Methodology . . . 6
1.6.1 Data sorting . . . 6
1.6.2 Jury testing . . . 6
1.6.3 Models training . . . 7
1.6.4 Evaluation & models comparison . . . 7
1.7 Thesis outline . . . 8

2 Theory **9**
2.1 Product noise measurement . . . 9
2.2 Psychoacoustics for sound analysis . . . 10
2.2.1 Loudness . . . 11
2.2.2 Critical bands . . . 12
2.2.3 Perceptual effects . . . 12
2.2.3.1 Sharpness . . . 12
2.2.3.2 Tonality . . . 13
2.2.4 Modulation . . . 13
2.2.4.1 Fluctuation Strength . . . 14
2.2.4.2 Roughness . . . 14

2.2.4.3 Modulation analysis . . . 15
2.3 Listening test design . . . 16
2.3.1 Test environment . . . 16
2.3.2 Participant management . . . 16
2.3.3 Test procedure . . . 17
2.3.4 Unbiasing . . . 18
2.4 Regression methods . . . 19
2.4.1 Multiple linear Regression . . . 19
2.4.2 Multiple polynomial regression & interactions . . . 19
2.5 Artificial neural networks . . . 20
2.5.1 Structure . . . 20
2.5.2 Learning process . . . 21
2.6 Random forest . . . 22
2.7 Audio data augmentation . . . 24

3 Methods **27**
3.1 Data gathering and preprocessing . . . 27
3.1.1 Initial dataset . . . 27
3.1.2 File structure and protocol compliance . . . 27
3.1.3 Diversity-based reduction . . . 29
3.1.4 Backrest Sounds . . . 31
3.2 Listening experiment . . . 32
3.2.1 Design . . . 32
3.2.2 Environment . . . 34
3.2.3 Organization . . . 36
3.3 Analysis of the test results . . . 37
3.3.1 Consistency . . . 37
3.3.2 Concordance . . . 38
3.3.3 Hierarchical Clustering . . . 39
3.3.4 Metrics ratings . . . 39
3.4 Feature Selection . . . 40
3.5 Predictions & Comparison . . . 40
3.5.1 Performance evaluation . . . 41
3.5.1.1 Experiment scale . . . 41
3.5.1.2 Volvo Scale . . . 42
3.5.2 Linear & polynomial regression . . . 44
3.5.3 Machine learning models . . . 44
3.5.3.1 Simple neural network . . . 44
3.5.3.2 Random forest . . . 45
3.6 Data augmentation . . . 45

4 Results **47**
4.1 Jury Testing . . . 47
4.1.1 Participant summary . . . 47
4.1.2 Consistency vs. Concordance . . . 48
4.1.3 Experiment ratings . . . 49
4.1.4 Effects of unbiasing . . . 51

4.1.4.1 Consistency & concordance . . . 51
4.1.4.2 Ratings . . . 51
4.1.5 Groups comparison . . . 52
4.2 Feature Choice . . . 56
4.2.1 Experiment's metrics . . . 56
4.2.2 Objective metrics . . . 57
4.3 Annoyance Predictions . . . 59
4.3.1 Linear regression . . . 60
4.3.2 Polynomial regression . . . 61
4.3.2.1 Original polynomial . . . 61
4.3.2.2 Simplified polynomial . . . 63
4.3.3 Simple neural network . . . 65
4.3.4 Random forest . . . 66
4.3.5 Effects of data augmentation . . . 69
4.3.5.1 On the metrics . . . 69
4.3.5.2 On the sounds . . . 72
4.3.6 Best model : RMSE comparison . . . 72

5 Discussion **75**
5.1 Interpretation of Key Results . . . 75
5.1.1 Models performances . . . 75
5.1.2 Dataset size and data augmentation . . . 76
5.2 Methodological reflections . . . 76
5.2.1 Listening test design . . . 76
5.2.2 Presence of a bias . . . 77
5.2.3 Feature selection . . . 77
5.3 Applications and future work . . . 78

6 Conclusion **79**

Bibliography **81**

List of Figures

1.1 Volvo rating scale used for jury testing 2
1.2 Example of the previous work done by Hasselström 5

2.1 Different possible measurement setups : Binaural head (left), Simplified Binaural (middle) and over-ear microphones (right) 10
2.2 Comparison of time signal and 3rd-octave spectrum for an electric motor (left) and a Mozart concert (right) - Almost identical 10
2.3 Equal Loudness Curves [35] 11
2.4 Critical bands rate 12
2.5 Sharpness of different sounds as a function of frequency 13
2.6 Different type of modulation parameters 14
2.7 Temporal masking pattern of a sinusoidally amplitude-modulated sound 14
2.8 Roughness relative to f_{mod} 15
2.9 Algorithm for the Modulation Analysis [20] 15
2.10 Example of a listening test interface using mixed evaluation methods 18
2.11 Nonlinear model of a neuron [19] 20
2.12 Illustration of a neural network with 10 inputs and 2 outputs [19] . . 21
2.13 Summary of a neural network training process [34] 22
2.14 Example of decision tree partitions on 2D data, where each rectangle represents a leaf region in the feature space. 23
2.15 Effects of pitch shifting on a rooster crowing 24

3.1 Waveform with annotated verbal cues preceding each movement . . . 28
3.2 Waveform without verbal annotations and with minor artifacts 28
3.3 Waveform of a sound from the final dataset 29
3.4 Distribution of candidate sounds in 3D psychoacoustic space (Red = Up, Blue = Down) 30
3.5 Final selected test sounds 30
3.6 Overview of additional psychoacoustic metrics considered 31
3.7 Distribution of backrest adjustment sounds (Red = Forward, Blue = Backward) 32
3.8 User interface of the listening test 33
3.9 Demographic questionnaire 34
3.10 Listening test setup at Volvo: car seats mounted on wooden pallets in a side-facing configuration. 35
3.11 Listening test setup at Chalmers: simplified environment without car seats. 35

3.12 Listening test invitation 36
3.13 Example of dendrogram trees without clusters 39
3.14 Example of a reference plot comparing predicted and original ratings 41
3.15 Comparison between prediction CI and original rating CI 42
3.16 Example of ratings comparison in the Volvo scale 43
3.17 Exemple of a reference plot after data augmentation 46

4.1 Distribution of participant with their acoustic background 47
4.2 Distribution of participants across age groups 48
4.3 Consistency vs. Concordance for all 4 metrics 48
4.4 Box plots ratings of annoyance 49
4.5 Box plots for loudness, modulation and sharpness from the listening test 50
4.6 Consistency vs. Concordance after unbiasing 51
4.7 Annoyance ratings before and after unbiasing 52
4.8 Correlation matrix between participant groups 53
4.9 RMSE matrix showing differences in ratings between participant groups 53
4.10 dendrogram for the each participant group 55
4.11 Correlation matrix between annoyance and test metrics 56
4.12 Correlation matrix between objective metrics and subjective annoyance ratings 59
4.13 Prediction results for the linear regression 60
4.14 Linear regression prediction over the annoyance box plots 60
4.15 Linear regression results in the Volvo scale 61
4.16 Prediction results for the MPR 62
4.17 MPR prediction over the annoyance box plots 62
4.18 MPR results in the Volvo scale 63
4.19 Prediction results for the Simplified MPR 63
4.20 Simplified MPR predictions over the annoyance box plots 64
4.21 Simplified MPR results in the Volvo scale 64
4.22 Prediction results for the neural network 65
4.23 Neural network predictions over the annoyance box plots 65
4.24 Neural network results in the Volvo scale 66
4.25 Prediction results for the random forest 66
4.26 Surface of RMSE values used to find the best hyperparameters for the random forest model 67
4.27 Training and testing loss versus number of trees, showing overfitting beyond a certain point 67
4.28 Example of a single decision tree used in the random forest 68
4.29 Random forest results in the Volvo scale 68
4.30 Predicted vs. actual annoyance with data augmentation: Top left – linear regression; Top right – polynomial regression; Bottom left – simplified polynomial regression; Bottom right – neural network. . . . 69
4.31 Neural network predictions over annoyance box plots 70
4.32 Random forest predictions after data augmentation 70
4.33 Hyperparameter tuning for the random forest with data augmentation 71

4.34 Sound clustering in 3D feature space using the random forest model trained on augmented data. 71
4.35 Prediction results for hypothesis 1 (left) and hypothesis 2 (right) . . . 72
4.36 Comparison of Train and Test RMSE for all models 73

List of Tables

3.1 Playback configuration 35
3.2 Loudness and sharpness related metrics 40
3.3 Modulation related metrics 40

4.1 Loudness correlation coefficients 57
4.2 Sharpness correlation coefficients 57
4.3 Modulation correlation coefficients 58

1
Introduction

This chapter lays the foundation for the thesis by presenting the context of the work, its objectives and its limitations.

1.1 Historical overview

Historically, acoustics have not been a primary concern for manufacturers, who have often focused more on performance, safety, and aesthetics. The concept of acoustic comfort began to take shape in the early 20th century, notably following the advent of telephone communication. Researchers at Bell Telephone Laboratories were among the first to study and characterize sound perception, exploring aspects such as loudness and articulation [14, 13]. Around the same time, studies revealed the psychological impact of sound on workers, particularly in relation to loudness and pitch [27, 41]. This growing awareness led to the establishment of acoustics research labs worldwide, including those at Chalmers University and RCA Laboratories, both build in 1943 [36].
The Hi-Fi boom of the 1970s marked a real shift in how people thought about sound. As high-fidelity audio systems became more popular, expectations for acoustic comfort started rising. This led to the emergence of Noise, Vibration, and Harshness (NVH) as a dedicated area of engineering. Around the same time, the field of psychoacoustics began to take shape, offering ways to describe and measure how we perceive sound [12, 33]. Researchers soon realized that these perceptions were closely tied to Sound Quality. [15, 24, 8]

Since then, NVH has become a key part of product design, especially in the automotive industry [16]. Listening test procedures have been standardized [38] and developed [7], to ensure repeatability and relevancy. Those kinds of tests are now incorporated in the product design.

Today, the focus is on finding ways to predict subjective impressions using objective measurements, making sound quality evaluations faster, more reliable, and less expensive [9, 1, 26, 3, 2].

1.2 Background & motivation

Sound quality is a crucial factor influencing user experience and comfort, making it a key consideration in the design of industrial equipment, particularly vehicles. In the case of a daily used object like a car, even minor issues such as a rattle when adjusting the seat or a squeak from the wipers can become major annoyances for users, despite the vehicle functioning perfectly otherwise.

To evaluate sound quality, product development teams often use jury testing. These tests involve human participants, typically experts, who listen to audio recordings and rate them using both predefined scales and their own acoustic standards. This feedback is then used to help optimize acoustic performance.

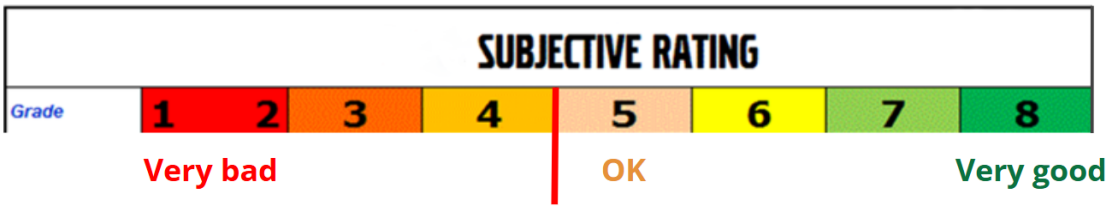


Figure 1.1: Volvo rating scale used for jury testing

However, the metrics rated by listening experts are quite distinctive, as these specialists apply their professional expertise to evaluations that may not always align with the average customer's perception. This, along with the time-intensive nature of the tests and the challenge of assembling qualified listening experts, explains why the jury testing process, despite its value, contains several inherent limitations. These constraints can impact both the feasibility of conducting regular evaluations and the applicability of results to real-world customer experiences.

An alternative approach involves using predictive models to link objective sound metrics with subjective quality assessments. Traditional linear regression methods offer a straightforward starting point [8, 3, 24, 1], establishing mathematical relationships between measurable sound parameters and quality ratings. They have been widely used for their simplicity and reliability. However, they may struggle with highly complex, non-linear acoustic relationships.

Building upon these foundations, machine learning techniques, particularly neural networks, present a more promising solution by identifying intricate patterns and relationships between these metrics. Renowned for their ability to approximate complex functions, neural networks can bridge the gap between objective data and subjective perception, reducing reliance on repetitive jury tests while capturing nuances that simpler statistical models might overlook. A wide range of architectures is available, from simple models [26, 28] to deep networks [46], with adjustable depth to suit the specific characteristics of the sounds being analyzed. Additionally, various data processing techniques can be employed to enhance performance [30, 39],

which make neural networks a truly powerful tool.

1.3 Aim

The aim of this thesis is to design a model that leverages the acoustic properties of signals—such as psychoacoustic metrics and other parameters—to provide meaningful predictions.

1.3.1 Research questions

This thesis aims to answer the following research questions :

- What's the best way to gather enough data to train a neural network in the context of sound annoyance ?
- How accurately can subjective annoyance ratings be predicted from objective sound quality metrics, using different methods ?
- Which psychoacoustic metrics (e.g., loudness, sharpness, roughness, tonality) are most influential in determining perceived annoyance in an automotive environment ?

1.3.2 Objectives

Thus, the objectives can be stated as follows :

- Develop a data acquisition and preprocessing strategy that ensures the robustness and scalability of annoyance prediction models.
- Design a methodology for modeling perceived sound annoyance in seat adjustment mechanisms using both traditional and machine learning techniques.
- Identify and extract relevant psychoacoustic features that effectively capture perceptual aspects of sound quality.
- Benchmark the performance of different predictive approaches and determine their suitability for use in product development.

1.4 Scope and limitations

This section present the limitation of this thesis work and its extent.

1.4.1 Simplification of the annoyance

In vehicle engineering, the primary metric used is often referred to as annoyance. However, the evaluation process is more nuanced: experts are not simply judging whether a sound is annoying, but whether it is acceptable enough for production without generating significant customer complaints.

The term annoyance will be used here for the sake of simplicity, with the understanding that it does not fully capture the complexity of expert assessments. This choice is especially practical when involving both experts and non-experts in listening tests, as annoyance is a term that is easily understood across audiences.

1.4.2 Electric seat adjustment mechanisms

The focus is placed on a single component, such as seat adjustment mechanisms, which are relatively straightforward to test in a vehicle setting. These systems vary significantly between brands and are less affected by external noise sources (e.g., fabric friction when adjusting a backrest). Limiting the scope in this way ensures a controlled environment for refining the model. Having multiple components at this stage would significantly increase complexity, particularly since sound preprocessing must still be done manually.

1.4.3 Dataset size

As is often the case in sound quality analysis, the available dataset is relatively limited. Consequently, the model developed in this thesis cannot be expected to achieve perfect accuracy or broad generalization. The focus will therefore be on collecting a diverse range of input sounds, varying in pitch, sound level, tonality, and other acoustic characteristics. To address the data scarcity, data augmentation techniques will be explored. The effectiveness of such methods is known to vary depending on the context [39, 26]. Their suitability was assessed in this thesis and found to have limited impact on performance. However, the experiments highlighted a reassuring level of robustness (small variations in input led to only minor changes in output) indicating consistent and stable behavior rather than chaotic responses.

1.5 Previous work

This thesis builds upon a benchmarking study conducted in 2022 by Hasselström [18], which proposed an initial methodology for predicting overall sound quality (OSQ) based on internal Volvo Cars data. Although a preliminary sorting of the sound recordings was carried out, the project was never completed due to the absence of a proper listening test to gather subjective ratings.

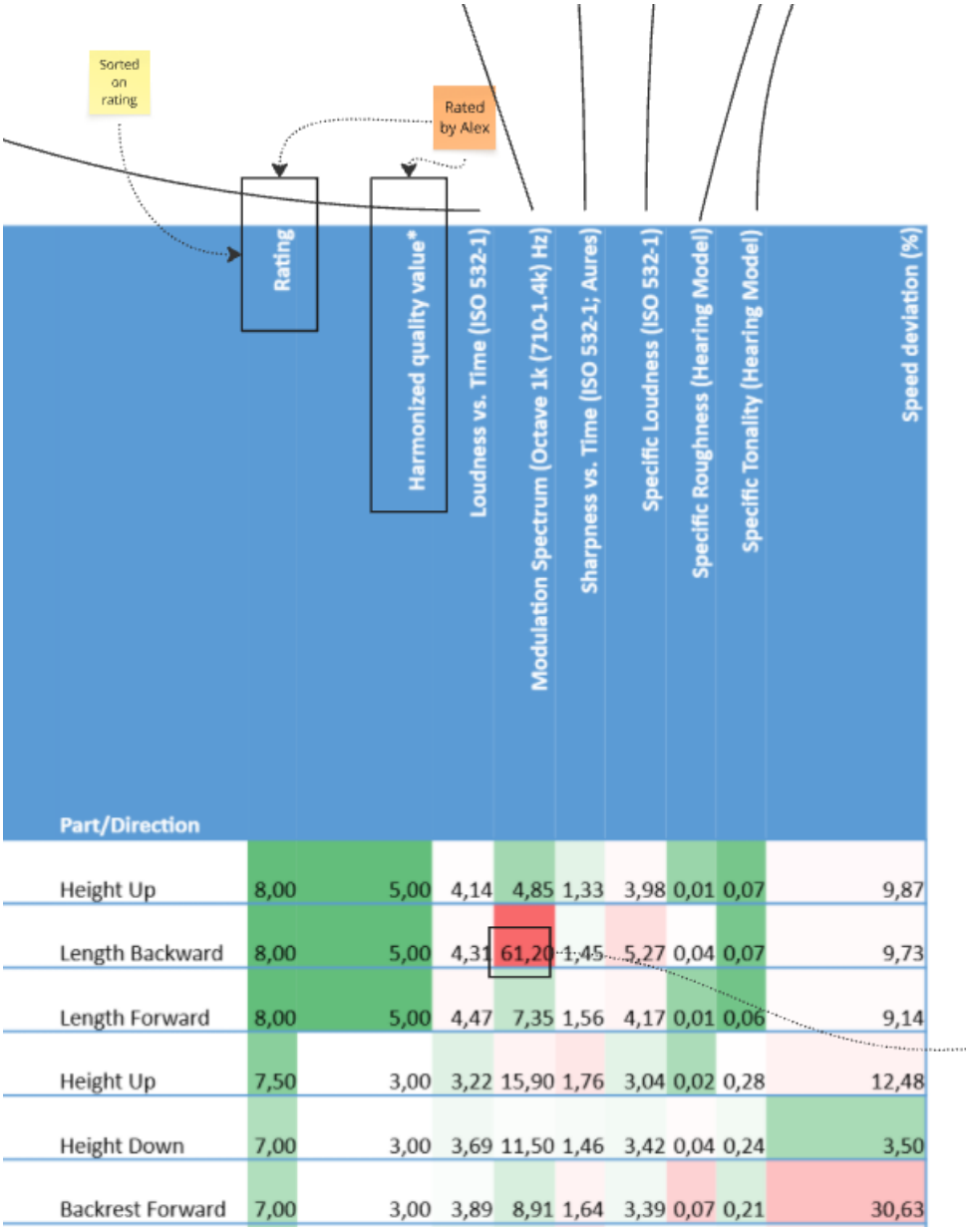


Figure 1.2: Example of the previous work done by Hasselström

The work is also inspired by a previous master's thesis by Kunte [26], which investigated the use of neural networks for sound quality prediction. However, that study was limited by a small dataset and several methodological constraints, which this thesis seeks to address and improve upon.

Moreover, earlier listening tests conducted at Volvo [2] will serve as a foundation for designing and executing the current evaluation procedure. These past efforts provide valuable insights and best practices to ensure high-quality subjective data collection in an efficient and reliable manner.

1.6 Methodology

This thesis adopts an experimental approach to predict sound annoyance using machine learning models trained on psychoacoustic features. The methodology is structured into four key stages: data preparation, listening test design, model training, and performance evaluation.

1.6.1 Data sorting

The initial step involves selecting and preparing a representative dataset of seat adjustment sounds. These recordings must be manually preprocessed to remove irrelevant segments, such as operator speech, background noise, or accidental clicks.They are then normalized to the same duration to ensure the consistency of the listening test [18] and the predictions.

Once cleaned, the sounds are analyzed to extract several psychoacoustic metrics, including loudness, sharpness, and tonality. This analysis, along with all metric computations, is performed using the Artemis Suite software by Head Acoustics.

From the larger pool of recordings, a subset of 33 samples is carefully selected to reflect a broad range of perceptual characteristics. This curated selection includes sounds with varying loudness levels, different degrees of sharpness, tonal and non-tonal content, as well as both modulated and stable signals. It ensures sufficient diversity to build efficient prediction models [25].

1.6.2 Jury testing

The next step involves designing a listening test in which participants rate the perceived annoyance of each sound, along with additional parameters useful for cross-correlation and analysis. Conducting this test represents a primary objective of the thesis. A sufficiently large and diverse participant pool is essential to ensure reliable and unbiased results. For this reason, the test must remain intuitive and accessible, regardless of participants' backgrounds [4, 6, 38].

To validate the test setup, several pilot sessions are organized with both NVH team members and external participants. Realism is enhanced through methods such as aurally accurate playback systems and the use of actual car seats, helping participants remain immersed in a context that resembles an in-vehicle experience, even within a controlled listening environment. The listening sessions are conducted at both Volvo Cars and Chalmers, using the SQala software from Head Acoustics.

The final version of the test includes clear instructions, a brief training phase, and an optimized interface to minimize confusion and cognitive load. Care is also taken to keep the overall duration reasonable in order to maintain participant engagement and reduce fatigue, contributing to more consistent and meaningful evaluations. An ideal test length would be around 20-30 minutes.

A custom scale from 1 (low annoyance) to 8 (high annoyance) is adopted for the listening test, offering a broad and interpretable range. As a result, the test scale is inverted relative to the Volvo scale (going from high to low annoyance). A method for mapping between these two scales is explored later in the thesis. For the additional perceptual metrics, a simpler scale from 1 to 5 is used, as it provides sufficient resolution without increasing cognitive load.

1.6.3 Models training

Once the listening test is completed, the collected data is organized and used to train several machine learning models. Initial sorting and cleaning are performed in Excel using Power Query, after which the data is imported into MATLAB for further processing and analysis. To simplify computation and ensure consistency across analyses, annoyance ratings originally collected on a 1–8 scale are rescaled to a 1–5 scale. Most results presented in this thesis refer to the rescaled 1–5 format.

Each psychoacoustic metric—such as loudness, sharpness, or modulation—can be computed using different methods and in accordance with various standards [21, 23]. To identify the most suitable calculation method for this study, correlation coefficients are computed between the objective metrics and the corresponding subjective ratings from the listening test. The method yielding the highest correlation is considered the most appropriate.

A baseline linear regression model is implemented in MATLAB to assess feasibility and highlight the most influential variables. In parallel, a more advanced model capable of capturing complex, non-linear relationships is developed in Python using dedicated machine learning libraries. This combined approach offers flexibility and enables a direct comparison between simpler and more sophisticated modeling strategies.

From the pool of 33 sound samples, 70% are allocated to model training, while the remaining 30% are reserved for testing. This training-testing split follows standard practice and is generally effective in ensuring both reliable learning and generalization performance [42].

1.6.4 Evaluation & models comparison

To evaluate model performance, root mean square error (RMSE) is used as the primary metric, providing a direct indication of the prediction accuracy. Although other metrics—such as mean absolute error or correlation coefficients—could also be considered, the focus here remains on simplicity and clarity.

Results are plotted with 95% confidence intervals for both the listening test ratings and the model predictions (when possible). Since the ultimate objective is to predict values on the original Volvo scale (ranging from 5 to 8), all results are rescaled

accordingly. Model outputs are then compared to historical ratings provided by the Volvo NVH team in recent years. The aim is not to replicate these values exactly. It is rather to observe whether similar trends emerge across the range of sounds.

In addition, the sounds are represented in a three-dimensional space defined by key psychoacoustic metrics. This visualization aims to reveal potential clustering patterns—ideally distinguishing between good, mid-level, and poor sound quality categories.

1.7 Thesis outline

The thesis begins with a presentation of the relevant background theory:

1. Sound quality and psychoacoustics
2. Listening test design principles
3. Prediction methods

In the method section, the selected approaches are presented and explained. This includes the data collection process, the selection of relevant features for prediction, and considerations for improving model performance.

The following section presents the results of the listening test, including annoyance ratings and participant data. The different models are compared, and the prediction outcomes are evaluated.

The discussion section offers an analysis of the results in relation to the research questions. Potential improvements and alternative approaches are also considered.

Finally, the conclusion summarizes the main findings and reflects on the overall contribution of the work.

2

Theory

This chapter provides an overview of the theoretical background relevant to this thesis. It describes all the needed knowledge to develop a sound quality metric, such as the fundamentals of psychoacoustics, the principles behind effective listening test design and the functioning of various prediction models.

2.1 Product noise measurement

A fundamental rule in noise measurement is the consistency of the measurement environment. To ensure the reliability and comparability of results, all measurements must be performed under the same conditions. This consistency is essential to reduce variability [40, 2] :

- **Same location**: identical room acoustics and background noise.
- **Same equipment**: same microphones, etc.
- **Same setup**: identical microphone positions, angles, and distances.
- **Same procedure**: consistent signal types, durations, and sequences.

In the context of seat adjustments, and to imitate human hearing perception, several measurement methods can be considered. Each has its own advantages and limitations depending on the available time, realism requirements, and resources:

1. **Artificial head (binaural recording):**
    - Closely mimics human auditory perception
    - Higher cost and more demanding setup.
2. **Microphones at the ear position (simplified binaural):**
    - Good approximation of ear-level sound reception.
    - Lower cost compared to artificial heads.
3. **Human subject with over-ear microphones:**
    - Allows real-time subjective feedback.
    - Hard to reproduce consistently and introduces variability

The method choice depends on the desired trade-off between realism, reproducibility, and cost (in term of time and money). Here is an illustration of the three measurement setups :

Figure 2.1: Different possible measurement setups : Binaural head (left), Simplified Binaural (middle) and over-ear microphones (right)

2.2 Psychoacoustics for sound analysis

Once the measurements are done, they have to be analyzed. Understanding sound solely through physical parameters such as pressure level or frequency spectrum is often insufficient when evaluating human perception. In reality , different kind of sources contribute to the noise : broadband and narrow-band noises, tonal components, modulated sounds, etc.

