U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Sound Quality Predictive Models: Annoyance in the Interior of a Propeller Aircraft

Bernardo Oliveira de Sá Lopes

Dissertation submitted to
Faculdade de Engenharia da Universidade do Porto
for the degree of:

Master in Mechanical Engineering

Supervisor at FEUP:
Prof. José Dias Rodrigues (Associate Professor)

Supervisors at Siemens PLM Software:
Eng. Claudio Colangeli (Researcher)
Eng. Giampiero Accardo (Researcher)

Departamento de Engenharia Mecânica
Faculdade de Engenharia da Universidade do Porto

Porto, 2018

The work presented in this dissertation was conducted
as a result of a partnership between FEUP and
Siemens PLM Software, in the company premises at
Leuven, Belgium

Bernardo Oliveira de Sá Lopes
E-mail: up201304484@fe.up.pt

Faculdade de Engenharia da Universidade do Porto
Departamento de Engenharia Mecânica
Rua Dr. Roberto Frias s/n, Sala M206
4200-465 Porto
Portugal