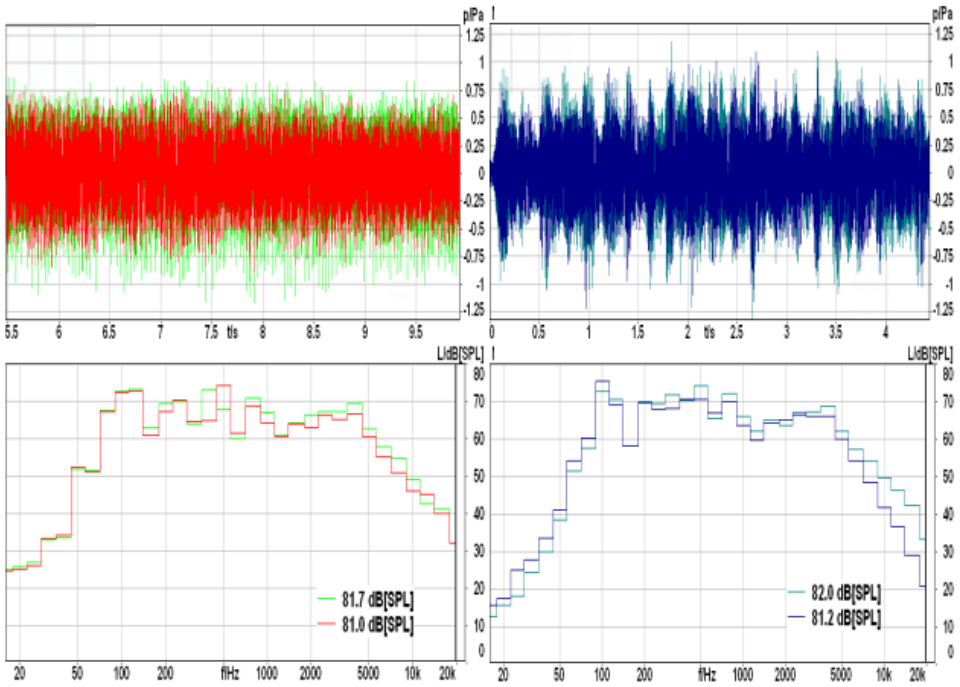


Figure 2.2: Comparison of time signal and 3rd-octave spectrum for an electric motor (left) and a Mozart concert (right) - Almost identical

Psychoacoustics bridges this gap by studying how humans interpret and respond to auditory stimuli. Several metrics have been developed to quantify perceived attributes of sound:

2.2.1 Loudness

Loudness is the perceived intensity of a sound, more precisely "the magnitude of an auditory sensation" [14]. Unlike sound pressure level (SPL), which is a physical quantity.

At the reference frequency of 1 kHz, the loudness level in phons is numerically equal to the sound pressure level in dB_{SPL}. For other frequencies, however, the phon scale compensates for the ear's frequency-dependent sensitivity. Several standards can be used to calculate loudness : DIN 45631/A1, ISO 532-3, ANSI S3.4 2007. Most recent methods account for factors such as spectral content, bandwidth, duration, and masking [23, 21]. This is illustrated by the equal-loudness contours :

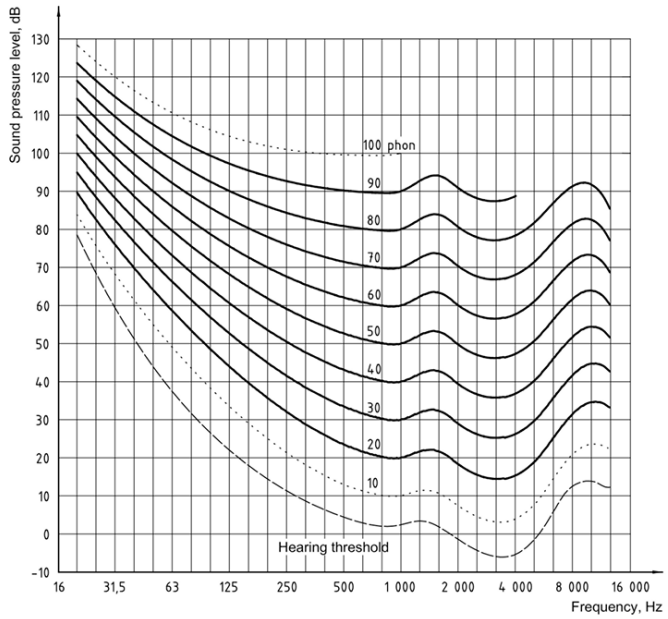


Figure 2.3: Equal Loudness Curves [35]

The *sone* is a linear unit of perceived loudness derived from the logarithmic phon scale. The relationship between loudness N (in sones) and loudness level L_N (in phons) is given by:

$$N = 2^{\left(\frac{L_N - 40}{10}\right)} \tag{2.1}$$

According to equation 2.1, every increase of 10 phons results in a doubling of the perceived loudness.

2.2.2 Critical bands

Before introducing the other parameters, it is essential to understand how the human ear processes sound. It does not perceive all frequencies with equal resolution, but instead is more sensitive to differences at low frequencies and less precise at higher frequencies. This frequency resolution is described using the unit *Bark*. Zwicker organized the audible frequency range into 24 critical bands, corresponding to a scale from 0 to 24 Bark [12], as illustrated in Figure 2.4.

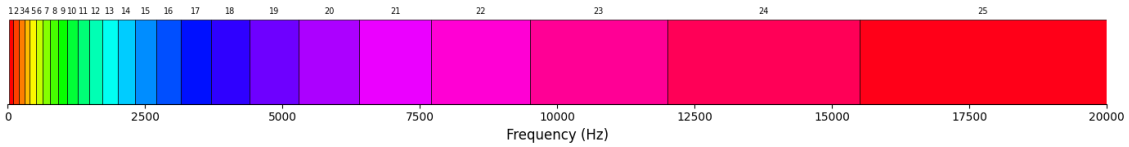


Figure 2.4: Critical bands rate

These critical bands form the basis for calculating *specific loudness*, a concept that describes how perceived loudness is distributed across the frequency spectrum (denoted as $N'(z)$, in *sone/Bark*). The total loudness is then obtained by integrating the specific loudness over all 24 critical bands:

$$N = \int_0^{24\ \text{Bark}} N'(z)\,dz \tag{2.2}$$

2.2.3 Perceptual effects

Loudness alone is not sufficient to fully characterize a sound. For the human ear, high-frequency components tend to be perceived as louder than low-frequency ones, even at the same sound pressure level. In addition, prominent tonal components—regardless of their absolute level—can significantly affect the perception and evaluation of sound, often increasing annoyance. Therefore, these perceptual effects must also be taken into account.

2.2.3.1 Sharpness

Sharpness quantifies the perceptual sensation of high-frequency content. A common model, proposed by von Bismarck and later refined by Zwicker [12, 21], defines sharpness S as:

$$S = \frac{\int_0^{24\ \text{Bark}} N'(z) \cdot g(z) \cdot z\,dz}{\int_0^{24\ \text{Bark}} N'(z)\,dz} \tag{2.3}$$

where:

- $N'(z)$ is the specific loudness in sone/Bark,
- z is the critical band rate in Bark,
- $g(z)$ is a weighting function that increases with frequency, typically with $g(z) = 1$ for $z \leq 16$, and exponential beyond.

The unit of sharpness is the *acum*, where 1 acum corresponds to the sharpness of a 1 kHz pure tone at 60 dB SPL. Sharpness generally increases with the frequency content of a sound. For different sounds (narrow-band noise (solid line), highpass noise (dashed line), and lowpass noise (dotted line)), the variation of sharpness as a function of frequency is illustrated in Figure 2.5.

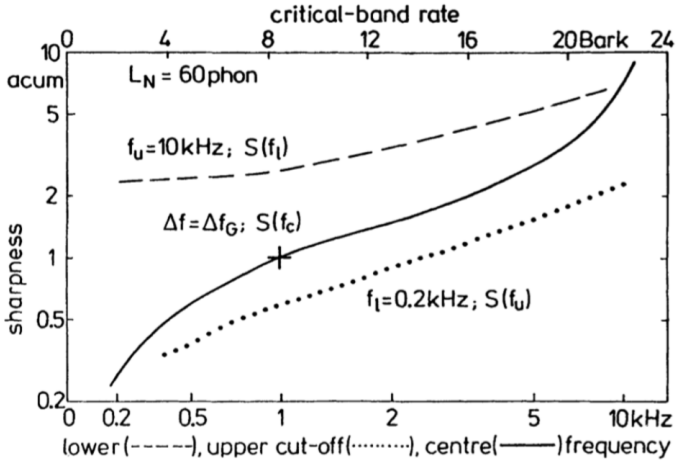


Figure 2.5: Sharpness of different sounds as a function of frequency

2.2.3.2 Tonality

Tonality (T) quantifies the perceptual prominence of tonal versus non-tonal components in a sound. It estimates how clearly tonal elements stand out relative to the overall signal. reference value of 1 *tu* (tonality unit) corresponds to the tonality of a 1 kHz pure tone at 40 dB SPL in a quiet environment [43].

Several methods exist for calculating tonality, but the most recent approach [43]—as defined in the ECMA-418-2 standard—incorporates key psychoacoustic principles, such as the hearing threshold, loudness-dependent masking, and frequency-specific masking effects. This method enables a more objective and perceptually relevant evaluation of audible tonal content [21].

2.2.4 Modulation

When listening to sounds such as those produced when adjusting a seat, the customer typically desires a stable sound that conveys a solid and well-crafted impression. Sounds that fluctuate or exhibit significant temporal variations are generally perceived as more annoying than a stable sound at the same loudness level. Therefore, it is important to analyze these temporal variations carefully. One key step is to distinguish slow variations from fast ones, as they are perceived differently by the human ear [21, 33].

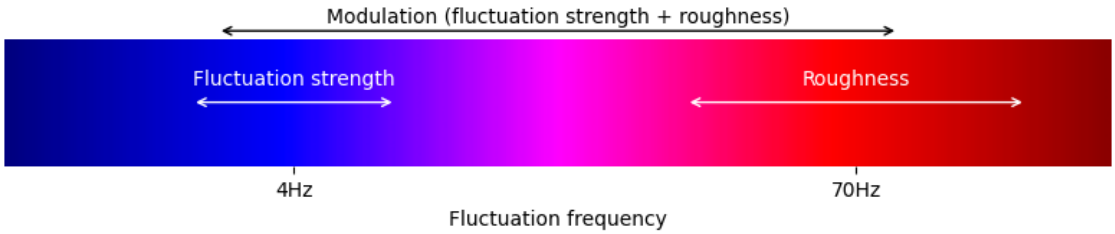


Figure 2.6: Different type of modulation parameters

2.2.4.1 Fluctuation Strength

Fluctuation strength describes how prominently we perceive slow modulations in a sound's amplitude or frequency—typically below 20 Hz. These slow variations occur over time and are directly related to how we perceive rhythm, speech, or musical phrasing [10, 11, 37].

Fluctuation strength (F) is related to the temporal masking depth ΔL, which quantifies the level difference between maxima and minima in the temporal masking pattern. Depending on the speed of the variations, the ear blend more or less the sounds together, which created a temporal masking pattern :

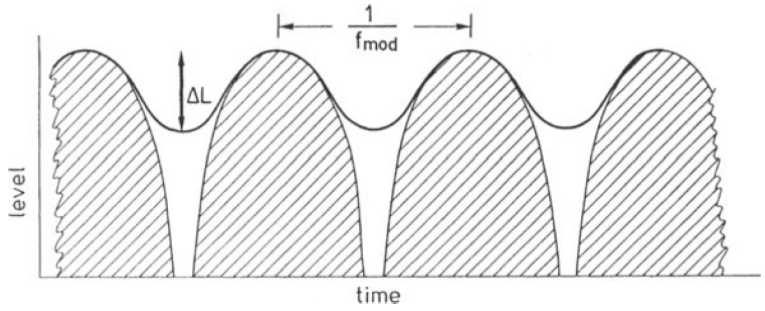


Figure 2.7: Temporal masking pattern of a sinusoidally amplitude-modulated sound

Using the Zwicker method [12], the fluctuation strength F is expressed in *Vacil*. A reference value of 1 Vacil corresponds to a 1 kHz tone at 60 dB SPL with 100% amplitude modulation at 4 Hz. It can be calculated as follows:

$$F \propto \frac{\Delta L}{\left(\frac{f_{\mathrm{mod}}}{4\,\mathrm{Hz}} + \frac{4\,\mathrm{Hz}}{f_{\mathrm{mod}}}\right)} \tag{2.4}$$

2.2.4.2 Roughness

Roughness describes the perception of rapid modulations in a sound's amplitude or frequency—typically between 20 Hz and 300 Hz. These faster fluctuations generate a sensation of harshness or grating in the sound and are strongly related to auditory dissonance and timbre perception. Roughness (R) is calculated in *Asper*, such that

1 Asper corresponds to a 1 kHz tone at 60 dBSPL with 100% amplitude modulation at 70 Hz [12]. An approximation is given in Equation 2.5 :

$$R \propto f_{\text{mod}} \Delta L \tag{2.5}$$

Since ΔL decreases when f_{mod} increases (due to the ear sensitivity), roughness exhibits a bell-like curve :

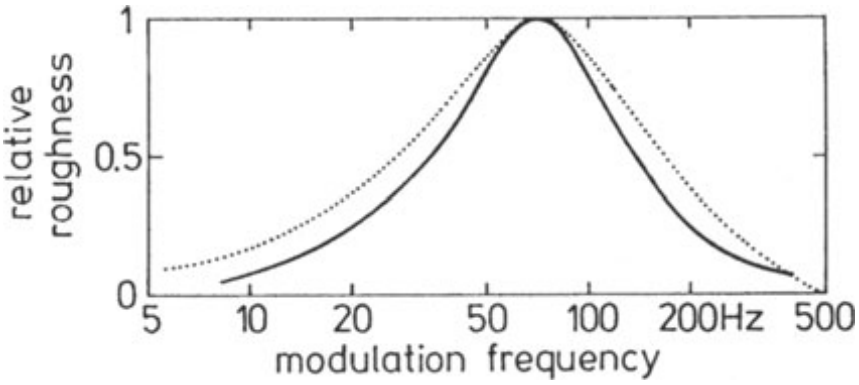


Figure 2.8: Roughness relative to f_{mod}

2.2.4.3 Modulation analysis

The Modulation Analysis (MA) [20] is designed to bridge the gap between roughness and fluctuation strength. By analyzing a broader frequency range and avoiding psychoacoustic weightings, it offers a better representation of temporal signal instability. This makes it particularly relevant for this thesis work.
The core of the method is illustrated in Figure 2.9 :

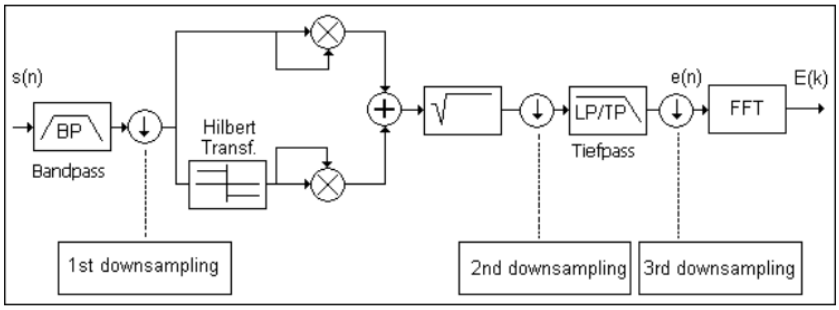


Figure 2.9: Algorithm for the Modulation Analysis [20]

This outputs a modulation spectrum, representing the signal's amplitude fluctuations across modulation frequencies. Unlike traditional psychoacoustic metrics, the modulation here is expressed in dB SPL, offering a direct link to the signal's acoustic energy distribution over modulation frequencies.

2.3 Listening test design

To establish the link between those metrics and perceptual qualities such as annoyance, subjective evaluation through listening tests is indispensable. These tests involve presenting recorded stimuli to human subjects and collecting their impressions or ratings. However, for the data to be valid and interpretable, the test must be carefully designed to eliminate biases and ensure reliability [38].

2.3.1 Test environment

The physical and perceptual context in which the sounds are played has a major influence on listener responses.

- **Realistic context:** The playback setting should recreate, as closely as possible, the original acoustic scene in which the sounds were recorded. For automotive applications, this often involves mimicking a car interior.
- **Proper calibration:** Accurate playback is essential. The system must be calibrated to ensure the sounds are played back at the same levels and spectral balance as in the real environment. Headphone playback is often preferred due to its reproducibility, but it should be supplemented with equalization (e.g., diffuse field or ID equalization).
- **Ambient conditions:** The room should have low background noise controlled temperature and comfortable lighting to minimize listener discomfort.

2.3.2 Participant management

Selecting, preparing, and guiding participants is as important as the technical setup.

- **Sample size and duration:** A balance must be found between the number of subjects and the cognitive load of the test. For simple evaluations, around 40 participants can be sufficient [6], though more may be needed for complicated tests. The overall duration should not exceed 30–45 minutes to avoid fatigue and loss of concentration.
- **Subject accommodation:** Before the actual test begins, participants should be familiarized with the range and nature of the sounds. This can take the form of a practice session with:
  - Representative sounds (e.g., worst, average, best).
  - Randomly selected examples.
  - Graded reference sounds that remain accessible during the test.

  This process reduces response variability and ensures that participants calibrate their internal judgment scales.
- **Clear instructions:** Test objectives, rating scales, and any interface used (computer, paper forms, controllers) should be clearly explained to prevent

misinterpretation. Instructions must be neutral to avoid framing effects or bias.

2.3.3 Test procedure

The way stimuli are presented and evaluated has a direct impact on data consistency and interpretability.

- **Consistency checks:** Some sounds should be repeated throughout the test to detect inconsistent responses or listener fatigue. If a subject provides significantly different ratings for the same sound, their data may be excluded or analyzed separately. The calculation of consistency is based on the average difference in ratings for the repeated sounds. It is described in detail in Section 3.3.1. Using the full set of responses, concordance can also be computed. It shows the degree of variation in ratings between participants. Calculations are further explained in Section 3.3.2.

- **Evaluation methods:** Several methods exist to collect subjective judgments:

  - *Paired comparison:* Subjects are presented with two sounds and asked which is more annoying, more pleasant, etc. This method is intuitive and avoids scale calibration issues.

  - *Semantic differential:* Subjects rate each sound on bipolar adjective scales (e.g., smooth–rough, quiet–loud). This method allows multi-attribute analysis but requires thoughtful design of the descriptor pairs.

  - *Rating scales:* Numerical or labeled scales (e.g., 1 to 10) may be used for global impressions like annoyance or comfort, although they are more sensitive to inter-subject variability.

  It is also possible to assess multiple perceptual attributes within a single listening test. For instance, annoyance is often considered as the primary metric. It can be evaluated using a numerical rating scale, while secondary attributes such as loudness, sharpness, or modulation can be assessed through simpler semantic differential scales.

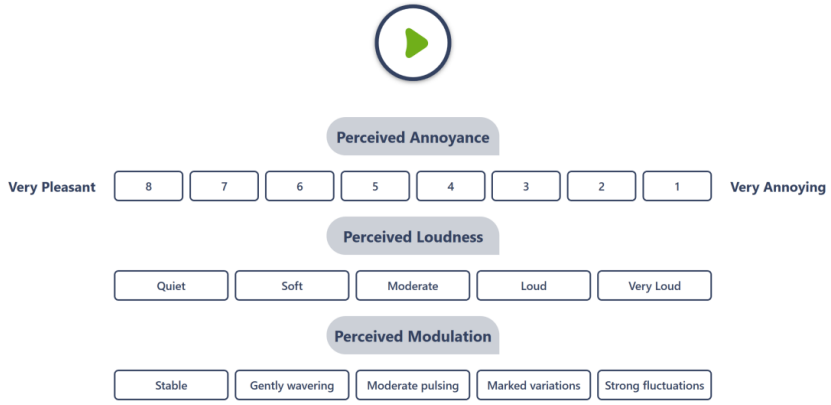


Figure 2.10: Example of a listening test interface using mixed evaluation methods

- **Randomization:** To avoid order effects and anchoring bias, the order of sounds should be randomized across participants.

2.3.4 Unbiasing

In listening tests, it's common practice to remove individual bias from participants' responses [2, 26]. Ideally, if the sound samples are well selected and balanced, each participant's average rating should lie near the midpoint of the scale. For example, on a scale from 1 to 8, the expected mean rating would be $(1 + 8)/2 = 4.5$. However, in the specific context of seat adjustment sounds (typically generated by small electric motors), it's rare for listeners to rate any sound as "not annoying at all." As a result, the responses tend to cluster in the upper half of the scale, reflecting a bias toward more annoying perceptions. In such cases, the lower part of the scale is rarely used, and the average rating for each participant may be consistently above 4.5.

To correct for this perceptual bias, an unbiasing algorithm is applied. This algorithm normalizes each participant's ratings by centering them around their own mean and scaling to unit variance, then re-projects them back onto the original rating scale:

Algorithm 1 Unbiasing algorithm

Require: $\{r_i,\ 1 \leq i \leq n_{\text{jurors}}\}$ ▷ r_i is the vector of all ratings from juror i
Require: $S = [S_{\text{start}}, S_{\text{end}}]$ ▷ The rating scale of the test
  for $i = 1, \ldots, n_{\text{jurors}}$ **do**
    $r_{\text{unbiased}_i} = \dfrac{r_i - \overline{r_i}}{\sigma(r_i)}$ ▷ Center ratings around 0 and scale to unit variance
    $r_{\text{unbiased}_i} = r_{\text{unbiased}_i} \cdot \left(\dfrac{S_{\text{end}} - S_{\text{start}}}{2}\right) + \left(\dfrac{S_{\text{start}} + S_{\text{end}}}{2}\right)$ ▷ Ratings in S
  end for
  return $r_{\text{unbiased}} = \{r_{\text{unbiased}i},\ 1 \leq i \leq n_{\text{jurors}}\}$

2.4 Regression methods

With the objective metrics extracted and the subjective ratings collected, the next step is to do the prediction. Among the various techniques available, regression remains a widely used and intuitive approach [16, 31, 26]. Linear and polynomial regressions offer a straightforward and understandable way to model the relationship between sound quality metrics and the corresponding subjective evaluations. When one variable is dependent of several parameters, this is called multivariate regression analysis (MRA)

2.4.1 Multiple linear Regression

The goal of this method [44] is to approximate the parameter y using all the predictors $x_1, x_2, ..., x_n$ such that :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \tag{2.6}$$

Where β_0 is the constant term called the intercept, and ε corresponds to the error.

The parameters β_i are estimated using the least squares method to minimize the sum of squared errors (SSE) between the predicted values $\hat{y}_i$ and the actual subjective ratings y_i:

$$\text{SSE} = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

where m is the number of observations.

2.4.2 Multiple polynomial regression & interactions

Greater precision can be achieved by including non-linear terms, such as squared predictors:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_{1_n} x_n + \sum_{i=1}^{n} \beta_{2_n} {x_n}^2 + \varepsilon \tag{2.7}$$

Furthermore, interaction terms can be included to account for dependencies between predictors. For example, the interaction between two variables x_1 and x_2 is modeled as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon \tag{2.8}$$

These extensions enhance the model's flexibility and are particularly useful in automotive sound analysis. For example, a low and modulated sound may be perceived as acceptable, whereas a similarly modulated but sharp sound may be judged far more unpleasant. However, such model complexity increases the risk of overfitting, especially in the absence of proper regularization or validation on unseen data. Consequently, techniques such as cross-validation, regularization, and dimensionality reduction are commonly employed to ensure robustness.

In the context of this thesis, the constant term can be problematic and may be omitted to ensure that annoyance is modeled solely as a function of psychoacoustic parameters. Including an intercept could introduce a bias that outweighs the contribution of the actual metrics, leading to a less precise model.

Overall, regression methods remain widely appreciated for their transparency, and simplicity. Those are especially valuable in industrial contexts where interpretability is critical.

2.5 Artificial neural networks

In the context of sound quality prediction, particularly for annoyance estimation, the relationship between psychoacoustic metrics and perceived annoyance can be non-linear and pretty complicated. For instance, a moderate increase in sharpness may have negligible impact on low-loudness sounds but can become highly influential at higher loudness levels. This behavior is trouble for traditional regression, unless specifically modeled through interaction or polynomial terms. However, increasing model complexity in this way raises the risk of overfitting, especially with limited datasets.

To address this, machine learning approaches offer a compelling alternative. By design, neural networks are capable of learning arbitrary non-linear mappings from data, without the need to manually define interaction terms or transformation functions.

The next section introduces neural networks, including their structure, training principles, and advantages in modeling perceptual sound attributes.

2.5.1 Structure

The artificial neuron is the core processing unit in a neural network. It performs a simple yet powerful transformation of inputs into an output signal. As visible in Figure 2.11, it functions in a 3-step process :

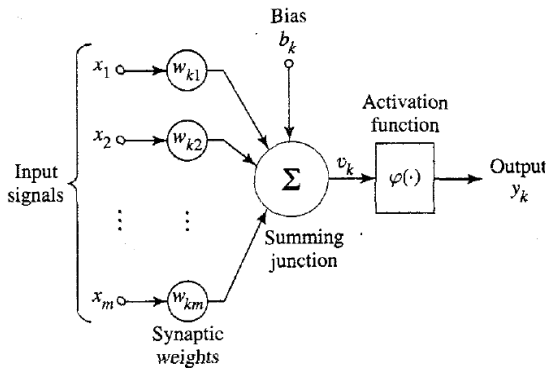


Figure 2.11: Nonlinear model of a neuron [19]

1. **Weights:** Each input x_j is transmitted to the neuron via a connection that carries a weight w_{kj}. As learning progresses, these weights are adjusted back and forth.
2. **Linear combination:** The neuron aggregates its inputs by computing a weighted sum of the incoming signals [29, 19] :

$$v_k = \sum_j w_{kj} x_j + b_k \tag{2.9}$$

   This value reflects the combined influence of all input signals before any non-linearity is applied.
3. **Activation function:** To introduce non-linear behavior and constrain the output range, the net input v_k is passed through an activation function $\phi(\cdot)$.

$$y_k = \phi(v_k) \tag{2.10}$$

   Common activation functions include the sigmoid, tanh, and ReLU, each shaping the output in different ways depending on the application.

By combining these three operations, a neuron transforms a set of input features into an output value, forming the building block of deeper neural architectures, as seen in Figure 2.12.

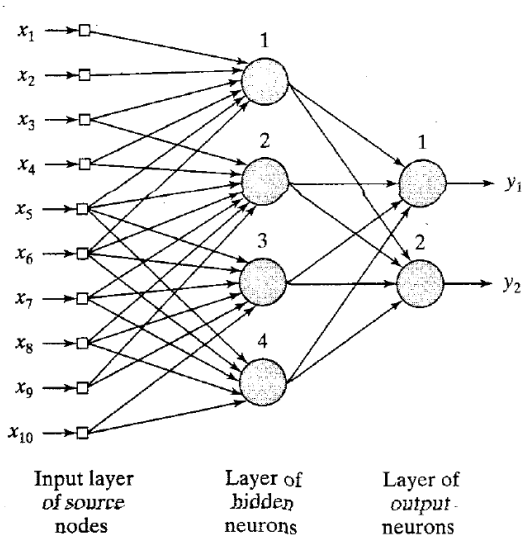


Figure 2.12: Illustration of a neural network with 10 inputs and 2 outputs [19]

2.5.2 Learning process

This project uses supervised learning with backpropagation [29]. The weights and biases are randomly initialized, and the model computes output values from the input data. A loss function (usually MSE or RMSE) is then calculated based on the difference between the predicted and actual values.

Since the model knows all the computations that led to the loss, it can adjust the weights and biases to reduce it. This is done using gradient descent. After many iterations, the loss ideally reaches a minimum, meaning the model has learned from the data.

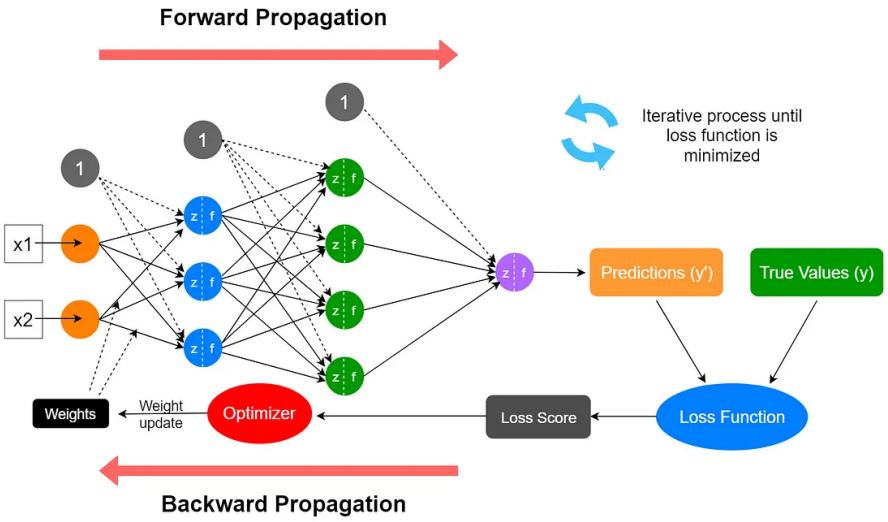


Figure 2.13: Summary of a neural network training process [34]

The backpropagation process requires tuning. The learning rate must be set carefully (too high causes instability, too low slows down training). Other parameters, like the number of hidden layers, neurons, and batch size, also affect performance and must be adjusted based on the task.

2.6 Random forest

One goal of this thesis is to effectively separate sounds of different quality (for example, good, medium, and bad) within a given feature space. Ideally, these sounds can form distinct clusters, and can be categorized just by plotting them in this space.

If such clusters appear when applying regression or neural network models, then the data can be partitioned into groups based on well-defined conditions. In this context, decision trees [32] offer a particularly suitable machine learning approach.

Decision tree methods aim to recursively divide the data into subsets, where each subset ideally corresponds to a specific category of sounds. This approach partitions the sounds based on thresholds (e.g. "if Loudness > 44, rating $= 8$), which means it is easy to interpret.

Mathematically, given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes feature vectors and y_i the corresponding target values, the decision tree iteratively splits the data at each node by selecting a feature j and threshold t to minimize the variance

within the resulting child nodes. Formally, the split partitions the data subset Q_m into

$$Q_{\text{left}} = \{\mathbf{x}_i \mid x_{ij} \leq t\} \quad \text{and} \quad Q_{\text{right}} = \{\mathbf{x}_i \mid x_{ij} > t\},$$

and chooses (j^*, t^*) to minimize

$$\frac{|Q_{\text{left}}|}{|Q_m|} \operatorname{Var}(Q_{\text{left}}) + \frac{|Q_{\text{right}}|}{|Q_m|} \operatorname{Var}(Q_{\text{right}}).$$

This recursive splitting continues until stopping criteria, such as maximum tree depth or minimum samples per leaf, are met. To predict, the model makes the sound go through the tree, and reach its final leaf, where the output is the average of target values of training samples in that leaf:

$$\hat{y}(\mathbf{x}) = \frac{1}{|Q_l|} \sum_{i \in Q_l} y_i.$$

This partitioning enables intuitive visualization and interpretability, as illustrated in Figure 2.14, which shows an example of decision boundaries formed by a decision tree on synthetic 2D data.

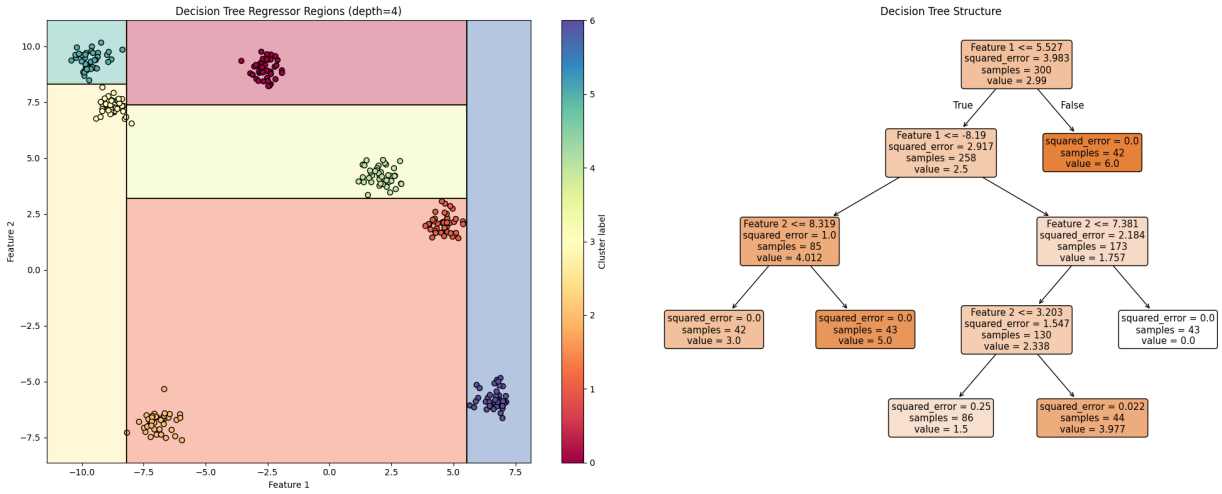


Figure 2.14: Example of decision tree partitions on 2D data, where each rectangle represents a leaf region in the feature space.

Despite their advantages, decision trees can be prone to overfitting and tend to have high variance. However, by creating multiple decision trees, each making its own prediction, the results can be combined through a majority vote. This gives a more stable and accurate model called Random Forests (RF) [5]. With improved generalization and prediction smoothness, RF are a powerful tool for sound quality classification tasks.

2.7 Audio data augmentation

ML models typically perform better with larger and more diverse datasets. However, in many real-world scenarios, such as in this thesis where Volvo Cars has recorded a limited number of vehicle interior sounds, the amount of available audio data is restricted. Furthermore, not all recordings are suitable for training mostly due to inconsistency.

To address this limitation, Data Augmentation (DA) can be employed. DA is a widely adopted technique in image processing, where transformations such as cropping, flipping, rotating, or adding noise help create new training samples from existing ones. These augmentation strategies aim to increase the variability of the training dataset while preserving the semantic meaning of the audio. Although the range of effective augmentation techniques in the audio domain is somewhat narrower than in images [30, 39], several methods have proven useful. In the context of audio classification, data augmentation methods can be broadly categorized into the following:

- **Time-based transformations:** These include time stretching (modifying speed without changing pitch), time shifting (slightly delaying the signal), or cropping.
- **Frequency-based transformations:** Methods like pitch shifting (changing pitch without affecting speed) and equalization alter the spectral content to simulate different recording conditions. An example of this is shown in Fig. 2.15
- **Noise injection:** Adding Gaussian noise, background sounds, or other perturbations can improve robustness to real-world conditions.

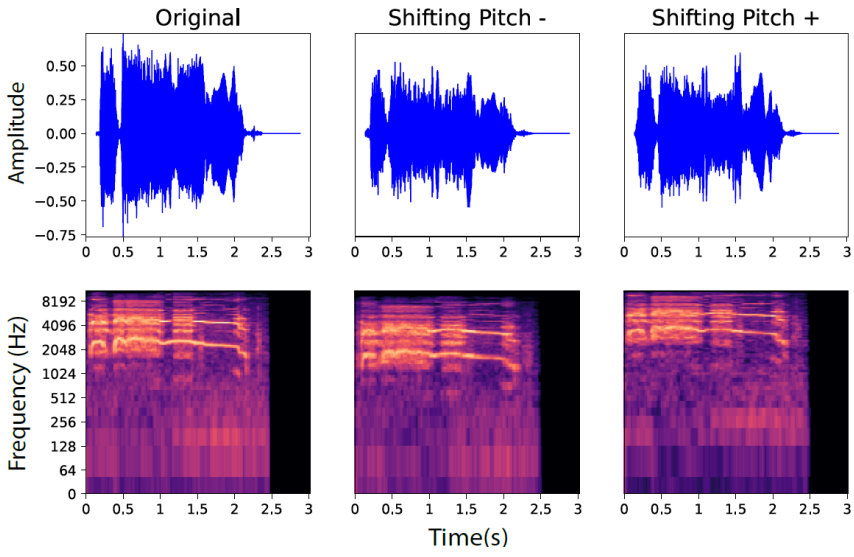


Figure 2.15: Effects of pitch shifting on a rooster crowing

These techniques are typically applied directly to the input sound. However, since the models are designed to predict from psychoacoustic parameters, it may be more appropriate to apply data augmentation directly to these features instead of the audio waveform.

A major challenge in this case is understanding how the augmentation impacts the corresponding label. For example, when background noise is added to a sound, it is unclear whether the perceived annoyance should increase or remain unchanged. Without a clear model of how the transformation affects perception, assigning new labels becomes speculative.

For example, in a linear regression model,if the original psychoacoustic feature vector $\mathbf{x}$ is altered by a perturbation $\boldsymbol{\delta}$, according to Eq. 2.6 the resulting prediction becomes (when ignoring the error term):

$$\begin{aligned} y_{\text{aug}} = \beta_0 + \sum_{i=1}^{n} \beta_i (x_i + \delta_i) &= \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n} \beta_i \delta_i \\ &= f(\mathbf{x}) + f(\boldsymbol{\delta}) - \beta_0 \\ &= y + f(\boldsymbol{\delta}) - \beta_0 \end{aligned} \tag{2.11}$$

Here, f is the linear regression model and $y = f(\mathbf{x})$ is the original prediction. Thus, the new "augmented" rating can be calculated in this case. It can also be calculated for a MPR using Eq. 2.7:

$$y_{\text{aug}} = f(\mathbf{x}) + f(\boldsymbol{\delta}) - \beta_0 + 2 \sum_{i=1}^{n} \beta_{2_i} x_i \delta_i \tag{2.12}$$

When using neural networks,the impact of augmentation on the target label becomes difficult to predict and must be treated with caution. In this thesis, only noise injection will be considered as a data augmentation method. This choice is motivated by the nature of the recorded sounds, which often exhibit strong time variations. Techniques such as cropping or time shifting may remove important temporal features, potentially the sound rating in a unpredictable way. The same goes for frequency based transformations. Noise injection will be explored both at the signal level and at the feature level, to investigate how it affects model robustness and prediction accuracy.

3
Methods

This chapter outlines the methodology used to develop the prediction models, based on the theoretical framework presented in the previous section. It details the process of sound gathering/preprocessing, the computation of psychoacoustic metrics, and the implementation of the predictive models.

3.1 Data gathering and preprocessing

As previously mentioned, this study focuses exclusively on **seat adjustment** sounds. Specifically, it examines the sounds produced by electrically powered vertical seat movements (i.e., moving the seat up and down). Although horizontal movements and backrest adjustment sounds were initially considered, they were excluded after the pilot listening tests due to inconsistent perceptual relevance.

3.1.1 Initial dataset

Volvo Cars has accumulated a large collection of seat adjustment sound recordings. As noted in Section 2.1, recording consistency is crucial for meaningful analysis. Unfortunately, some of the recordings were made under varied acoustic conditions and with different equipment, limiting their comparability. Recordings from 29 vehicles were used in this thesis, and they were recorded in semi-anechoic chambers.

For each vehicle recorded under these controlled conditions, two sound samples were retained: one for upward seat movement and one for downward. This results in a curated dataset of 58 sound samples, spanning from 2011 to 2025.

3.1.2 File structure and protocol compliance

All recordings were conducted according to a standardized protocol designed to ensure consistency. To facilitate post-processing, it is common practice for the recording technician to speak briefly before each movement, announcing actions such as "down" or "up." These verbal cues are visible in the waveform, as shown in Figure 3.1:

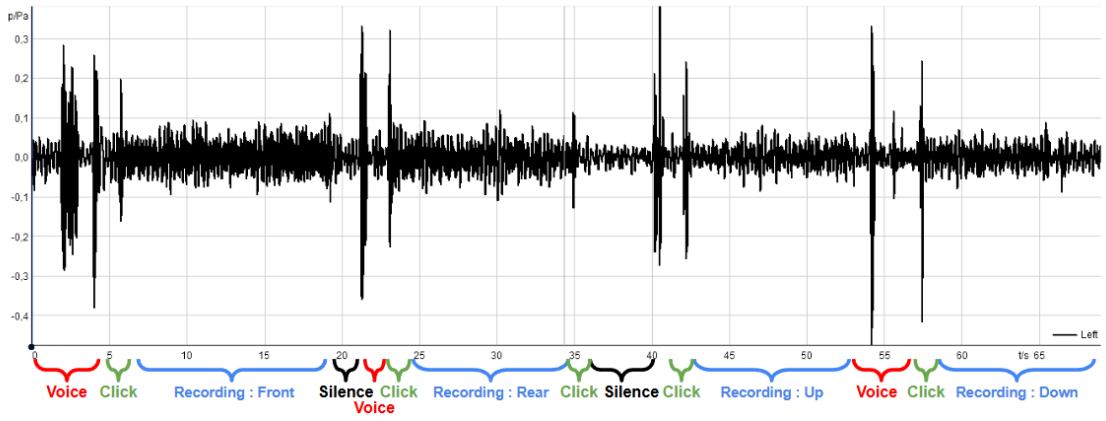


Figure 3.1: Waveform with annotated verbal cues preceding each movement

However, not all recordings follow this practice. Some recordings strictly adhere to the protocol without any verbal annotation. In addition, certain recordings exhibit unwanted artifacts such as extraneous noises or unexpected silences. An example of such a waveform without verbal cues is shown in Figure 3.2:

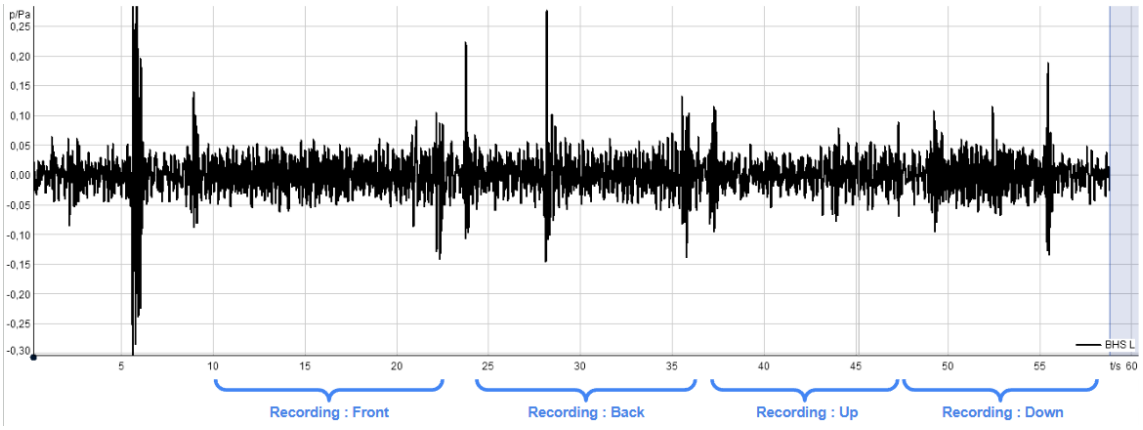


Figure 3.2: Waveform without verbal annotations and with minor artifacts

Here, there is no voice, but some peaks can still be seen at the start of the recording. In this case, this is just the sound of the car being turned on. These kinds of artifacts are present in many files. As explained by Hasselström [18], this inconsistency makes automatic segmentation infeasible. Thus, manual segmentation was performed on all selected files, including the removal of silence and parasitic sounds such as background noise or mechanical clicks. This is done in Artemis Suite using the Mark Editor.

From the original measurements, only sound samples involving "Up" and "Down" seat movements were retained. These directions were chosen over "Forward" and "Backward" due to their greater acoustic variability and clearer perceptual distinctiveness, as observed in pilot listening tests. After preprocessing, all 58 sounds were trimmed to a uniform duration of 6 seconds .They didn't contain any silence or any artifacts.

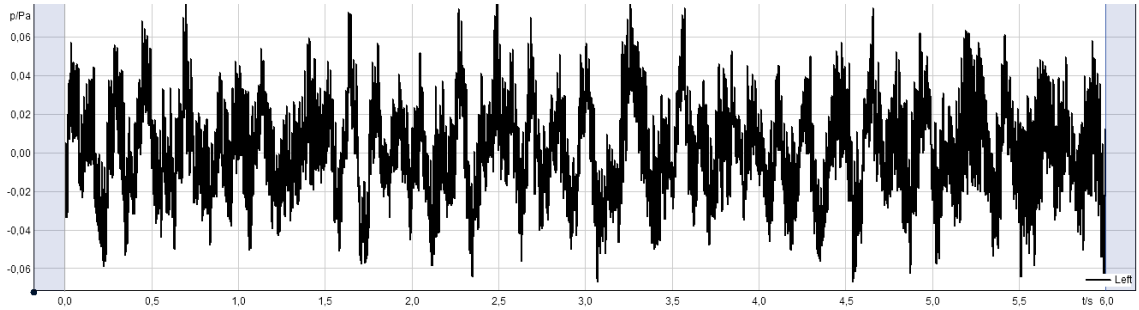


Figure 3.3: Waveform of a sound from the final dataset

Nonetheless, this quantity still exceeded the acceptable limit for a single listening session. Ideally, the test should last less than 30 minutes, so that the listener doesn't get fatigue. Therefore, a reduction strategy was employed, prioritizing psychoacoustic diversity.

3.1.3 Diversity-based reduction

To build a model with the broadest generalization capability, it must be exposed to a wide variety of sounds. This is why the sound selection was based on their psychoacoustic characteristics, with the goal of covering as many perceptually distinct cases as possible.

Key metrics were computed for all candidate sounds, based on previous studies [26, 2, 28]:

- Average Loudness vs. Time – ISO 532-1 (*sone*)
- Average Sharpness vs. Time – Aures (*acum*)
- Average Modulation Spectrum Level – Octave Bands (*dB SPL*)

A 3D plot was generated to visualize the distribution of the candidate sounds in psychoacoustic space (Figure 3.4).

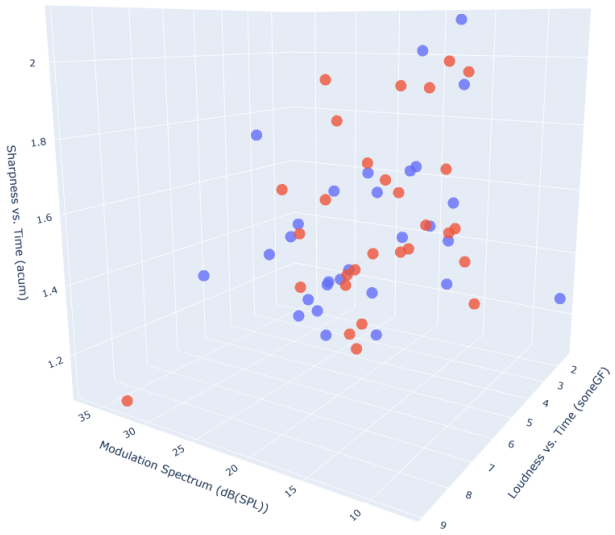


Figure 3.4: Distribution of candidate sounds in 3D psychoacoustic space (Red = Up, Blue = Down)

From this distribution, a manual selection was performed:

- 29 sounds were selected for the listening test (including 3 consistency checkers)
- 3 additional sounds were used for the test introduction
- 1 sound served as the reference "average" sound during the test

This resulted in a final dataset of 33 test sounds:

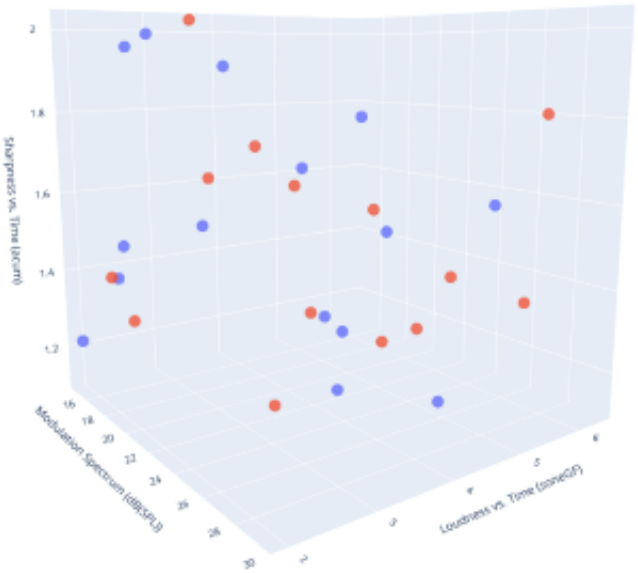


Figure 3.5: Final selected test sounds

The selected samples span a wide range of perceptual profiles—some with high loudness, others with pronounced modulation or sharpness. Although only three key metrics were used for the 3D visualization, additional psychoacoustic parameters were carefully examined to ensure that the final dataset captures the full diversity of subjective experiences encountered across different vehicles and adjustment mechanisms.

Direction	Level (dB(SPL))	Loudness vs. Time (soneGF)	Specific Loudness (soneGF)	Specific Roughness (Hearing Model) (asper)	Modulation Spectrum (dB(SPL))	Sharpness vs. Time (acum)
Down	38,08	1,33	1,11	0,15	7,25	1,26
Up	43,30	3,09	2,20	0,45	15,52	1,48
Down	47,51	2,96	2,62	0,15	15,40	1,57
Up	49,32	2,73	2,57	0,10	16,61	1,67
Down	62,16	5,74	4,85	0,42	25,36	1,50
Up	61,83	5,32	4,82	0,46	25,63	1,50
Down	62,32	3,14	2,86	0,22	14,84	2,11
Up	62,56	3,28	3,05	0,30	17,39	1,92
Down	58,51	2,93	2,81	0,15	14,76	1,93
Up	58,58	3,23	3,03	0,33	21,74	1,63
Down	66,81	5,34	4,92	0,54	23,77	1,27
Up	65,97	5,74	4,88	0,52	19,49	1,25
Down	53,01	3,55	3,24	0,51	22,86	1,65
Up	51,92	4,05	3,69	0,69	26,19	1,57
Down	63,59	4,73	4,32	0,23	26,37	1,27
Up	63,30	3,67	3,41	0,23	16,70	1,52

Figure 3.6: Overview of additional psychoacoustic metrics considered

A conscious effort was also made to include samples from a broad range of vehicles and years.

3.1.4 Backrest Sounds

A parallel selection process was attempted for backrest adjustment sounds, having them rated in a separate listening test to compare with the "up and down" test. However, those were found to include fabric friction noise, making them unsuitable for perceptual testing. Figure 3.7 shows their distribution in psychoacoustic space.

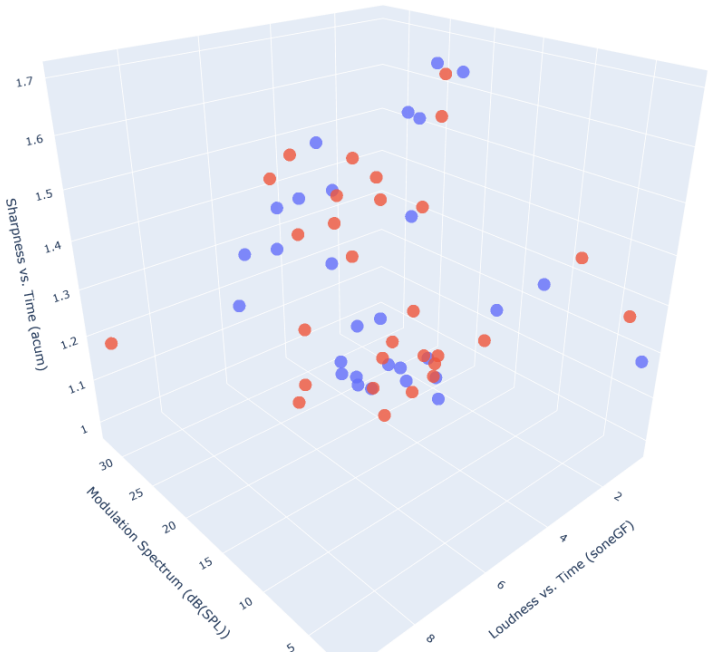


Figure 3.7: Distribution of backrest adjustment sounds (Red = Forward, Blue = Backward)

3.2 Listening experiment

Once all the sounds were gathered, the listening test could begin. The test design was developed in parallel with the sound collection over a period of two weeks. Several pilot tests were conducted to refine the setup before launching the official test. Combining time spent at both Chalmers and Volvo, the entire testing phase lasted five weeks.

3.2.1 Design

To ensure consistency and ease of use, the listening experiment was conducted using the SQala platform developed by *Head Acoustics*. Consequently, all aurally correct recordings were readily available in the ".hdf" format, and staying within the same ecosystem significantly streamlined the integration process.

In addition to assessing overall annoyance, it was deemed valuable to include ratings of key psychoacoustic attributes to explore potential correlations. Given that some participants were non-experts in acoustics, the chosen metrics had to be easy to understand and to interpret visually. Based on previous studies [2, 31, 38, 9], loudness, sharpness, and modulation (either roughness or FS, depending on context) have been shown to significantly influence user comfort. Thus, they were selected alongside annoyance for evaluation.

Annoyance was rated on an 8-point categorical scale to enable finer differentiation

between levels of discomfort. The remaining three attributes were evaluated using 5-point semantic differential scales, capturing participants qualitative impressions. Using sematic differential scale has been shown to be more effective when rating such sensations. [38, 27]

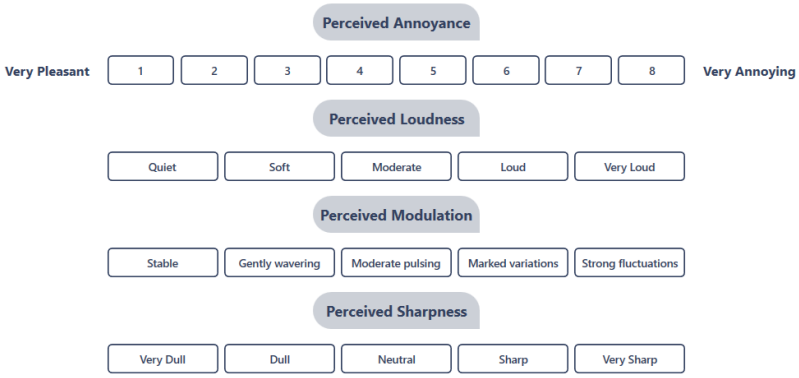


Figure 3.8: User interface of the listening test

The listening test followed this structure:

1. **Introduction**
   To ensure consistency across participants and to align their understanding of the task, the experiment began with a brief introduction. It explains what the participants might hear and which characteristics could contribute to perceived annoyance.

   "Keep in mind that annoyance can stem from various factors.
   Please pay particular attention to the following aspects:

   - Unstable motor speed / modulation: This may sound as if the motor is struggling to move the backrest smoothly.
   - Creaking or clicking noises: These are often due to friction or mechanical interactions.
   - Sharp and loud noises: Sudden, harsh sounds that can be jarring to the ears.
   - Rumbling noises: Low-frequency, rough sounds that may be especially unpleasant in a vehicle setting.

   These are just examples—ultimately, only you can decide what truly annoys you."

2. **Training phase**
   Participants were presented with three example sounds to familiarize themselves with the rating scale: one rated as “good” (score 1–2 by the Volvo team),

two “medium” (scores 3–5), and one “bad” (score 7–8). During this phase, listeners could replay the sounds as many times as needed to get accustomed to the scale.

3. **Demographic questionnaire**
   Participants completed a short demographic survey to support deeper analysis of the results. This allowed comparisons between different groups—e.g., Volvo vs. non-Volvo employees, NVH experts vs. non-experts, and how age might influence annoyance perception.

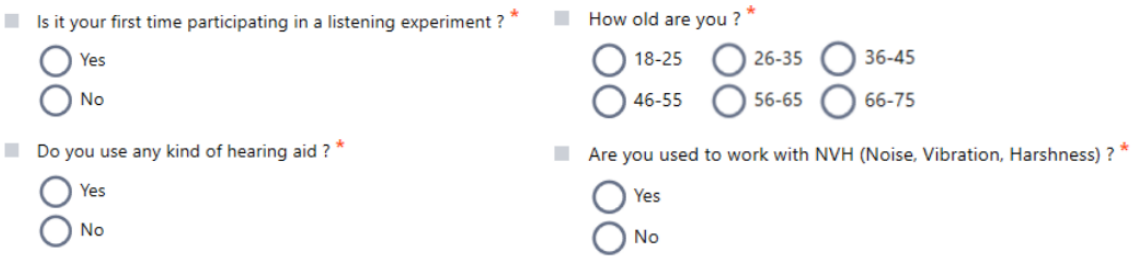


Figure 3.9: Demographic questionnaire

4. **Main test**
   The main test consisted of five batches, each containing 6–7 sound samples. To lower listener fatigue, a short break was provided between batches. Each of them was preceded by a reference sound to help recalibrating. The inclusion of a single reference per batch (rather than a persistent comparison option) was forced by SQala, which do not support continuous access to a reference sound.

   The full test duration was kept under 20 minutes to maintain participant focus and comfort. Counting the introduction and feedback time, as well a some discussion time after the experiment, this amounts to around 30 minutes.

5. **Feedback**
   At the end of the experiment, participants were invited to share open-ended feedback. This helped gather insights for improving future possible tests. It also gave ideas to better interpret or nuance the results of the tests.

3.2.2 Environment

Participants were recruited from various Volvo departments and from Chalmers University, with the aim of gathering a diverse pool of listeners. The goal was to balance the sample, ideally achieving a 50/50 split between experts and non-experts.

At Volvo, the experiment was conducted in a dedicated listening room. Car seats were mounted on wooden pallets and positioned sideways to simulate the enclosed environment of a vehicle. This arrangement also allowed two participants to take part simultaneously, which encouraged participation—people were more likely to

join if they could come in pairs.

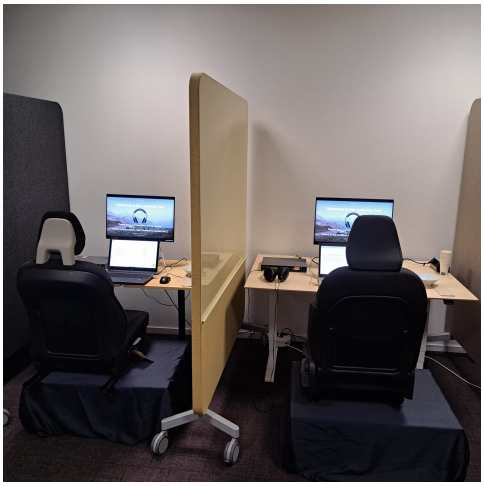


Figure 3.10: Listening test setup at Volvo: car seats mounted on wooden pallets in a side-facing configuration.

At Chalmers, a simpler setup was used, as transporting car seats was impractical. This setup provided an opportunity to assess whether the testing environment significantly influenced participants' responses.

Figure 3.11: Listening test setup at Chalmers: simplified environment without car seats.

In both environments, the playback equipment was identical:

Audio	**Playback**
ID (Independent of Direction) equalization performed using a PEQ V equalizer from *Head Acoustics*.	Open-back headphones (Sennheiser HD600) were used to reproduce a spatial auditory experience, similar to that of a car interior.

Table 3.1: Playback configuration

The two rooms used for the experiment were well-isolated: the listening studio at Volvo and the audio lab at the Chalmers Acoustic Department. The background noise level in both rooms was approximately 25 dB(A). Since all test sounds were played at levels above 42 dB(A), this resulted in a signal-to-noise ratio (SNR) of around 17 dB(A), which is sufficient for this kind of perceptual evaluation tasks.

3.2.3 Organization

Several pilot sessions were conducted with the project team, master's students at Volvo, and user experience (UX) experts to evaluate and refine the listening test procedure. These sessions were highly valuable, as they helped to resolve multiple issues, such as :

- Removing backrest sounds
- Standardizing the duration of all audio samples
- Using real car seats to provide a more realistic test environment.
- Improving the initial briefing
- Introducing small gestures like offering cookies and water for participant comfort

Figure 3.12: Listening test invitation

Once the official experiment began, the procedure was kept consistent to ensure the reliability of the results. While the oral introduction at the beginning of the test was slightly adjusted over the course of the five weeks, all other elements remained unchanged. Many participants came in pairs, which encouraged discussions about the experiment and this thesis work.

Once the experiment was completed, a summary of the methodology and results was shared with the Volvo team through an interactive Power BI report.

3.3 Analysis of the test results

Using the demographic questionnaire (Section 3.9), the results of the listening test can be sorted by age group, experience with listening experiments, and expertise in NVH. Gender differences were not considered in the analysis, as several previous studies have shown that gender has no significant impact on auditory perception in similar contexts [2, 26, 9]. To compare groups, $RMSE$ and Pearson's correlation coefficient are used, and displayed through matrices.

To assess participants' ability to provide consistent and accurate responses, two specific metrics were used.

3.3.1 Consistency

To be considered "valid" for the experiment, each listener must demonstrate consistent responses. This means that when presented with the same sound multiple times, the listener should provide similar ratings each time. Some participants may be unfocused or misunderstand the task, leading to unreliable answers. Measuring consistency helps ensure that the collected data truly reflects perceptual judgments rather than random guesses or confusion.

In the listening test, 3 sounds were each repeated 3 times, in different batches. The consistency per participant (also called Singular Consistency SC_s) is calculated such that, for the i^{th} juror :

Algorithm 2 Calculation of Singular Consistency

Require: $\{r^c_{i,j},\ 1 \leq j \leq 3\}$ ▷ $r^c_{i,j}$ is all 3 ratings from juror i on sound j
Require: $S = [S_{\text{start}}, S_{\text{end}}]$ ▷ The rating scale of the test
  $SC_s \leftarrow 0$
  for $j = 1, \ldots, 3$ **do**
    $$SC_s = SC_s + \left(1 - \frac{\sigma(r^c_{i,j})}{S_{\text{end}} - S_{\text{start}}}\right)$$
  end for
  $$SC_s \leftarrow \frac{SC_s}{3}$$
  return SC_s

This consistency score provides a simple and interpretable measure of agreement between participants for each sound. A value close to 1 indicates high agreement, while lower values show poor consensus. As an example :

- **Sound 1:** Ratings = 4, 3, 3
  Standard Deviation: $\sigma = 0.58$
  Consistency: $1 - \frac{0.58}{7} = 0.917 \approx 91.7\%$
- **Sound 2:** Ratings = 6, 5, 4
  Standard Deviation: $\sigma = 1.00$

Consistency: $1 - \frac{1.00}{7} = 0.857 \approx 85.7\%$

- **Sound 3:** Ratings = 4, 4, 6
  Standard Deviation: $\sigma = 1.16$
  Consistency: $1 - \frac{1.16}{7} = 0.834 \approx 83.4\%$
- $SC_s = \frac{0.917+0.857+0.834}{3} = 0.869 \approx 87\%$

3.3.2 Concordance

Another way to evaluate juror performance is by assessing how much a participant's rating for a given sound deviates from the overall average rating provided by all participants for that sound. This metric is known as concordance. It reflects the degree to which an individual's perception aligns with the group consensus. By averaging this deviation across all test sounds, the Singular Concordance (SC_d) of a participant is obtained. A SC_d value close to 1 indicates closer agreement with the group, while a value closer to 0 suggests greater deviation from the collective ratings.

The concordance for a given sound s is computed as follows [2]:

- If the participant's rating is within ± 1 unit of the mean, the concordance is mapped linearly between 1 (perfect agreement) and 0:

$$C_{i,s} = 1 - \frac{|\bar{r}_s - r_{i,s}|}{\bar{r}_s}$$

- If the deviation is greater than 1 unit, the concordance is set to 0:

$$C_{i,s} = 0$$

Where :

- $r_{i,s}$ is the rating given by participant i for the sound s,
- $\bar{r}_s$ is the mean rating for the sound s across all participants,
- $C_{i,s}$ is the concordance score of participant i for the sound s.

As an example for one sound :

- Average rating for the sound: $\bar{r}_s = 5.1$
- Participant's rating: $r_{i,s} = 6$
- Concordance: $C_{i,s} = 1 - \frac{|6-5.1|}{5.1} = 0.82$

Since the test consists of 26 different sounds, each participant receives 26 concordance values. The overall SC_d for participant i is then defined as the average of their individual concordance values:

$$SC_d(i) = \frac{1}{26} \sum_{s=1}^{26} C_{i,s} \tag{3.1}$$

3.3.3 Hierarchical Clustering

To investigate potential clustering among participants based on their responses, Ward's hierarchical clustering was used [45]. It identifies groups of participants who exhibit similar listening response patterns, and dendrogram trees are used for visualization.

Ward's method is a variance-minimizing clustering technique. At each iteration, the algorithm combines the two clusters that result in the smallest increase in total within-cluster variance. Each participant is represented as a 26×4-dimensional vector corresponding to their 4 ratings across all audio stimuli.

In a dendrogram, the vertical axis shows the linkage distance, i.e., the dissimilarity between merged clusters. The horizontal axis lists the individual participants. By marking specific groups (e.g., NVH experts) with a symbol on this axis, one can quickly check whether the clustering results are well distributed or not. Figure 3.13 shows an example of such a tree :

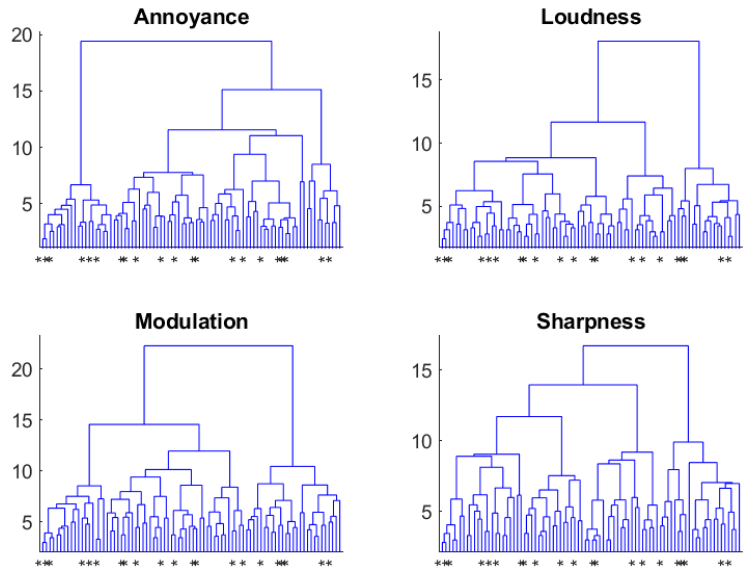


Figure 3.13: Example of dendrogram trees without clusters

3.3.4 Metrics ratings

For all four metrics evaluated in the listening test, the results are presented using box plots. Each metric is analyzed separately, and for each sound, both the arithmetic mean and the geometric mean of participant ratings are displayed. To improve the visual clarity of the box plots, a small amount of random noise (approximately 1%) is added to the data. This slight perturbation helps avoid issues such as overlapping points or compressed whiskers, which can occur when ratings are limited to a narrow discrete scale (e.g., from 1 to 5). This noise is applied solely for visualization purposes and is not used in any further analysis.

In addition to the individual metric analysis, all test metrics are compared with one another using a correlation matrix. This allows for identifying potential relationships between those parameters.

3.4 Feature Selection

Using the metrics from the test, the optimal methods from different standards can be selected. This selection process employs correlation analysis to identify the most effective methods. The following calculation methods are compared with the test results :

Loudness	**Sharpness**	
ISO 532-3 (Cubic Average)	Sharpness | DIN45631; DIN 45692 (S5)	Tonality | ECMA 418-2 (T5)
ISO 532-3 (Max)	Sharpness | DIN 45631; DIN 45692 (Cubic Average)	Tonality | ECMA 418-2 (T10)
ISO 532-3 (N5)	Sharpness | DIN 45631; DIN 45692 (S10)	
ISO 532-3 (N10)	Sharpness | ISO 532-3; Aures (S5)	
ISO 532-1 (N5)	Sharpness | ISO 532-3; Aures (Cubic Average)	
ISO 532-1 (Cubic Average)	Sharpness | ISO 532-3; Aures (S10)	
ISO 532-1 (Max)	Sharpness | ISO 532-1; Aures (S5)	
ISO 532-1 (N10)	Sharpness | ISO 532-1; Aures (S10)	

Table 3.2: Loudness and sharpness related metrics

Modulation		
Modulation Spectrum | Octave 1k 710-1.4k Hz (Average)	Fluctuation Strength (F5)	Roughness | ECMA 418-2 (3rd) (R5)
Modulation Spectrum | Octave 1k 710-1.4k Hz (Quad Average)	Fluctuation Strength (Max)	
Modulation Spectrum | Octave 1k 710-1.4k Hz (Min)		
Modulation Spectrum | Octave 1k 710-1.4k Hz (Max)		
Modulation Spectrum | Octave 1k 710-1.4k Hz (L5)		
Modulation Spectrum | Octave 1k 710-1.4k Hz (L10)		
Modulation Spectrum | 1/3 Octave 315 280-355 Hz (Quad Average)		
Modulation Spectrum | 1/3 Octave 100 90-112 Hz (Quad Average)		
Modulation Spectrum | Full bandwidth (Quad Average)		

Table 3.3: Modulation related metrics

The calculations can also be combined (Sharpness x Tonality for example) if necessary to better correlate with the results from the listening test.

3.5 Predictions & Comparison

Once the methods are selected, the additional metrics from the listening test are no longer required. The training dataset includes 26 annoyance ratings, each paired with the corresponding objective metrics.

The test dataset consists of 7 data points: the 3 consistency sounds and 4 additional sounds rated by the Volvo team. Those ratings were not obtained through the formal listening test and therefore may lack precision. However, it still provides a useful indication, as the estimated ratings are roughly consistent with expectations.

3.5.1 Performance evaluation

The performance of each model is evaluated visually through plots, and quantitatively by calculating the RMSE for both the training and test datasets. This evaluation is done using both the scale from the listening test and the original Volvo scale 1.1.

3.5.1.1 Experiment scale

The models are first evaluated using the original scale from the listening experiment. The initial 1–8 annoyance scale is mapped to a 1–5 scale to be consistent with the other metrics, as it simplifies computations. Consequently, all resulting plots display annoyance ratings within the 1–5 range, and the model predictions are made accordingly.

As a first step in evaluating model performance, a reference plot is used: both the actual and predicted ratings are displayed on the same graph. Ideally, perfectly predicted points align along the diagonal midline. An example is shown in Fig. 3.14.

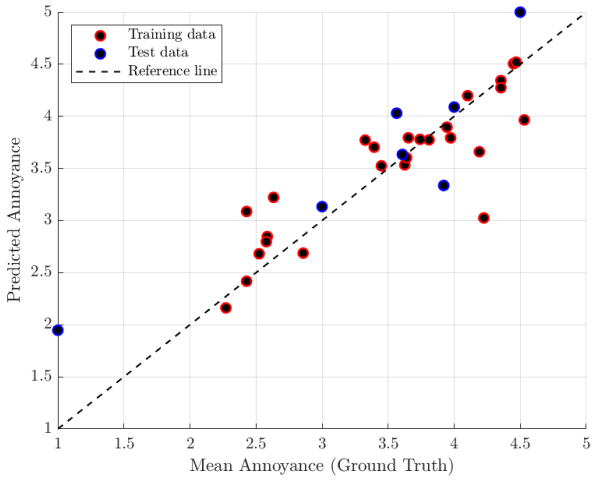


Figure 3.14: Example of a reference plot comparing predicted and original ratings

To complement this, the predicted annoyance values are overlaid on the original box plots from the listening test. This allows a visual check of whether the predicted values fall within the interquartile range of the experimental ratings.

In addition to the interquartile range comparison, it is also valuable to check whether the original mean annoyance ratings and their confidence intervals lie within the model's prediction confidence intervals. When this is the case, it indicates a higher degree of model robustness. An example is shown in Fig. 3.15.

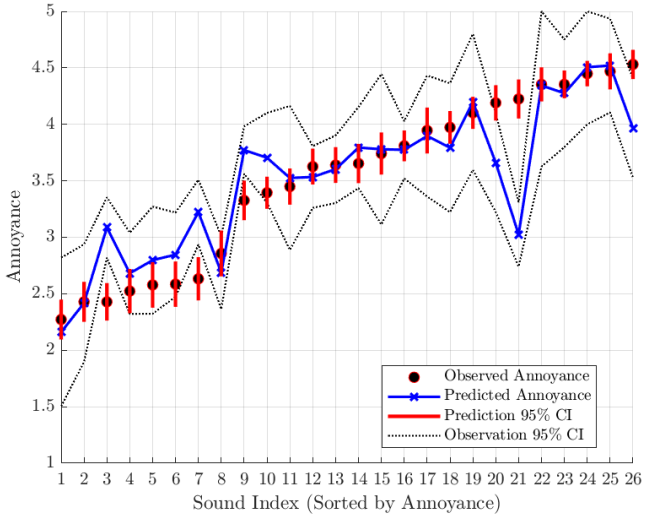


Figure 3.15: Comparison between prediction CI and original rating CI

The RMSE calculation is made using this 1-5 scale.

3.5.1.2 Volvo Scale

While the primary objective of this thesis is to assess annoyance using objective metrics, it is more practical for Volvo if the predictions are expressed in their internal rating scale. Therefore, the predicted annoyance values are mapped back to Volvo's 1-8 scal. Recordings used in this thesis are rated between 5 to 8 in Volvo's scale..

However, there is a fundamental difference between the two scales: in the experimental scale, 1 represents low annoyance and 5 high annoyance, whereas in the Volvo scale, 5 is the worst (highest annoyance) and 8 is the best (lowest annoyance). As such, the scale must be inverted during the transformation.
This inversion and rescaling is performed using the following formula, applied to each predicted rating $\hat{r}_i$:

$$\hat{r}_{i,\text{volvo}} = 5 + 3\left(1 - \frac{\hat{r}_i - \min(\hat{r})}{\max(\hat{r}) - \min(\hat{r})}\right) \tag{3.2}$$

Here, $\hat{r}$ is the vector of predicted annoyance ratings in the 1-5 experimental scale, and $\hat{r}_{i,\text{volvo}}$ is the prediction of the i^{th} converted to the Volvo 5-8 scale.

The reasoning behind this formula is:

1. $\hat{r}_i - \min(\hat{r})$ shifts the predictions so that the minimum value becomes 0, resulting in values within the range $[0, \max(\hat{r}) - \min(\hat{r})]$.
2. Dividing by $\max(\hat{r}) - \min(\hat{r})$ normalizes the values to the $[0, 1]$ range.
3. Subtracting this quantity from 1 inverts the scale: higher annoyance ratings become closer to 0 (better in the Volvo sense), while lower annoyance ratings

become closer to 1 (worse).

4. Multiplying by 3 and adding 5 maps the values from the normalized $[0, 1]$ scale to Volvo's $[5, 8]$ scale.

Equation 3.2 accounts for the fact that the entire experimental scale is not used in practice. It ensures that the worst predicted annoyance rating in the experiment maps to 8 in the Volvo scale, and the best maps to 5, thereby fully utilizing the available 5-8 range.

For this reason, unbiasing the ratings will have limited impact in this study. While unbiasing can help compensate for subjects who consistently rate lower or higher than others, in this case, most participants tend to use the high end of the annoyance scale. As a result, unbiasing mainly shifts the ratings downward uniformly, without changing too much the relative distribution of predictions. Thus, it does not significantly affect the final Volvo-scaled results. This observation is confirmed and discussed in more detail in Chapter 4.

Those Volvo scaled ratings are compared with ratings from the Component NVH team. As shown in Fig. 3.16, the predicted ratings follow the same trend as those expert ratings.

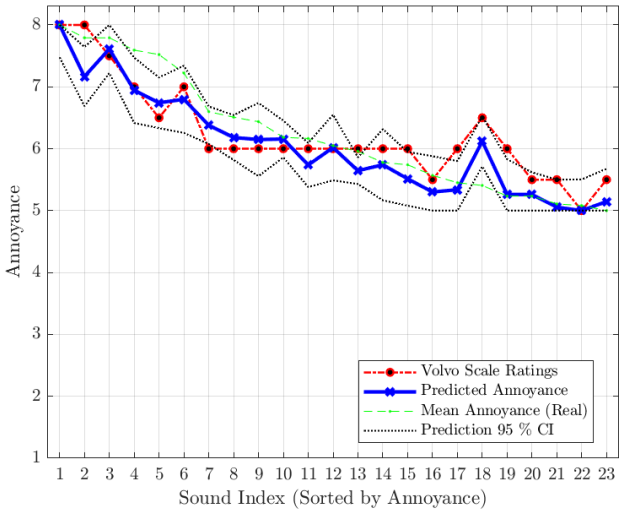


Figure 3.16: Example of ratings comparison in the Volvo scale

Finally, the sounds are visualized in a 3D space defined by the interaction between selected objective metrics. Each point in this space represents a sound and is associated with its corresponding rating. To facilitate interpretation, the points are grouped into three categories :

- **Uncomfortable:** ratings from 5 to 6
- **Acceptable:** ratings from 6 to 7
- **Comfortable:** ratings from 7 to 8

The title page cover shows an example of this clustering.

3.5.2 Linear & polynomial regression

Both regressions are implemented in `Matlab` using stepwise regression. The intercept is fixed at zero, and interactions between metrics are allowed. Unlike machine learning models, these regression approaches can provide confidence intervals for their predictions, which is particularly useful, as illustrated in Fig. 3.15.

Linear regression is a well-established method and is often found to perform well in predicting annoyance [17, 15]. Therefore, it is expected to yield reasonable training and test RMSE values.

Polynomial regression, on the other hand, is more prone to overfitting, especially when interaction terms are included and the dataset is limited. To mitigate this, the resulting polynomial model is simplified by removing certain terms, which improves generalization. This leads to a lower training RMSE than before, but the test RMSE increases a lot.

3.5.3 Machine learning models

The simple neural network is implemented in `Matlab`, while the random forest model is developed in `Python`.

3.5.3.1 Simple neural network

The neural network is intentionally kept simple, consisting of two hidden layers with i and j neurons respectively, and a single output representing the predicted annoyance. The number of inputs corresponds to the number of objective metrics used to assess annoyance, which is found to be 5 in this study.

To determine the optimal number of neurons in each hidden layer, a grid search was performed. For each possible combination (from 1 to 10 neurons in each layer), the model was trained and tested. The selection criterion was the sum of the training RMSE, test RMSE,and the difference of the two. This helps to have a more balanced model and ignore overfitting models.
The following `Matlab` code snippet shows the implementation of the grid search and training procedure :

```
max_neurons_1 = 10; max_neurons_2 = 10;

train_rmse = zeros(max_neurons_1, max_neurons_2);
test_rmse = train_rmsezeros(max_neurons_1, max_neurons_2);

best_total_rmse = inf; best_model = [];

for i = 1:max_neurons_1
    for j = 1:max_neurons_2
        net = fitnet([i, j]);
        net.trainParam.showWindow = false;
```

```
12
13          [net_trained, ~] = train(net, train_data, annoyance_train);
14
15          pred_train = net_trained(train_data);
16          pred_test = net_trained(test_data);
17
18          train_rmse(i, j) = rmse(annoyance_train, pred_train);
19          test_rmse(i, j) = rmse(annoyance_test, pred_test);
20
21          diff = abs(train_rmse(i, j) - test_rmse(i, j));
22          total_rmse = train_rmse(i, j) + test_rmse(i, j) + 0.8*diff;
23
24          if total_rmse < best_total_rmse
25              best_total_rmse = total_rmse; best_model = net_trained;
26              best_i = i; best_j = j;
27          end
28      end
29  end
```

Listing 3.1: Grid search for optimal neural network architecture

3.5.3.2 Random forest

The random forest model has two main hyperparameters that must be tuned: the number of trees and the maximum tree depth. Increasing the number of trees generally improves the model's performance by reducing variance and increasing stability. However, beyond a certain point, adding more trees leads to overfitting and can lead to unnecessarily long computation times.

Similarly, increasing the maximum depth of each tree allows the model to capture more complex patterns in the training data. However, if the trees are too deep, the model can become overly specific to the training data and overfits. To identify the optimal combination of these parameters, a grid search is performed. Due to the relatively small dataset and model size, the computational cost remains reasonable.

As part of the analysis, one of the individual decision trees within the trained random forest is visualized to illustrate the decision-making process. In the tree, the root node represents the most important decision criterion influencing the annoyance prediction. The following splits reflect less dominant, but still relevant, features. This is very useful to interpret which objective metrics are prioritized by the model.

3.6 Data augmentation

Two types of data augmentation were investigated: adding noise directly to the objective metrics, and adding noise to the sounds themselves.

For the first method, noise was introduced to the metric values by computing the range of each metric across the dataset and adding a noise amplitude corresponding to one tenth of that range. This was applied to all the 29 sounds rated in the test. For these altered data points, the corresponding annoyance ratings were increased by 10% to simulate the expected effect of the degraded sound quality.

For the second method, noise was added directly to the audio signals using Artemis Suite. This process was performed manually: each sound was carefully listened to, and a level of noise was selected such that it slightly degraded the perceived quality. Doing this ensures that the change would not significantly alter a juror's rating during a listening test.

Overall, both augmentation strategies had only a minor impact on the model performance. They did not lead to any noticeable improvement. In the best case, it simply increased the number of data points in the plots, with the original rating distribution remaining largely unchanged. In the worst case, it led to overfitting.

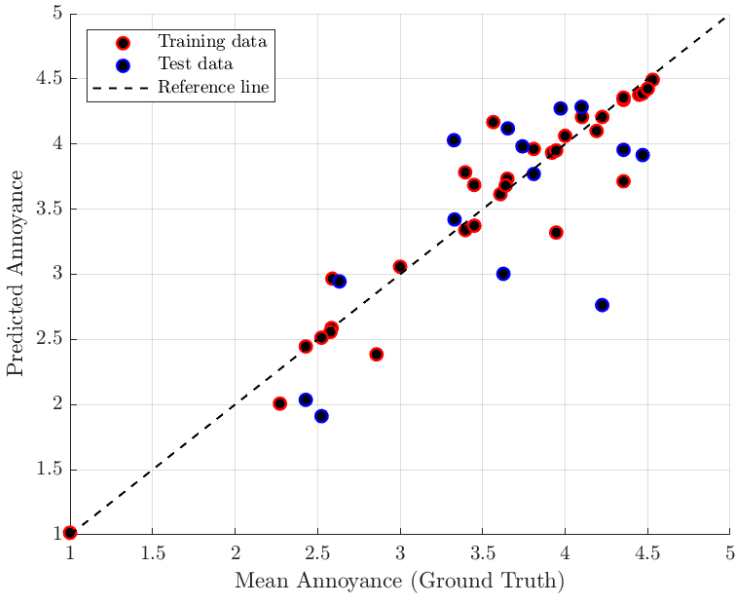


Figure 3.17: Exemple of a reference plot after data augmentation

4
Results

This chapter presents the results of the listening experiment and evaluates the performance of the different models based on this dataset.

4.1 Jury Testing

The results of the jury testing are presented in this section. It begins with a summary of the participants, followed by an overview of the overall ratings (including the original bias). Subsequently, comparisons between different groups are made, and finally, the effects of unbiasing are discussed.

4.1.1 Participant summary

After five weeks of running the experiment, a total of 85 participants took part. One participant was wearing a hearing aid and was therefore considered an outlier and excluded from the analysis. This results in a final dataset of 84 participants: 74 from Volvo Cars and 10 from Chalmers. Among the participants, 57% had previously taken part in a listening experiment, and exactly half of them were familiar with working with NVH.

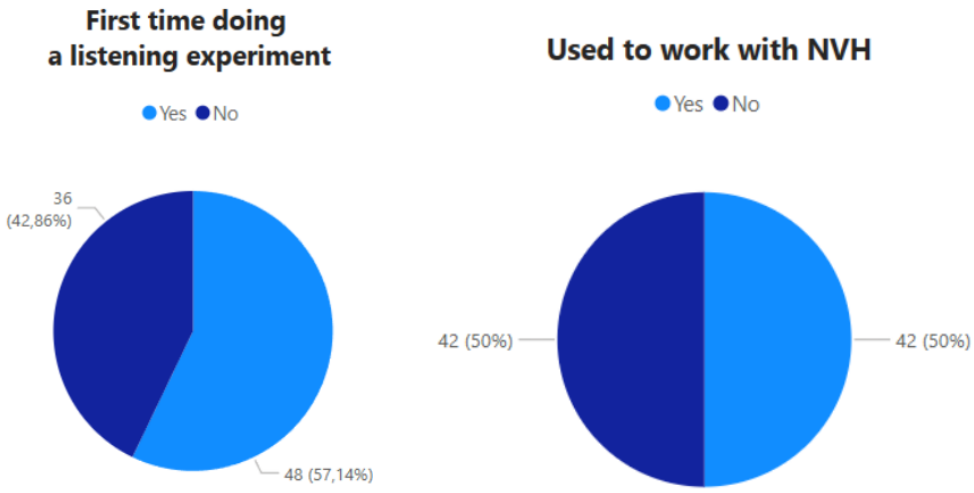


Figure 4.1: Distribution of participant with their acoustic background

Several age groups were represented in the panel, with a fairly balanced distribution: approximately 50% of the participants were under 45 years old. The 56-65 age group was under-represented compared. This limited representation leads to weaker correlation results for this age group in the next group comparisons.

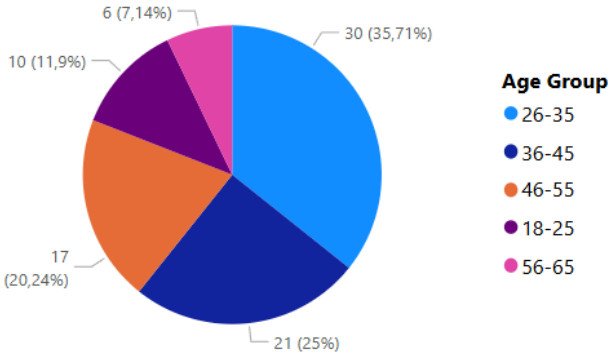


Figure 4.2: Distribution of participants across age groups

4.1.2 Consistency vs. Concordance

The singular consistency and concordance were calculated for each participant and plotted against each other for all four metrics of the experiment. As a rule of thumb, any participant located in the top-right corner for all four metrics is considered valid for the test.

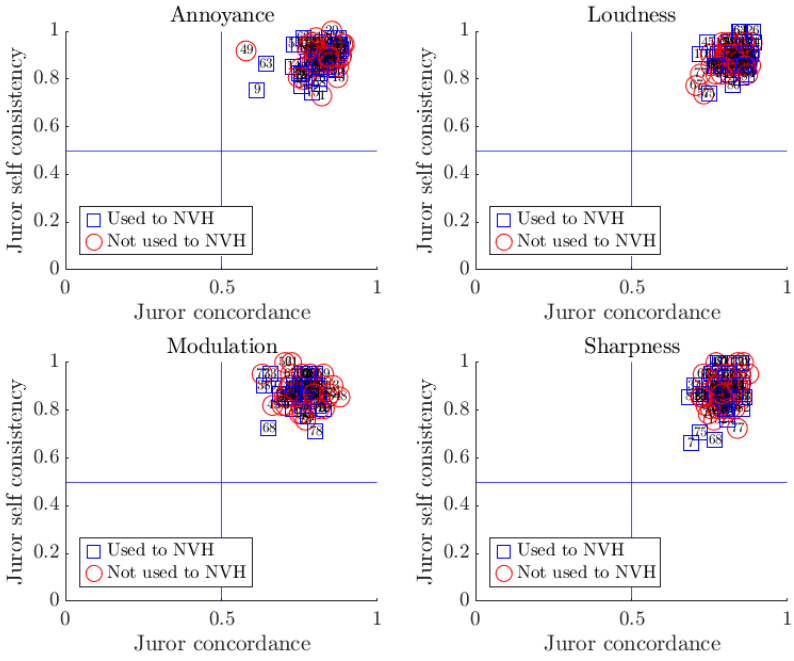


Figure 4.3: Consistency vs. Concordance for all 4 metrics

Even though some participants showed lower concordance in their annoyance ratings, these values are diluted by the overall number of participants, and the overall consistency remains high. Tests were conducted by removing these lower-concordance cases, and the results showed no change.

As expected, loudness showed the most consistent results, as it is the easiest metric to perceive and evaluate. In contrast, modulation and sharpness are more difficult to interpret, which likely explains the greater dispersion in the ratings. Moreover, there is no significant difference between NVH experts and non-experts in those plots. This shows that NVH professionals are not more consistent in their ratings than non-NVH participants.

4.1.3 Experiment ratings

Using this consistency and condordance test, no outliers were removed. The results of the test can now be plotted, in the form of box plots. There was 5 batches of sounds, and they sounds use for consitency calculation are now shown in th plots. Every sounds in there consitutues the training datastet for th emodels later.

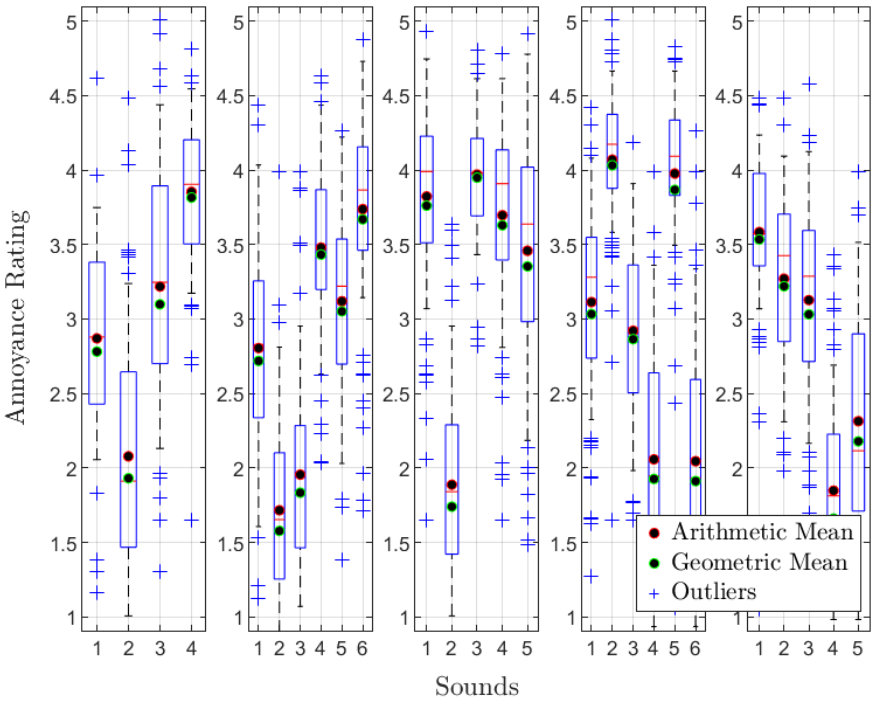


Figure 4.4: Box plots ratings of annoyance

The average annoyance ratings range from 1.7 to 4.1. As expected, only a portion of the scale is used, with a slight bias toward higher values: 16 out of the 26 sounds are rated above 3 (the midpoint of the scale).

Out of 84 participants, each sound has on average 10 outliers, which represents around 12% of the sample — an acceptable rate for such a large group. The interquartile range (IQR) rarely exceeds 1, indicating that the ratings are quite precise. For instance, a sound rated around 3 could realistically be perceived as 2.5 or 3.5 depending on the individual, which reflects real-life variation.

Another observation is that the ratings tend to be more dispersed for sounds with low annoyance. This is understandable: when a sound is pleasant, it becomes more difficult to differentiate between scores like 1, 1.5, or 2. Additionally, such sounds are often quieter, offering fewer perceptual cues to interpret. In contrast, more annoying sounds provide clearer elements for evaluation, making them easier to rate consistently.

Fig. 4.5 shows the box plots for the other metrics :

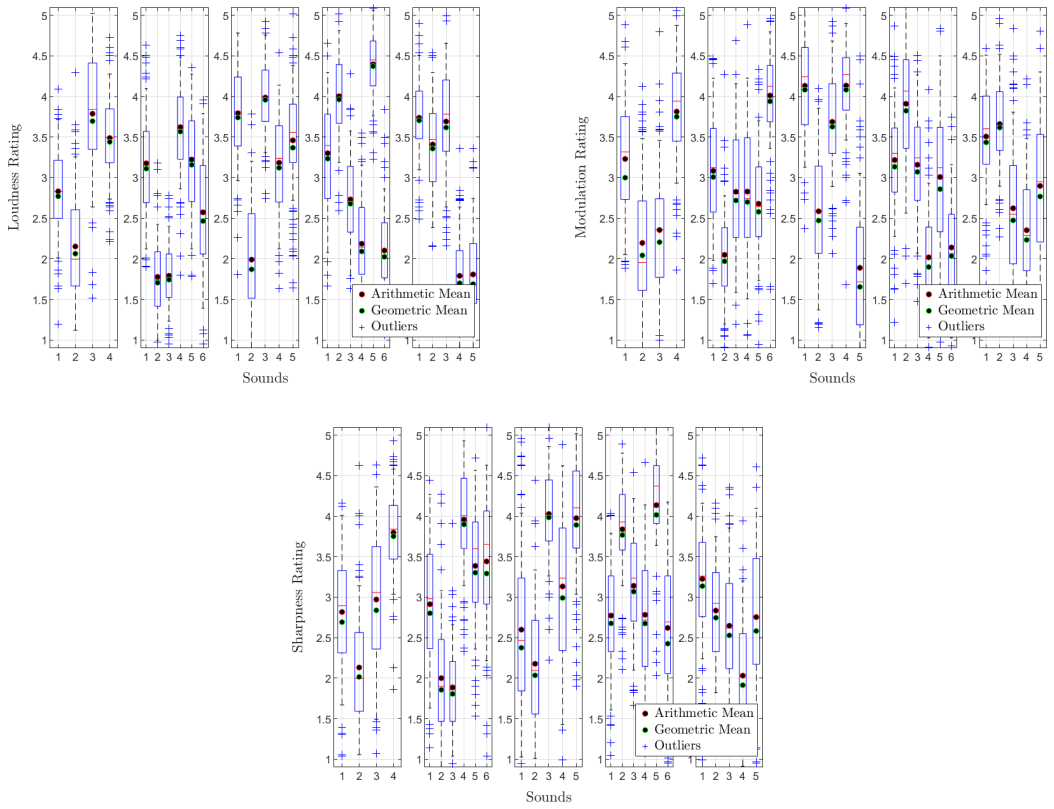


Figure 4.5: Box plots for loudness, modulation and sharpness from the listening test

The same observations made for annoyance ratings also apply to the other metrics. Loudness tends to produce more consistent ratings with a lower IQR, while modulation and sharpness show greater dispersion, particularly at lower values. Seeing that there is a bias toward higher values for all metrics, it seems important to observe the effects of unbiasing, as explained in Section. 2.3.4.

4.1.4 Effects of unbiasing

Unbiasing was applied to all jurors so that each participant's average response across all sounds was 3 out of 5.

4.1.4.1 Consistency & concordance

This unbiasing process shifts participants' ratings toward the center of the scale, which is expected to improve concordance values without significantly affecting consistency.

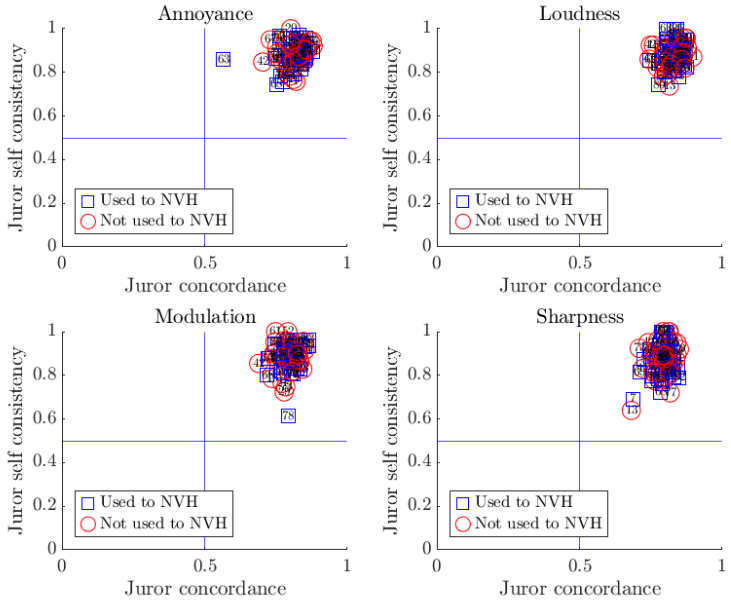


Figure 4.6: Consistency vs. Concordance after unbiasing

The number of potential outliers is reduced, and participant responses are more tightly clustered. While some individuals still exhibit low concordance or consistency, removing them has negligible impact on the overall results (similar to the original analysis). As before, loudness continues to show strong grouping, and concordance values are generally slightly higher, as expected.

4.1.4.2 Ratings

To better understand the effect of unbiasing, it's useful to compare box plots of annoyance ratings before and after the process. In this comparison, the original box plot is shown without any added noise, since unbiasing is applied directly to the raw ratings and not to the noise-augmented versions used for visualization purposes.

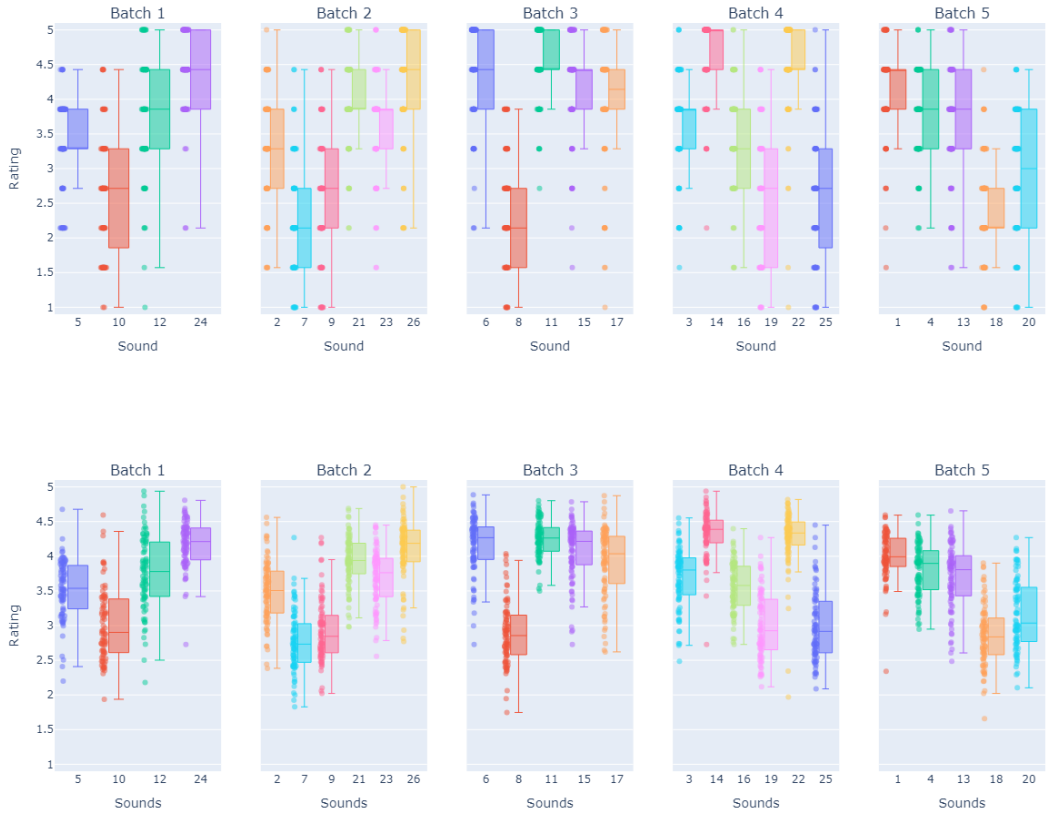


Figure 4.7: Annoyance ratings before and after unbiasing

Unbiasing clearly reduces variability in the responses while preserving the overall trend. When comparing each group of sounds before and after unbiasing, the relative order and rating pattern remain consistent, but the interquartile range (IQR) becomes smaller in the unbiased case. However, this compression also results in a narrower scale, with average values now ranging from 2.7 to 4.4. While some form of rescaling could be considered, it's more insightful to examine the correlation between the original and unbiased ratings.

The Pearson correlation coefficient between the original (biased) and unbiased ratings is 0.998 for each metric, indicating a very strong linear relationship. This confirms that unbiasing primarily performs a linear shift on participants' ratings, preserving the overall structure and trends of the data. For this reason, it was decided that unbiasing would not be further used, and that using the original results is more authentic in this case.

4.1.5 Groups comparison

Using the original ratings for each group, the Pearson correlation coefficient between groups was computed.

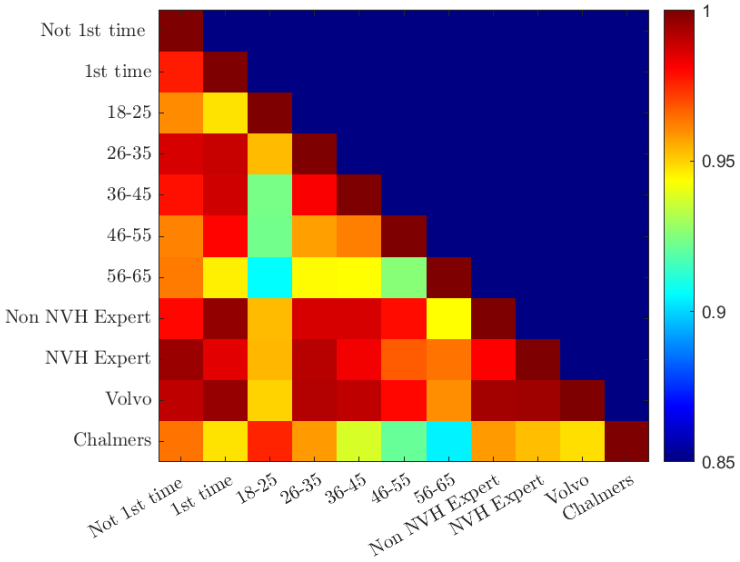


Figure 4.8: Correlation matrix between participant groups

Figure 4.8 shows consistently high correlation coefficients between all groups, with a minimum value of 0.9. As expected, the 56–65 age group shows the lowest correlation with the others, particularly with the 18–25 group (mainly composed of Chalmers students). Interestingly, the correlation between first-time and non–first-time participants is 0.97, as is the correlation between NVH experts and non-experts. This indicates that all groups followed a very similar trend in their ratings, even if absolute values may differ slightly.

To further analyze group differences, a RMSE matrix was also generated:

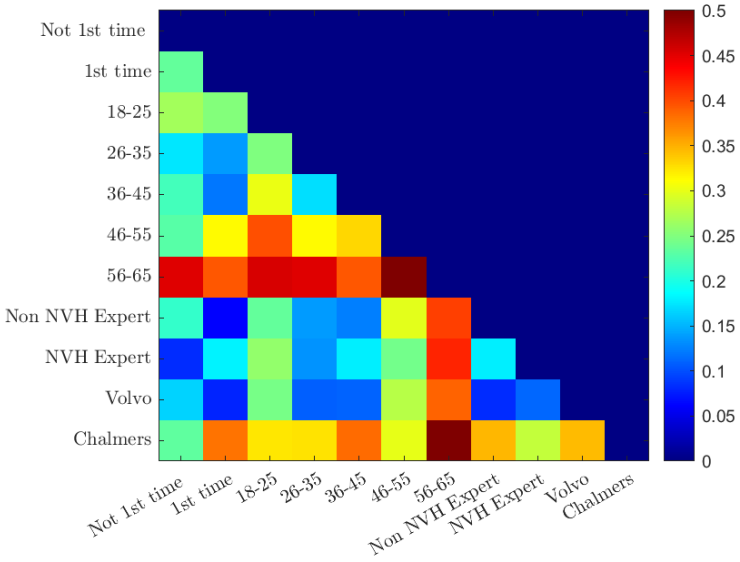


Figure 4.9: RMSE matrix showing differences in ratings between participant groups

Figure 4.9 helps to measure the differences between groups ratings. The average RMSE is around 0.2 on a 5-point scale, which is relatively low. The highest RMSE, approximately 0.5, occurs between the 56–65 age group and other groups. However, this age group had significantly fewer participants and is therefore not very representative. Since the correlation values remain high and trends are preserved, this group was retained in the analysis.

To further investigate if clusters or subgroups exist within the participant responses, hierarchical clustering was performed using dendrogram, as shown in Fig. 4.10.

A dendrogram divide participant responses based on similarity. The most relevant aspect of these plots is the distribution of group members along the x-axis. In each plot, black dots indicate participants belonging to a specific group. If a distinct trend were present within a particular group, the corresponding points would be expected to cluster together in a specific region of the x-axis. However, this is not observed: the points are consistently distributed across the entire axis in all cases.

This analysis confirms the absence of any major differences in responses between participant groups. It indicates that factors such as age, background, or experience do not significantly influence how the sounds were rated.

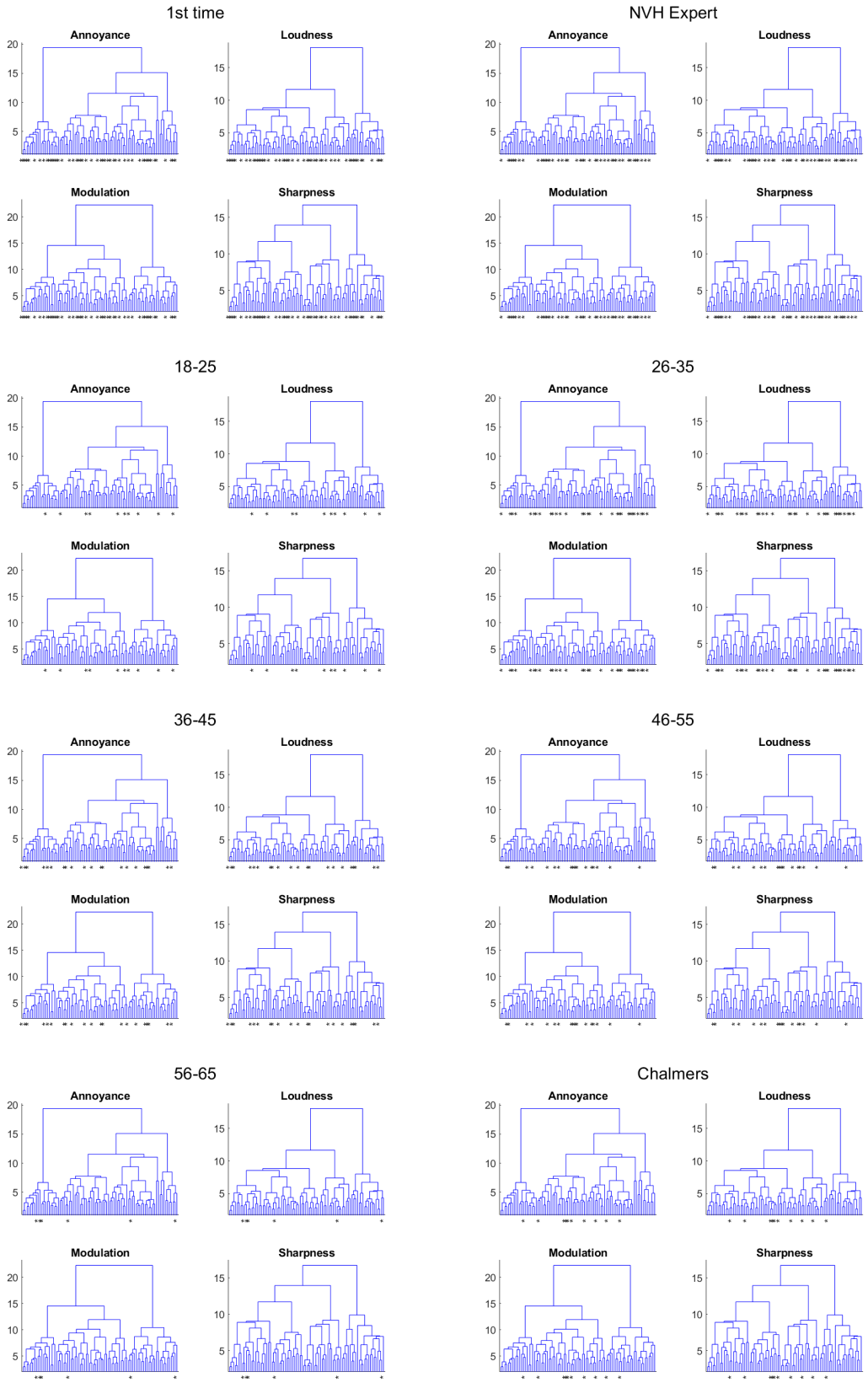


Figure 4.10: dendrogram for the each participant group

4.2 Feature Choice

The jury test results have been validated, showing no significant outliers or subgroups of participants with divergent behavior. Thus, it is now possible to move forward with identifying the objective metrics that best correspond to the test metrics. Despite being named after objective properties such as loudness, modulation, or sharpness, the test metrics were rated subjectively by participants. As a result, they may not correspond directly to the formal definitions used in psychoacoustic models.

4.2.1 Experiment's metrics

Before selecting features or modeling approaches, it is essential to verify that all test metrics are related to annoyance. As in previous steps, a correlation matrix is used to assess the strength of these relationships.

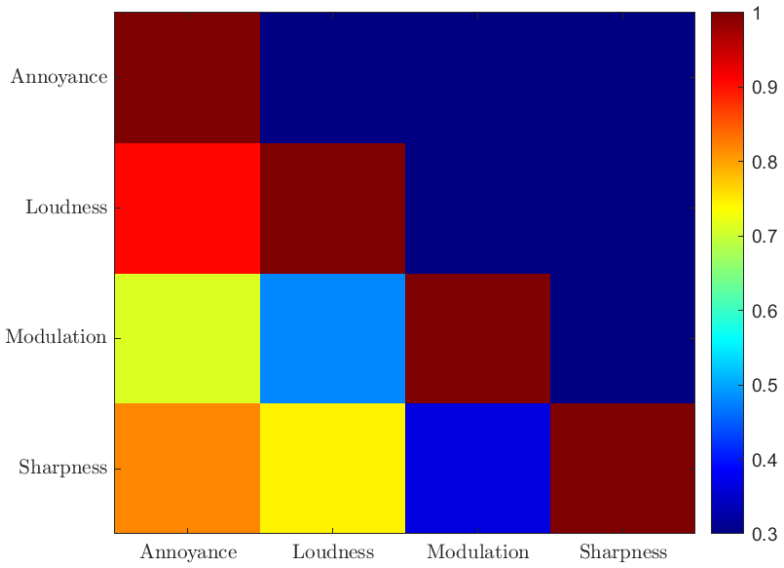


Figure 4.11: Correlation matrix between annoyance and test metrics

As shown in the first column of Figure 4.36, annoyance exhibits a correlation coefficient of at least 0.7 with each of the tested metrics. Loudness shows the highest correlation, followed by sharpness, and then modulation. This trend aligns with expectations and will likely be reflected in feature importance rankings when using models such as decision trees in later stages. The plot also reveals some expected relationships between the objective metrics. For instance, loudness and modulation show low correlation, indicating they capture different perceptual aspects of the sounds (which is the case by definition). In contrast, sharpness and loudness are more strongly correlated, which is consistent with the fact that sharpness is calculated as a weighted function of the loudness spectrum.

The relatively strong correlations indicate that all three metrics are significantly related to annoyance and are therefore suitable candidates for use in predictive modeling. The next step involves selecting the most appropriate calculation methods for these metrics.

4.2.2 Objective metrics

For each sound, objective metrics were computed using various calculation methods. The correlation between these objective values and the corresponding subjective test metrics was then evaluated.

Loudness	Correlation coefficient
ISO 532-3 (Cubic Average)	0.907
ISO 532-3 (Max)	0.862
ISO 532-3 (N5)	0.877
ISO 532-3 (N10)	0.893
ISO 532-1 (N5)	0.887
ISO 532-1 (Cubic Average)	0.909
ISO 532-1 (Max)	0.757
ISO 532-1 (N10)	0.896
ECMA 418-2 3rd (Cubic Average)	**0.918**
ECMA 418-2 3rd (N5)	0.901
ECMA 418-2 3rd (N10)	0.903

Table 4.1: Loudness correlation coefficients

All objective methods show a strong correlation with the subjective ratings of loudness from the test. This serves as a positive indication that the test methodology (particularly the use of reference sounds and the equalization) was effective. Given that loudness is generally intuitive for participants to assess, a weak correlation would have raised concerns about the validity of the testing approach. The Hearing Model with the cubic average approach is the most performant method in this case. For the sharpness, the results were way more worriying.

Sharpness	Correlation coefficient
Sharpness | DIN45631; DIN 45692 (S5)	-0.132
Sharpness | DIN 45631; DIN 45692 (Cubic Average)	-0.129
Sharpness | DIN 45631; DIN 45692 (S10)	-0.097
Sharpness | ISO 532-3; Aures (S5)	-0.170
Sharpness | ISO 532-3; Aures (Cubic Average)	-0.171
Sharpness | ISO 532-3; Aures (S10)	-0.127
Sharpness | ISO 532-1; Aures (S5)	-0.128
Sharpness | ISO 532-1; Aures (S10)	-0.082
Tonality | ECMA 418-2 (T5)	**0.680**
Tonality | ECMA 418-2 (T10)	0.675

Table 4.2: Sharpness correlation coefficients

This table reveals that the sharpness ratings from the test reflect more of a tonality perception. Sharpness is known to be a challenging metric for listeners to assess

accurately. It is influenced by both loudness and high-frequency content, and when averaged over time, it may not fully capture transient sharp sounds. For example, if a sound contains a sharp component for only one second within a six-second interval, participants may still rate the overall sound as sharp, even though the averaged sharpness value would be relatively low.

This suggests that the test ratings may represent a mixed attribute combining sharpness and tonality. To verify this hypothesis, sharpness calculation results were multiplied by the tonality values (T5). Using the ISO 532-1 (S10) sharpness method combined with tonality resulted in a correlation coefficient of 0.739 with the test ratings. Therefore, even if this metric si actually mostly tonality, for the remainder of the analysis, it will named "sharpness $\times$ tonality. The last metric to check is the modulation.

Modulation	Correlation coefficient
Modulation Spectrum | Octave 1k 710-1.4k Hz (Average)	**0.447**
Modulation Spectrum | Octave 1k 710-1.4k Hz (Quad Average)	0.407
Modulation Spectrum | Octave 1k 710-1.4k Hz (Min)	0.323
Modulation Spectrum | Octave 1k 710-1.4k Hz (Max)	0.398
Modulation Spectrum | Octave 1k 710-1.4k Hz (L5)	0.439
Modulation Spectrum | Octave 1k 710-1.4k Hz (L10)	0.424
Modulation Spectrum | 1/3 Octave 315 280-355 Hz (Quad Average)	0.255
Modulation Spectrum | 1/3 Octave 100 90-112 Hz (Quad Average)	0.125
Modulation Spectrum | Full bandwidth (Quad Average)	0.242
Fluctuation Strength (F5)	0.3821
Fluctuation Strength (Max)	0.1205
Roughness | ECMA 418-2 (3rd) (R5)	0.2388

Table 4.3: Modulation correlation coefficients

A similar situation occurs with the modulation metric, although the effect is less pronounced than with sharpness. The fluctuations perceived by participants may be caused by the electric seat motor changing speed during operation. Since the motor does not rotate very fast, these fluctuations are likely to be lower in frequency and may not correspond directly to pure modulation. Instead, the metric may represent a combination of modulation and FS in this case.

This hypothesis was tested by multiplying the best modulation spectrum method with the F5 FS method, which increased the correlation coefficient to 0.564. In contrast, multiplying with roughness alone resulted in a maximum correlation of 0.4, indicating that roughness is less representative here. However, these improvements are relatively minor, and the overall correlation remains moderate, so metric selection should not rely solely on these values.

An alternative approach is to compare all objective metrics directly against the annoyance ratings from the test to identify which ones have the strongest correlations. Selecting metrics based on their relationship with annoyance would provide a more relevant basis for modeling.

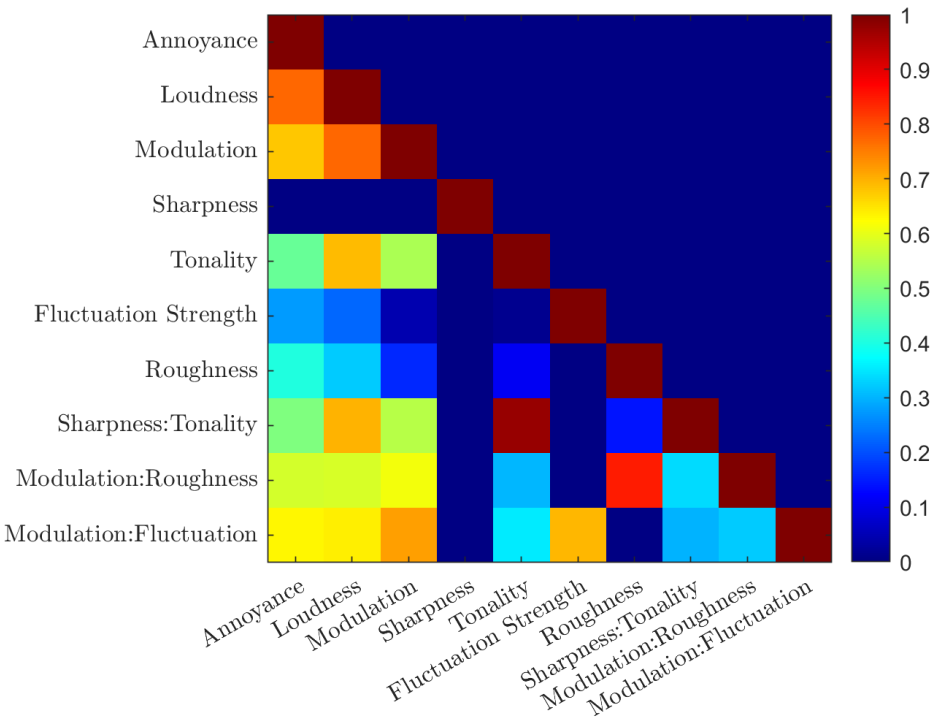


Figure 4.12: Correlation matrix between objective metrics and subjective annoyance ratings

Fig. 4.12 demonstrates that loudness is the primary factor influencing annoyance, followed by the modulation spectrum. Therefore, using modulation alone, without incorporating roughness or fluctuation strength (FS), appears to be more relevant. Since the ultimate goal is to predict annoyance, it's logical to focus on the metrics that have the greatest impact on it.

To maximize the flexibility of the models and facilitate ease of use—such as directly applying them after calculating the objective metrics in ArtemiS Suite, the models will be provided only with the original metrics. However, they will be allowed to consider interactions between these metrics. Thus, the inputs of the predictions models are : **Loudness, Sharpness, Tonality, Modulation, Fluctuation Strength**.

With the inputs ready, the prediction can now be done.

4.3 Annoyance Predictions

The results of the different models are presented and compared in this section, starting from regression models to neural networks and random forests. The data is presented in the experiment's scale, then in the Volvo scale. The RMSE for all methods are compared at the end of the section.

4.3.1 Linear regression

The predicted ratings from the linear regression are pretty close to the original ratings for the training set, and there seems to be no overfitting :

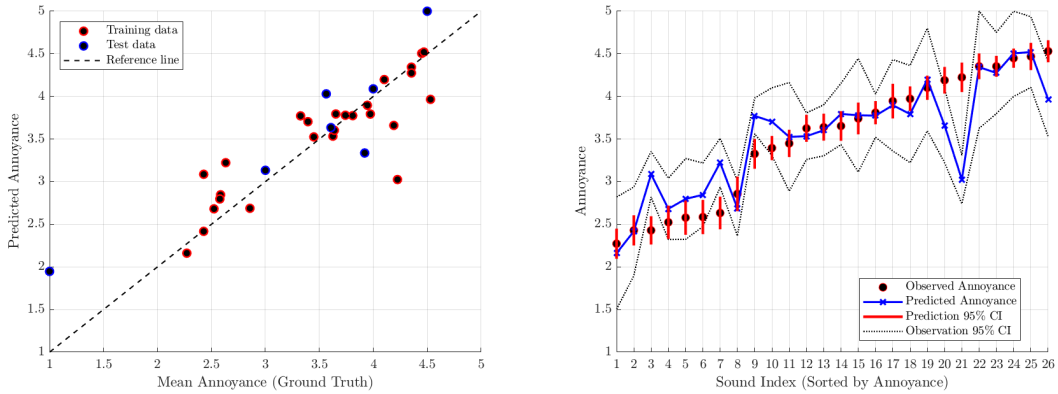


Figure 4.13: Prediction results for the linear regression

The linear regression model performs reasonably well for most sounds but fails completely for certain cases, such as sounds 3 and 21. The prediction accuracy on the test sounds is comparable to that on the training set, which indicates a degree of robustness. However, the overall precision of the model remains limited and clearly leaves room for improvement. This lack of accuracy becomes even more apparent in Figure 4.14:

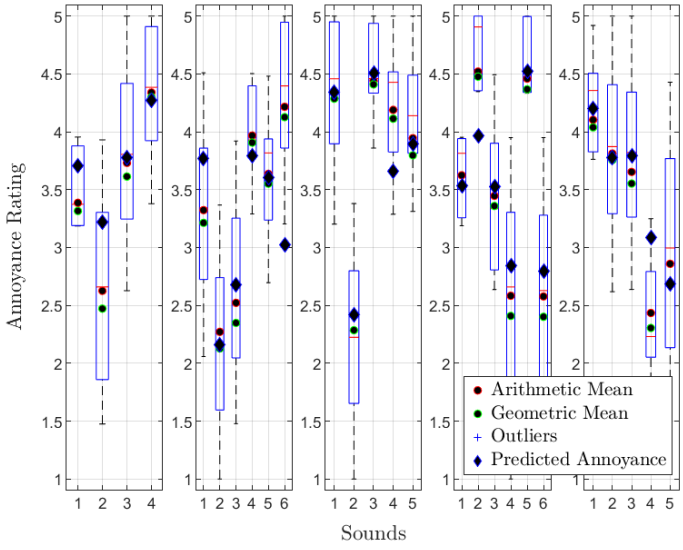


Figure 4.14: Linear regression prediction over the annoyance box plots

Five out of the 26 sounds fall entirely outside the IQR, highlighting notable prediction errors. Despite this, the linear regression model provides a reasonable overall performance and serves as a baseline for comparison with more advanced models. The results can now be scaled back to the Volvo scale.

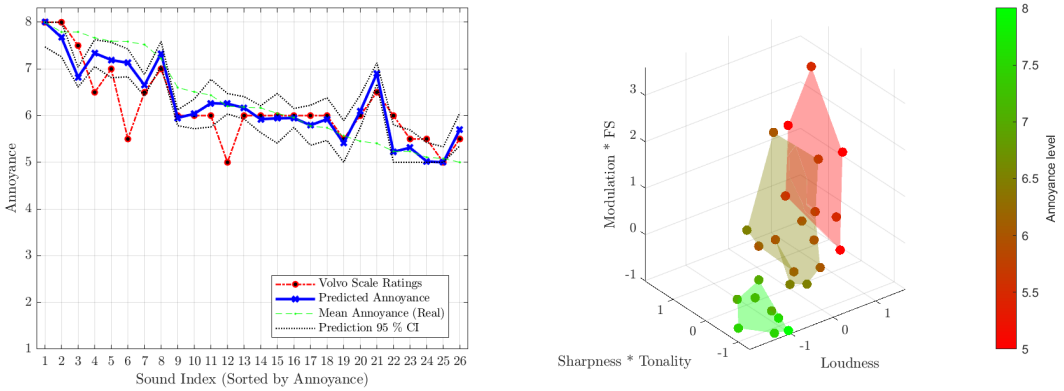


Figure 4.15: Linear regression results in the Volvo scale

In the Volvo ratings, sounds 4, 6, and 12 appear to have been rated unusually harshly. Upon verification, each of these sounds was evaluated by only one or two individuals, which likely explains why their ratings fall so far outside the expected range. For the remainder of the visualizations using the Volvo scale, these sounds will be treated as outliers in the Volvo scale and excluded from plots to improve clarity.

Despite these outliers, the clustering in the 3D space is very distinct: three separate groups of sounds are clearly identifiable. The high-quality sounds are characterized by low loudness, modulation, and sharpness. The average sounds show slightly higher values, particularly with more variation in modulation and sharpness. The lower-quality sounds generally exhibit high loudness, with sharpness and modulation values that vary. As anticipated, loudness emerges as the primary factor influencing the overall rating of the sounds.

4.3.2 Polynomial regression

To try and address the inherent precision problem of the linear regression, multivariate polynomial regression (MPR) is used.

4.3.2.1 Original polynomial

he predicted ratings from the (MPR) model are highly accurate for the training set. However, there is clear evidence of overfitting, as the model fails to generalize well to the test data.

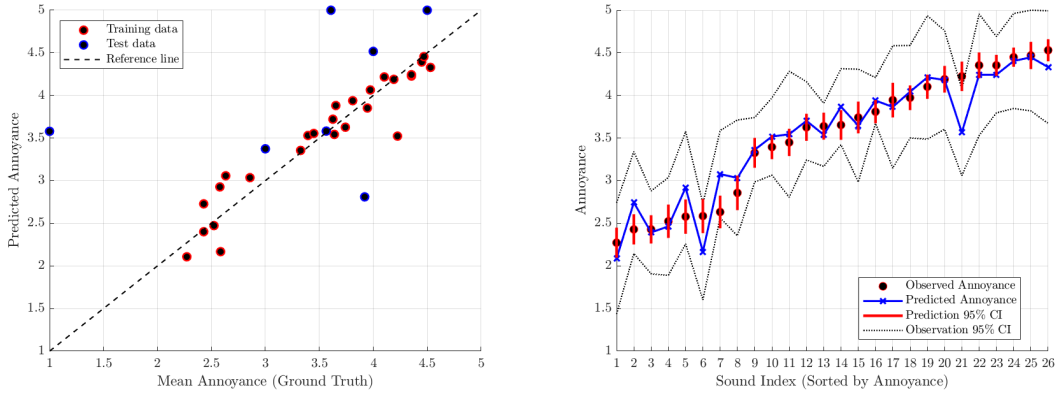


Figure 4.16: Prediction results for the MPR

Even with a highly precise model, sound 21 remains outside the prediction confidence interval. This indicates that the sound has a specific characteristic not captured by the model. It is likely a short or abrupt sound, something that is not reflected in the average values of the objective metrics. This particular case will be discussed in more detail in the following chapter.

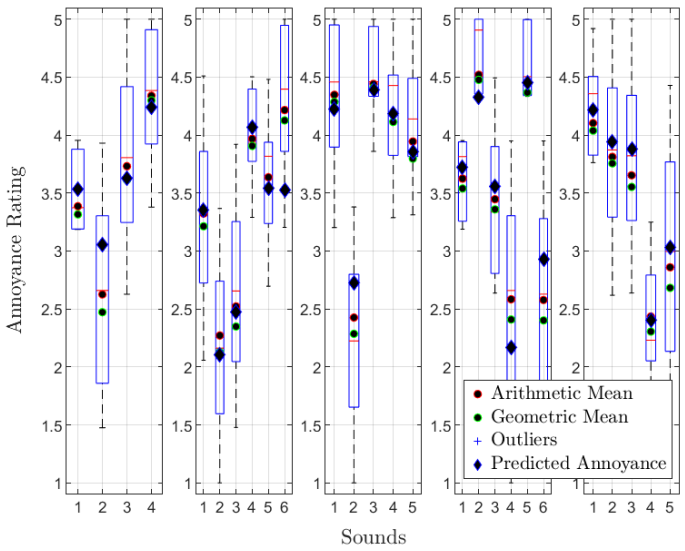


Figure 4.17: MPR prediction over the annoyance box plots

The model performs well on the training data, with only one sound falling outside the IQR, indicating that it successfully learned the patterns within the training set. However, the poor accuracy on the test data shows a clear overfitting issue. This suggests that the model is too complex relative to the available data and is likely fitting to noise or irrelevant variations. To improve its robustness and predictive performance, the model must be simplified—either by reducing the number of features, lowering the model's complexity, or applying regularization techniques. Here, the number of features will be reduced by hand.

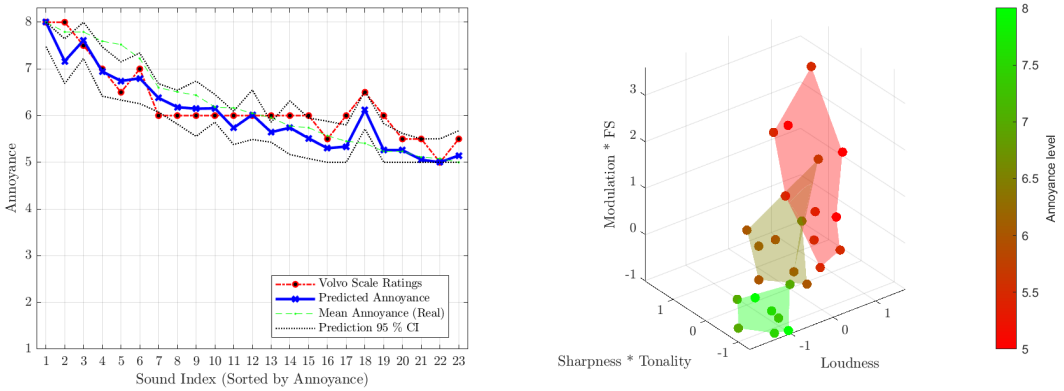


Figure 4.18: MPR results in the Volvo scale

The clusters remain clearly defined, and the model tends to give slightly harsher ratings than the linear model, pushing more sounds into the "bad" category. Once the Volvo outliers are removed, the predicted results closely align with the ratings given by the component NVH team. This indicates that a well-calibrated model can reflect expert judgment accurately and could be applied in practice with minimal adjustment.

4.3.2.2 Simplified polynomial

From the original polynomial regression model, 3 parameters were removed : Sharpness, Modulation × FS and Modulation × Tonality.

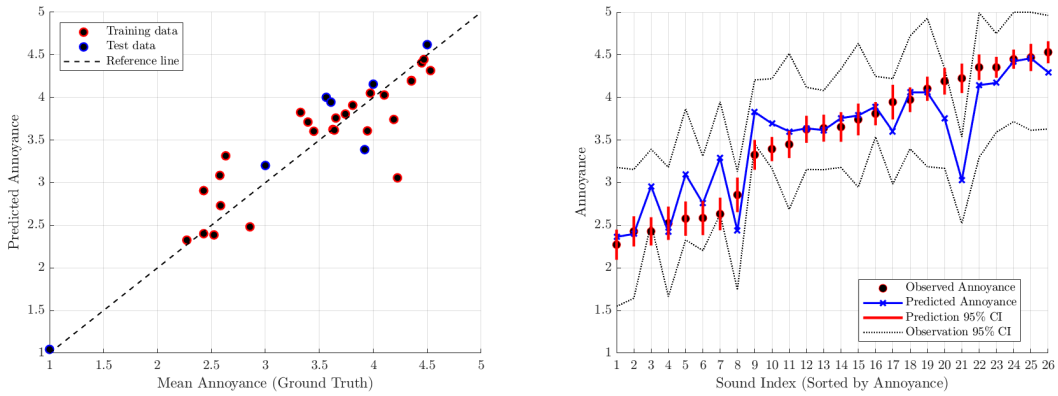


Figure 4.19: Prediction results for the Simplified MPR

This adjustment effectively addressed the overfitting issue, at the cost of reduced precision on the training data. The model is now less confident in its predictions, as reflected by wider confidence intervals compared to previous versions. However, this trade-off results in significantly improved test accuracy. Overall, this marks a step forward in achieving a more balanced and reliable framework.

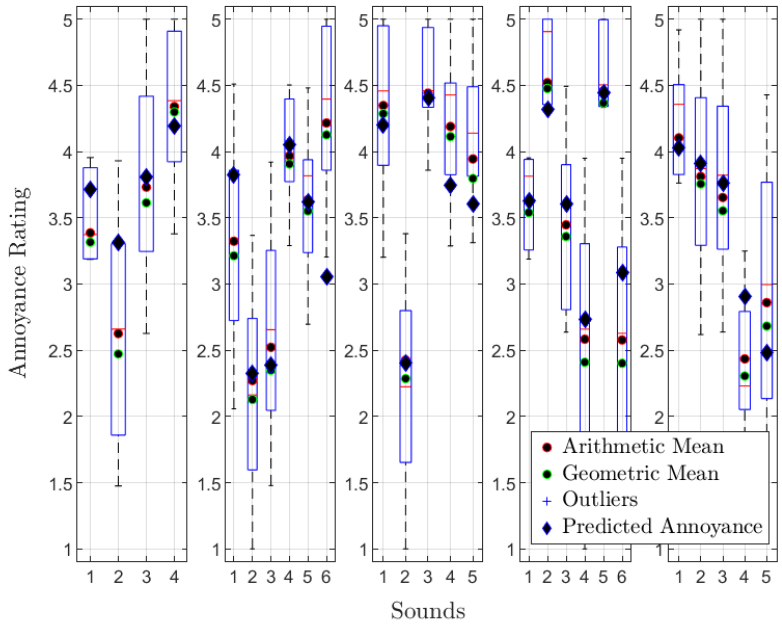


Figure 4.20: Simplified MPR predictions over the annoyance box plots

Four sounds fall outside the interquartile range, though three of them are very close to the boundary. All in all, the model looks already sufficiently reliable for practical use.

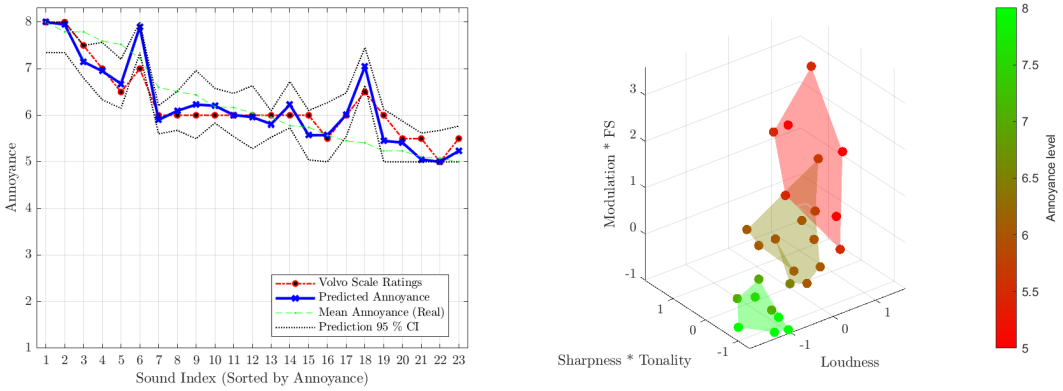


Figure 4.21: Simplified MPR results in the Volvo scale

The model continues to follow the trend observed in expert ratings, and the clusters are now even more clearly defined. This suggests that, after loudness, modulation has become the second key factor distinguishing average sounds from poor ones. With these improved results, the model's performance can now be compared against that of a neural network.

4.3.3 Simple neural network

The neural network produces highly unstable predictions in this case. As shown in Fig. 4.22, the predicted annoyance values deviate significantly from the actual ones. To help the model better understand the range of possible ratings, two sounds were included in the training set : those with the highest and lowest annoyance scores. However, the results remained unchanged when these extreme cases were excluded, indicating that their presence did not improve the model's stability or performance.

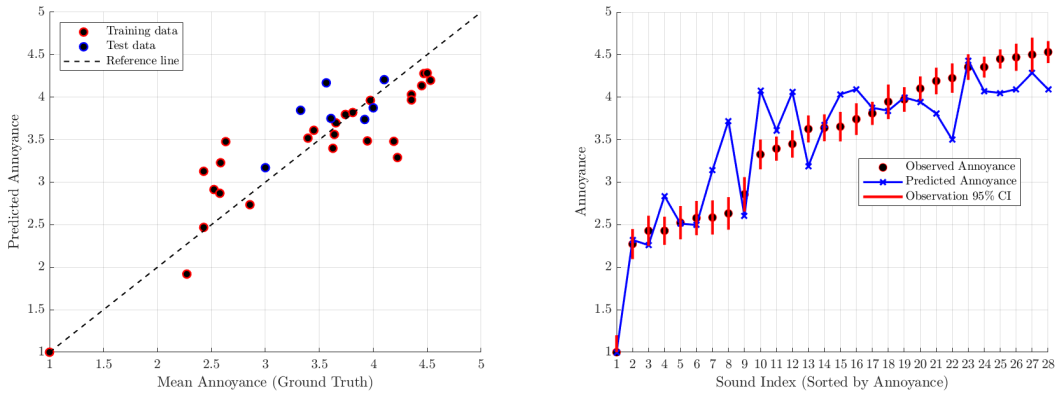


Figure 4.22: Prediction results for the neural network

The box plots confirm this lack of precision and instability :

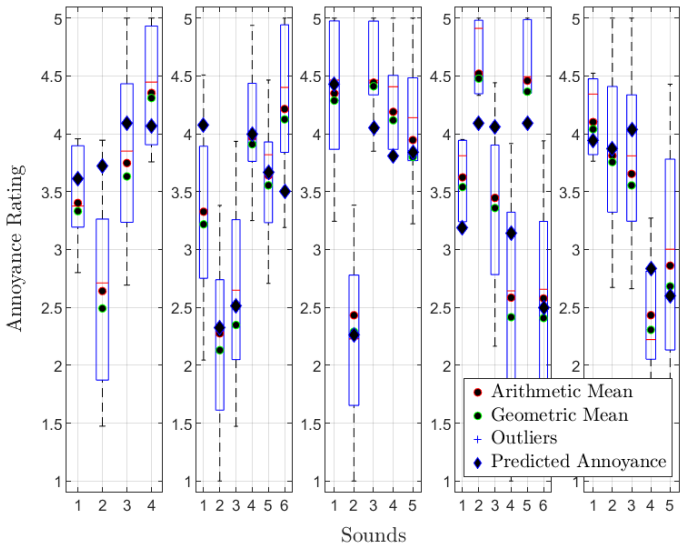


Figure 4.23: Neural network predictions over the annoyance box plots

In the Volvo scale, the model continues to follow the general trend of expert ratings and produces well-defined clusters. However, a key issue with the neural network becomes apparent in the 3D space: modulation appears to dominate the grading,

whereas loudness should be the primary factor. This suggests a misalignment in feature weighting within the network. Despite this, the overall clustering structure remains consistent with previous models, maintaining the distinction between good, average, and poor sound categories.

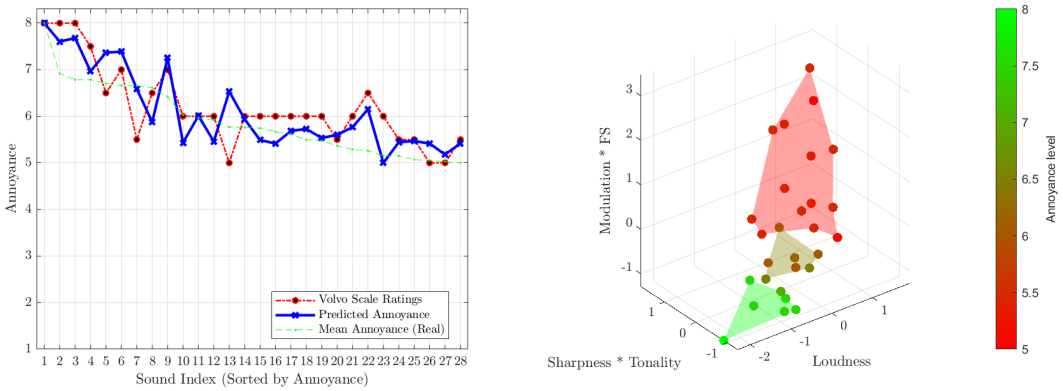


Figure 4.24: Neural network results in the Volvo scale

The clear separation of sounds observed across all prediction models strongly suggests the potential for defining distinct regions within the 3D space. Each region would correspond to a specific annoyance rating. This motivates the use of a model capable of partitioning the feature space in a structured and interpretable manner. A random forest model is particularly well-suited for this task.

4.3.4 Random forest

The random forest model proves to be the most suitable for this regression task. It captures specific patterns from the training data while maintaining avoiding overfitting. Fig. 4.25 shows the performance of the model using the scale of the listening experiment.

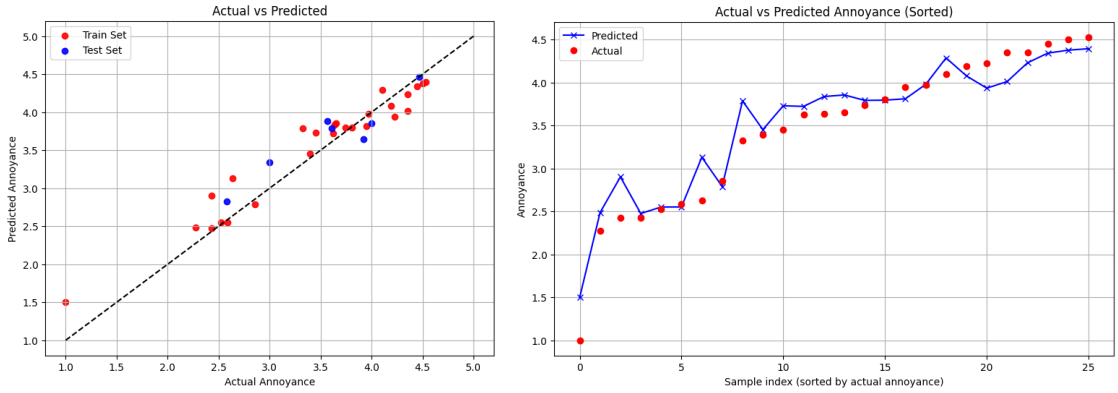


Figure 4.25: Prediction results for the random forest

To calculate the most optimal hyperparameters, a grid search was performed. It can be visualized as finding the minimum point coordinates of this 3D surface :

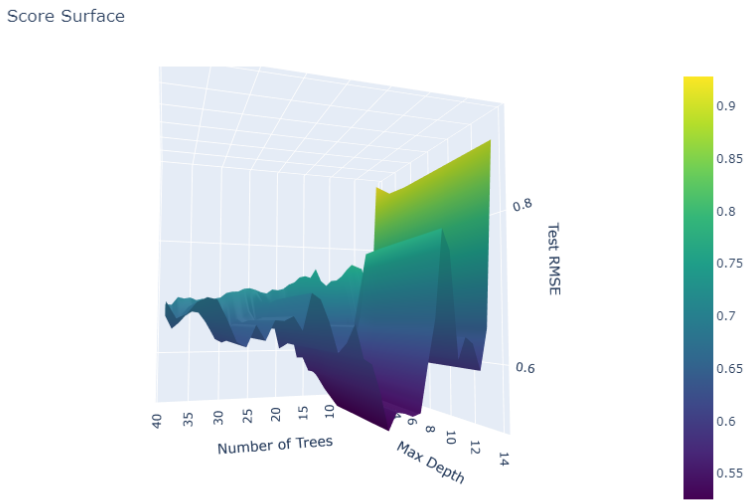


Figure 4.26: Surface of RMSE values used to find the best hyperparameters for the random forest model

The overfitting effect becomes clearly visible when the number of trees is too high. As shown in Figure 4.26, the maximum depth of the trees does not significantly affect the model's performance. Instead, the RMSE is primarily influenced by the number of trees.

Figure 4.27 illustrates how the optimal number of trees (15 in this case) can be identified: it corresponds to the point where the training and test RMSE are approximately equal, which is ideal. Beyond this point, increasing the number of trees causes the training RMSE to plateau while the test RMSE starts to increase.

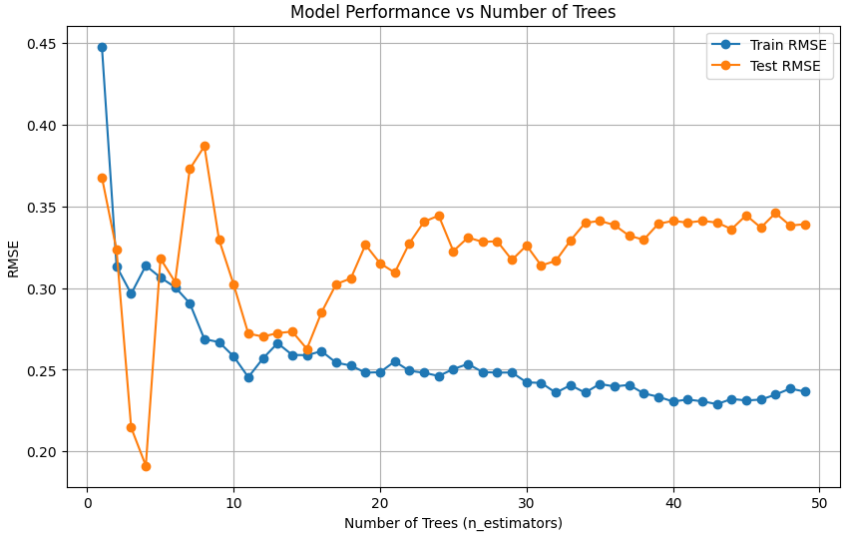


Figure 4.27: Training and testing loss versus number of trees, showing overfitting beyond a certain point

The main advantage of the random forest approach lies in its interpretability. By plotting one of the decision trees, the structure of the decision-making process becomes clear:

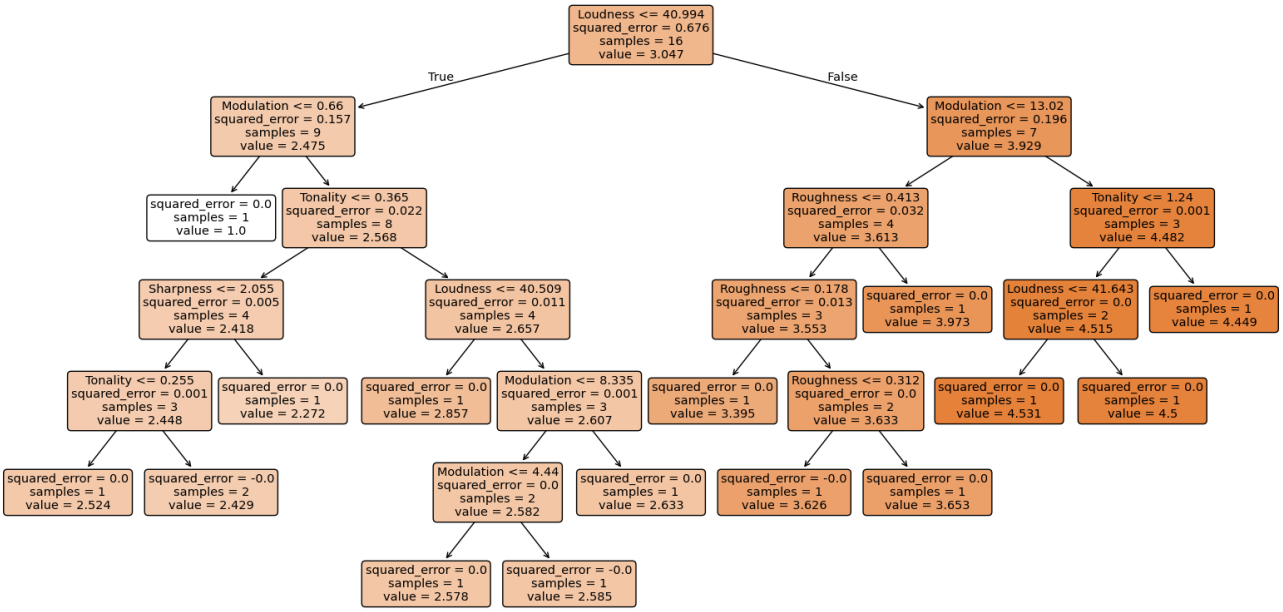


Figure 4.28: Example of a single decision tree used in the random forest

The first main split in the decision tree is based on the most influential metric. As expected, loudness is identified as the primary factor. The model then considers modulation and tonality. This order of importance is both logical and consistent with acoustic intuition. The first three splits are the most significant, as they effectively divide the dataset into three main categories. Using this model leads to the following results in the Volvo scale :

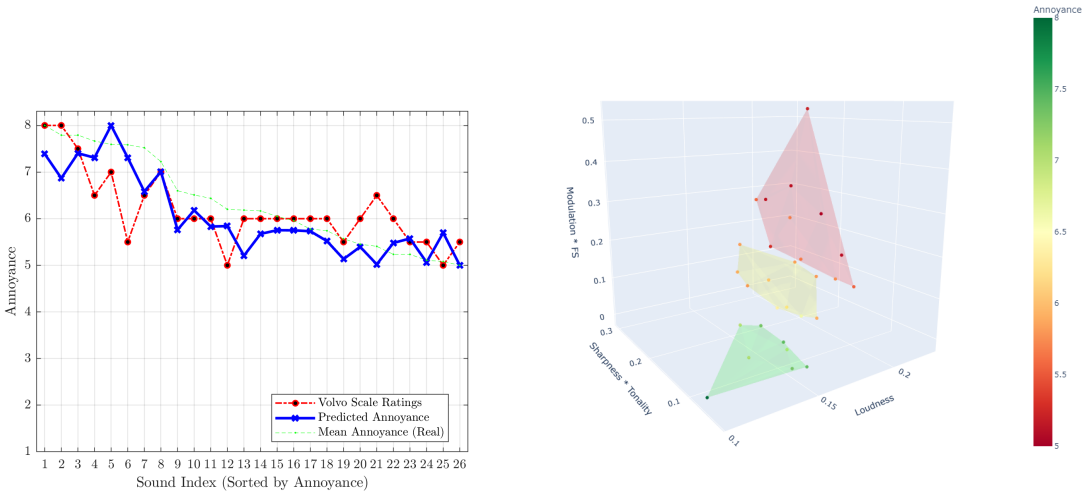


Figure 4.29: Random forest results in the Volvo scale

The random forest model follows the same general direction as the expert ratings, although it does not exactly replicate them. Given how accurately the model performs using the average ratings from the 84 participants, this suggests that the current approach to rating sounds could benefit from refinement, potentially using this model as a reference. In this case, the clusters are clearly defined: **good** sounds exhibit low loudness, sharpness, and modulation; **acceptable** sounds show increased loudness and modulation; and **poor** sounds present even higher values in both, with sharpness becoming a decisive factor in distinguishing between acceptable and poor.

4.3.5 Effects of data augmentation

To try and imporve the results,as mentioned in Section 3.6, data augmentation was tried both on the parameters themselves and on the sounds.

4.3.5.1 On the metrics

Out of the 33 original sounds, 27 were randomly selected and augmented to create close neighbors. Combined with 15 of the new metrics, this forms a training set of 42 sounds. The remaining 6 originals and 12 augmented counterparts make up the 18-sound test set. Annoyance for the augmented sounds was increased linearly, following Equation 2.11.

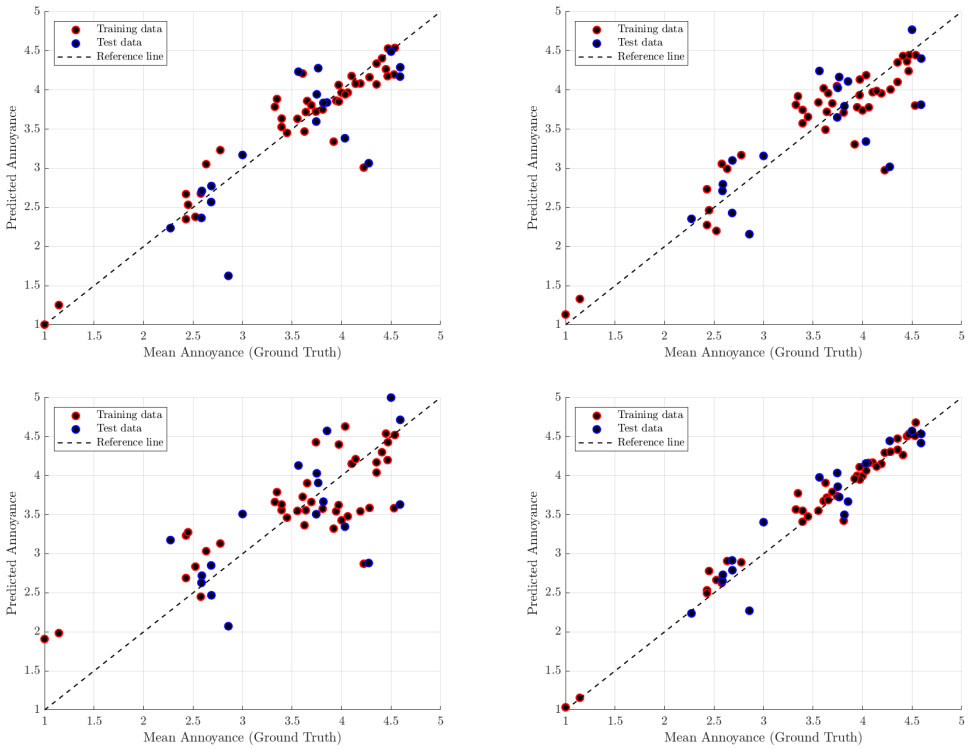


Figure 4.30: Predicted vs. actual annoyance with data augmentation: Top left – linear regression; Top right – polynomial regression; Bottom left – simplified polynomial regression; Bottom right – neural network.

Data augmentation does not significantly improve performance for the regression models. Instead, it confirms the models' stability: when two sounds have similar parameters, their predicted annoyance is also similar. This reinforces the idea that regression models are consistent : small changes in input yield small changes in output.

However, data augmentation greatly enhances the neural network's performance. With a larger and more diverse dataset, the network architecture could be increased to two hidden layers with 16 and 20 neurons, respectively. The earlier model likely lacked sufficient data to learn effectively. Here, the neural network benefits from the added variation without signs of overfitting. Plotting the neural network predictions over the original box plots in Fig. 4.31 further highlights the model's precision:

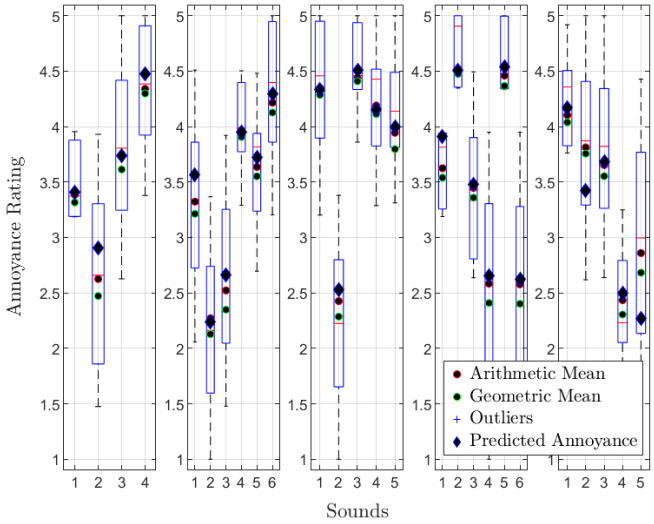


Figure 4.31: Neural network predictions over annoyance box plots

All the predictions fro the training set are inside the IQR, showing very good precision. Finally, the random forest model was tested with the augmented data:

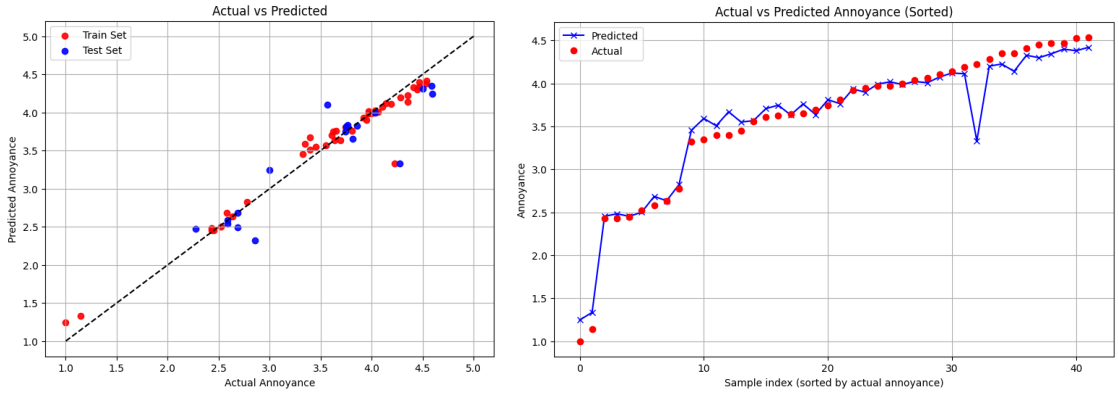


Figure 4.32: Random forest predictions after data augmentation

Even though the model already performed well, data augmentation helped consolidate the prediction structure. As observed earlier, tree depth has limited impact, while the number of trees remains the most critical parameter:

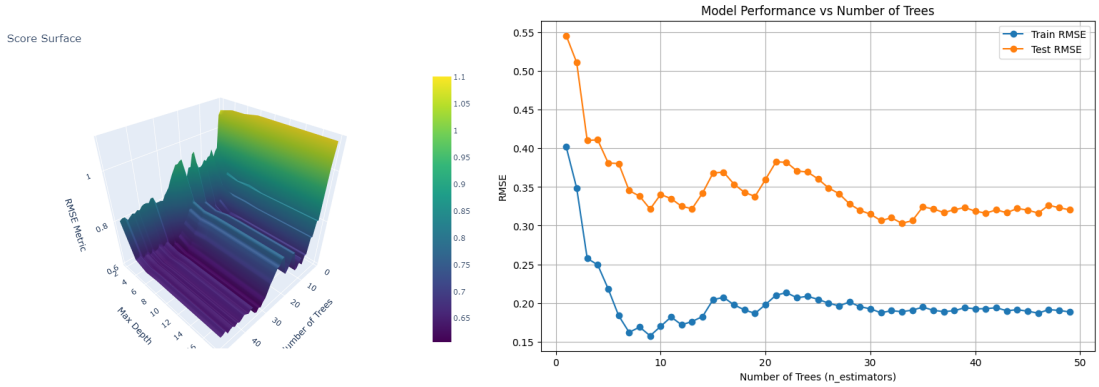


Figure 4.33: Hyperparameter tuning for the random forest with data augmentation

Using data augmentation slightly blurs the boundary between average and good sounds in the 3D feature space. However, the overall clustering structure remains consistent with the non-augmented case, preserving the logical separation of sound categories:

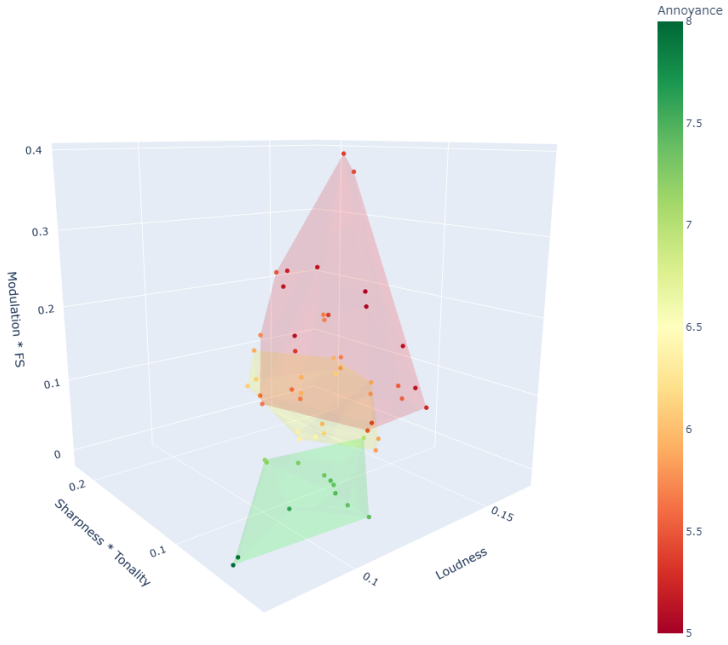


Figure 4.34: Sound clustering in 3D feature space using the random forest model trained on augmented data.

4.3.5.2 On the sounds

Given that data augmentation had a noticeable impact primarily on more complex models, this second method was tested only with the neural network. Out of the 33 original sounds, 13 were selected and modified by adding slight background noise using ArtemiS Suite. This process was time-consuming, which limited the number of augmented samples. A total of 32 sounds were used for training and the remaining 14 for testing.

Two hypotheses were considered regarding how the added noise might affect perceived annoyance:

- Hypothesis 1: The annoyance remains unchanged.
- Hypothesis 2: The annoyance slightly increases (up to 5%), as the added noise level did not exceed 5% of the original sound level.

The optimal neural network structure was again determined using a grid search. The resulting performance for both hypotheses is shown in Fig. 4.35.

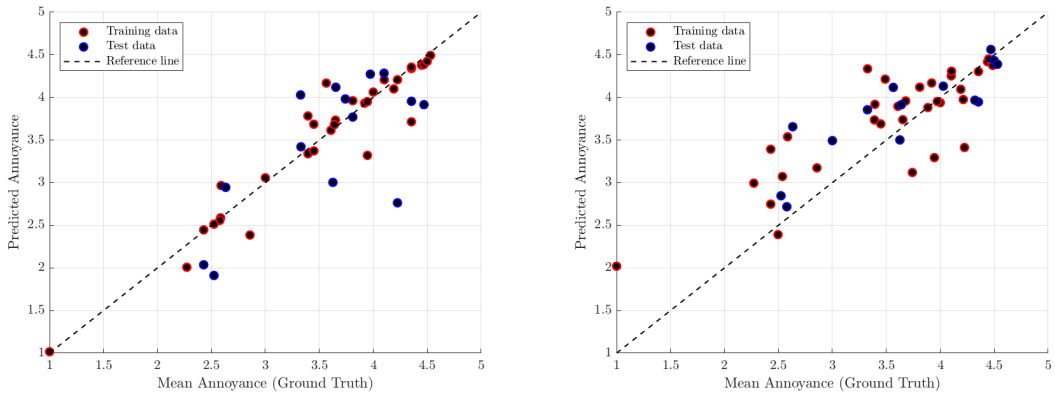


Figure 4.35: Prediction results for hypothesis 1 (left) and hypothesis 2 (right)

Neither hypothesis led to an improvement in the neural network's performance. In fact, slight signs of overfitting were observed. Consequently, this method was discarded.

4.3.6 Best model : RMSE comparison

After testing and visually analyzing all models and methods, the initial observations can be confirmed by comparing the training and test RMSE values across all approaches.

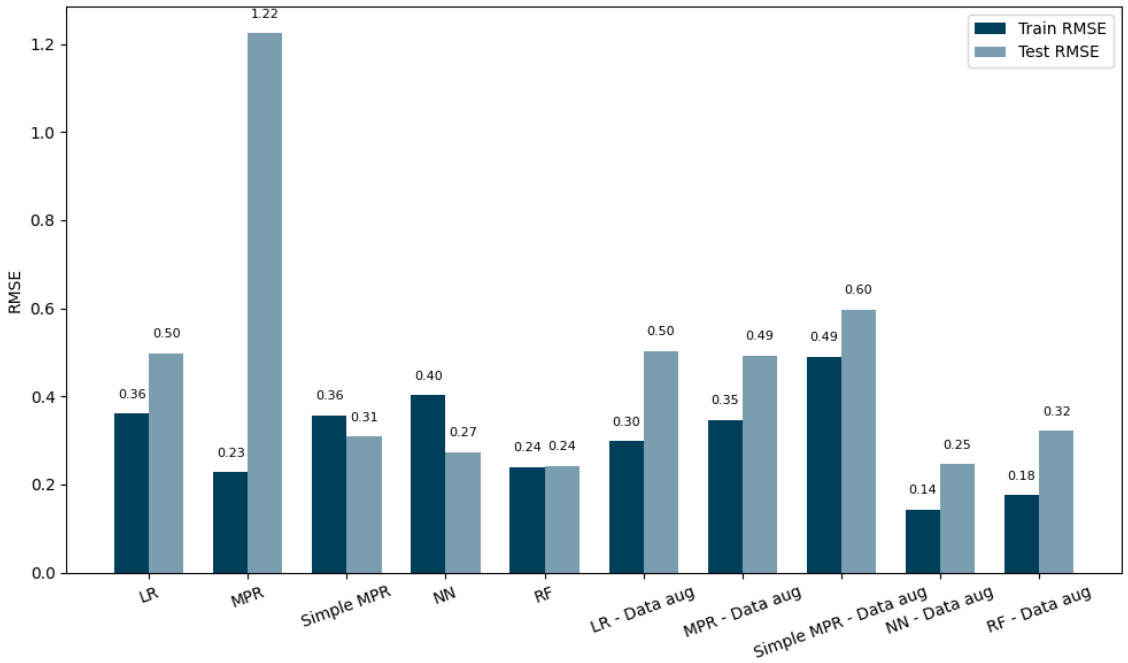


Figure 4.36: Comparison of Train and Test RMSE for all models

Among all tested models on the original dataset, the random forest performs the best, achieving equal RMSE on both training and testing sets. This indicates excellent generalization and robustness without overfitting.

From a research perspective, the neural network with data augmentation offers the best raw performance and is therefore the most suitable for pushing the boundaries of accuracy. However, in an industrial context, interpretability is a key requirement. RF models provide transparency that allows predictions to be double-checked and understood, which is crucial in practical applications. For this reason, the RF trained on the original data remains the preferred choice in real-world deployments.

It's important to note that these RMSE values were calculated on a [1,5] scale. Achieving RMSE values below 0.25, and even as low as 0.2 for some models, indicates a high degree of precision. These results demonstrate that predicting subjective ratings using objective metrics can be an effective and reliable alternative to time-consuming human jury tests.

5

Discussion

In this chapter, the results shown in the previous section are further discussed. The discussion involves methodological considerations and implications of the key results.

5.1 Interpretation of Key Results

This section interprets the main outcomes of the study. It relates them to participant perception and model performance.

5.1.1 Models performances

The central aim of this research was to determine whether objective psychoacoustic metrics could reliably predict subjective annoyance ratings for seat adjustment sounds. The results demonstrate that such prediction is indeed feasible, with loudness and modulation emerging as the most significant predictors of perceived annoyance. During the listening test feedback session, many participants noted that loud sounds were consistently bothersome. Additionally, sounds with strong modulation were often associated with poor-quality or malfunctioning components. Participants emphasized that a stable, consistent sound reflects better build quality, especially since the internal mechanisms of the seat and motor are not directly visible. In this context, sound serves as the main indicator of quality. The prediction results closely align with these user perceptions and expected trends.

Linear regression remains, as expected, a robust and reliable baseline model: it resists overfitting and consistently delivers reasonable results, while still leaving ample room for improvement with more sophisticated approaches. However, polynomial regression models, despite their increased complexity, failed to provide substantial improvements over simpler approaches.

The random forest model exhibited superior generalization capabilities, achieving nearly identical RMSE values across training and testing datasets. This consistency suggests robust model performance with minimal overfitting. Neural networks, while capable of marginally better raw performance through data augmentation, displayed a lack of precision in the original dataset and lacked the interpretability crucial for industrial applications. This finding suggests that tree-based methods may adequately capture the relationship between psychoacoustic features and annoyance

perception, particularly when feature interactions are handled automatically within the model architecture.

5.1.2 Dataset size and data augmentation

While the results are promising, the study is limited by its dataset size (33 original sounds). The restricted variability in the soundscape might constrain the generalization of the findings. Additional data, especially from varied seat types and brands, would enable more robust models and better capture outliers or rare perceptual phenomena. This is a common observation in studies of this nature. The core challenge lies deeper: capturing high-quality sound recordings is both time-consuming and costly. For example, in this study, the process involved transporting a car to a semi-anechoic chamber, setting up multiple microphones and recording equipment, and conducting extensive recording sessions. Future work could involve generating synhtetic sounds through deep generative models [22].

Moreover, only basic models were explored due to the dataset size. Future research could examine attention-based architectures, or transfer learning from related audio domains, for example using spectrograms as inputs for a Convolutional Neural Network (CNN), which can be more easily augmented that audio signals for example.

Data augmentation strategies (specifically, injecting noise into the input parameters) proved effective for the simple neural network but were unsuccessful in other models. In contrast, directly adding noise to the audio signals yielded no benefit. The limited impact of noise injection across most models may be attributed to the relatively small size of the training dataset. These findings underscore a key challenge in psychoacoustic prediction: while data augmentation is a standard technique in many machine learning domains, its application in perceptual audio tasks must be approached with caution to ensure that the integrity of the subjective labels is maintained.

5.2 Methodological reflections

This section reviews the quality of the listening test and analysis techniques.

5.2.1 Listening test design

The listening test was designed and executed with high methodological rigor. The inclusion of reference sounds, repetition for consistency measurement, and efforts to minimize bias through normalization significantly strengthen the validity of the collected subjective data. Results like high loudness correlation strengthen the validity of the test.

The decision not to exclude outliers unless they are clearly inconsistent helped preserve a robust sample size (84). This enhanced the statistical power of the analysis. This sample size is notably larger than typical for such studies, which often include no more than 40 participants. An alternative approach could have involved introducing a second test with new sounds midway through the experiment, allowing the first test to be rated by 42 participants and the second by all 84.

However, this option was considered too uncertain, given that the final number of participants was not guaranteed and the outcome of the test was still unclear(some revisions might need to be done midway). Therefore, a more conservative and controlled approach was adopted. For future research, assuming a sufficient number of sounds are available, this work suggests that conducting multiple tests with diverse stimuli could yield promising outcomes and contribute to the creation of a comprehensive dataset.

Also, the use of annoyance as the primary metric in this thesis was a deliberate choice. As a negative perceptual response, annoyance does not capture aspects related to comfort or ease of listening. Therefore, extending the framework to include positive perceptual dimensions—such as comfort, satisfaction, or perceived quality—could offer a more comprehensive and nuanced evaluation of component sound quality.

5.2.2 Presence of a bias

The process of unbiasing participant responses improved overall concordance without significantly affecting the underlying trends in the data. This indicates that, although some bias was present, it did not substantially distort perceptual judgments and supports the validity of the chosen rescaling method from an analytical perspective. In this instance, unbiasing proved unnecessary. However, this may not hold true in other studies. As such, bias correction should be evaluated on a case-by-case basis in future work.

5.2.3 Feature selection

Among the five primary psychoacoustic features (loudness, sharpness, modulation, tonality, and fluctuation strength), loudness emerged as the most reliable predictor, both statistically and perceptually. This is unsurprising given its intuitive and direct relationship with perceived annoyance. Modulation also played a key role, likely reflecting mechanical irregularities or unstable motor speeds that jurors were explicitly instructed to consider.

Metrics like sharpness and fluctuation strength, while relevant, displayed lower consistency and weaker correlations in this study. However, several previous studies have emphasized their significance in the evaluation of sound quality, especially in automotive and consumer product applications [16, 12]. This discrepancy suggests that the implementation or presentation of these metrics in the listening test may

have influenced participants' responses. For example, sharpness is a perceptual attribute that is notoriously difficult to isolate and rate, especially without extensive auditory training. It requires the listener to focus on high-frequency content and ignore loudness or modulation artifacts. This is an inherently non-intuitive task for non-expert participants

This observation does not negate the potential relevance of sharpness or fluctuation strength but highlights the critical role of perceptual clarity and interface design in jury testing. With improved metric communication and test calibration, their influence on perceived annoyance might be more accurately captured.

5.3 Applications and future work

From a practical standpoint, the random forest model's ability to offer transparent and interpretable predictions is a major advantage. In industrial contexts where engineers need to justify design decisions or trace anomalies, interpretability is often more valuable than marginal gains in accuracy. The success of the RF model suggests that sound quality assessments could be automated effectively, reducing reliance on time-consuming and subjective jury testing.

However, for cutting-edge product development or research into next-generation seat mechanisms, the higher precision of neural models may still be leveraged, especially if the data pool is expanded.

Furthermore, considering the availability of a Virtual Reality (VR) studio, future research could explore the use of immersive VR environments to simulate full vehicle cabins. This would allow for the study of how multimodal cues interact with auditory stimuli to influence the perceived annoyance.

6
Conclusion

The aim of this thesis has been to improve the prediction of subjective annoyance caused by seat adjustment sounds in vehicles. It has also aimed to evaluate the suitability of various regression and machine learning models to serve as predictive tools in the context of sound quality (SQ) assessments. By relating annoyance scores to objective psychoacoustic metrics, several models have been developed to estimate user perception without the need for repeated jury testing. These models were trained and validated using data from a carefully designed listening test and a processed dataset of seat adjustment sounds.

Four main approaches were implemented and compared: a multiple linear regression model, a multivariate polynomial regression model, a neural network, and a random forest. These models were evaluated both in terms of predictive accuracy and industrial interpretability. The random forest model yielded the most balanced results, showing good accuracy while maintaining transparency in decision-making. The linear regression model performed consistently and is suitable for baseline prediction and feature influence analysis. The neural network showed the potential to capture more subtle interactions but was prone to overfitting, especially given the limited dataset size. Its accuracy can be improved with a larger and more varied dataset, although its black-box nature limits its interpretability in practical applications. An important contribution of this work lies in the execution of a large-scale listening test, which involved 84 participants—an unusually high number for this type of psychoacoustic study. This strong empirical basis adds robustness to the model evaluation and reflects a high degree of methodological rigor. Additionally, the exploration of feature selection, unbiasing strategies, and perceptual metric correlations provides a solid framework that others in the field can build upon.

That said, the study is not without limitations. The relatively small number of distinct sound samples (33) constrains the possible generalization of the results, and complex perceptual metrics like sharpness or fluctuation strength may have been underrepresented due to their difficulty to explain and rate reliably. Future work should prioritize expanding the sound dataset and exploring advanced augmentation or synthetic generation techniques. Further improvements may also be realized through the use of multimodal platforms, allowing the ratings to be done in a more immersive environment. Additionally, the inclusion of more perceptual dimensions, like positive feelings, could enrich the model and align more closely with real customer impressions.

In conclusion, this thesis presents a structured and data-driven approach to predicting perceived annoyance in seat adjustment sounds. While the current models offer strong potential, their broader application across automotive sound design and quality assessment will depend on continued development, richer data sources, and deeper integration into perceptual testing environments. The methods and findings presented here provide a foundation for such future work and may extend to other components within the vehicle soundscape.

Bibliography

[1] E. Arvidsson, E. Nilsson, D. Bard-Hagberg, and O. J. I. Karlsson, *Subjective Experience of Speech Depending on the Acoustic Treatment in an Ordinary Room*, International Journal of Environmental Research and Public Health, 18 (2021), p. 12274.

[2] M. Bayani, C. Wickman, and R. Söderberg, *Analysis of sound characteristics to design an annoyance metric for rattle sounds in the automotive industry*, Int. J. Vehicle Noise and Vibration, 17 (2021), pp. 137–161.

[3] A. Benghanem, O. Valentin, P.-A. Gauthier, and A. Berry, *Objective quantification of sound sensory attributes in side-by-side vehicles using multiple linear regression models*, Frontiers in Acoustics, 2 (2024), p. 1477395.

[4] M. M. Bradley and P. J. Lang, *Measuring emotion: The self-assessment manikin and the semantic differential*, Journal of Behavior Therapy and Experimental Psychiatry, 25 (1994), pp. 49–59.

[5] L. Breiman, *Random Forests*, Machine Learning, 45 (2001), pp. 5–32.

[6] R. Budiu and K. Moran, *How Many Participants for Quantitative Usability Studies: A Summary of Sample-Size Recommendations*.

[7] J. Charbonneau, C. Novak, and H. Ule, *Validating a binaural head for use in jury testing*, Montreal, Canada, 2013, pp. 050163–050163.

[8] W. Ellermeier, M. Mader, and P. Daniel, *Scaling the Unpleasantness of Sounds According to the BTL Model: Ratio-Scale Representation and Psychoacoustical Analysis*, 90 (2004).

[9] W. V. Enigk H, *Subjective and Objective Evaluation of the Air Conditioning Sound*, Journal of Ergonomics, 04 (2014).

[10] H. Fastl, *Fluctuation strength and temporal masking patterns of amplitude-modulated broadband noise*, Hearing Research, 8 (1982), pp. 59–69.

[11] H. Fastl and E. Zwicker, *Fluctuation Strength*, in Psychoacoustics, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 247–256.

[12] ——, *Psychoacoustics : Facts and Models*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2nd edition ed., 2007.

[13] H. Fletcher and J. C. Steinberg, *Articulation Testing Methods*, Bell System Technical Journal (1929), pp. 806–854.

[14] Fletcher, Harvey and W. Munson, *Loudness, Its Definition, Measurement and Calculation*, The Journal of the Acoustical Society of America, V (1933), pp. 82–108.

[15] K. Genuit, *Possibilities of psychoacoustics to determine sound quality*, Jan. 1994, pp. 751–756. ADS Bibcode: 1994nce..conf..751G.

[16] ——, *The sound quality of vehicle interior noise: a challenge for the NVH-engineers*, International Journal of Vehicle Noise and Vibration, 1 (2004), p. 158.

[17] K. Genuit, *Objective Evaluation of Acoustic Quality Based on a Relative Approach*, in NOISE CONTROL - THE NEXT 25 YEARS, Liverpool, UK, Feb. 2024, Institute of Acoustics.

[18] A. Hasselström, *Benchmark work: correlation between subjective evaluation and objective measure of seat adjustment sound in various cars.*

[19] S. S. Haykin, *Neural networks: a comprehensive foundation*, Prentice Hall, Upper Saddle River, N.J, 2nd ed ed., 1999.

[20] Head Acoustics, *Application Note / Modulation Analysis*, (2018).

[21] ——, *Application Note / Tonality (Hearing Model)*, (2018).

[22] M. Huzaifah and L. Wyse, *Deep generative models for musical audio synthesis*, Nov. 2020. arXiv:2006.06426 [eess].

[23] International Organization for Standardization, *Acoustics - Method for calculating loudness level*, 1975.

[24] H. I. Jo, B. B. Santika, H. Lee, and J. Y. Jeon, *Classification of sound environment based on subjective response with speech privacy in open plan offices*, Applied Acoustics, 189 (2022), p. 108595.

[25] K. Koutini, H. Eghbal-zadeh, F. Henkel, J. Schlüter, and G. Widmer, *Over-Parameterization and Generalization in Audio Classification*, July 2021. arXiv:2107.08933 [cs].

[26] D. S. Kunte, *Sound Quality Prediction Using Neural Networks.*

[27] D. A. Laird and Coye, Kenneth, *Psychological Measurements of Annoyance as Related to Pitch and Loudness*, Journal of the Acoustical Society of America, (1929).

[28] S.-K. Lee, T.-G. Kim, and U. Lee, *Sound Quality Evaluation Based on Artificial Neural Network*, in Advances in Natural Computation, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, L. Jiao, L. Wang, X.-b. Gao, J. Liu, and F. Wu, eds., vol. 4221, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 545–554. Series Title: Lecture Notes in Computer Science.

[29] M. Li, *Comprehensive Review of Backpropagation Neural Networks*, Academic Journal of Science and Technology, 9 (2024), pp. 150–154.

[30] Lucas Ferreira-Paiva, Elizabeth Alfaro-Espinoza, Vinicius M. Almeida, Leonardo B. Felix, and Rodolpho V. A. Neves, *A Survey of Data Augmentation for Audio Classification*, Online, Oct. 2022.

[31] R. Lyon, *Designing for Product Sound Quality*, CRC Press, 0 ed., June 2000.

[32] I. D. Mienye and N. Jere, *A Survey of Decision Trees: Concepts, Algorithms, and Applications*, IEEE Access, 12 (2024), pp. 86716–86727.

[33] B. C. J. Moore, *An introduction to the psychology of hearing*, Emerald, Bingley, 6th ed ed., 2012.

[34] Mostafa Ibrahim, *Decoding Backpropagation and Its Role in Neural Network Learning | ml-articles – Weights & Biases*, 2025.

[35] M. Möser, *Engineering Acoustics: An Introduction to Noise Control*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.
[36] H. F. Olson, *Acoustic Laboratory in the New RCA Laboratories*, The Journal of the Acoustical Society of America, 15 (1943).
[37] A. Osses Vecchi, R. García León, and A. Kohlrausch, *Modelling the sensation of fluctuation strength*, Buenos Aires, Argentina, 2016, p. 050005.
[38] N. Otto, S. Amman, C. Eaton, and S. Lake, *Guidelines for Jury Evaluations of Automotive Sounds*, May 1999, pp. 1999–01–1822.
[39] H. Pettersson and O. Stensöta, *Classifying Brachycephalic Obstructive Airway Syndrome (BOAS) in Dogs*.
[40] Roland Sottek, *Product Sound Quality and Metric Development*, 2024.
[41] H. J. Sabine and R. A. Wilson, *The Application of Sound Absorption to Factory Noise Problems*, The Journal of the Acoustical Society of America, 15 (1943), pp. 27–31.
[42] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, *Trade-off between training and testing ratio in machine learning for medical image processing*, PeerJ Computer Science, 10 (2024), p. e2245.
[43] R. Sottek, *ECMA-418-2, 3rd edition, December 2024*, 2024.
[44] G. K. Uyanik and N. Güler, *A Study on Multiple Linear Regression Analysis*, Procedia - Social and Behavioral Sciences, 106 (2013), pp. 234–240.
[45] J. H. Ward, *Hierarchical Grouping to Optimize an Objective Function*, Journal of the American Statistical Association, 58 (1963), pp. 236–244.
[46] W. Yu, *The Estimation of Acoustic Parameters and Representations based on Room Impulse Responses*.

DEPARTMENT OF APPLIED ACOUSTICS
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY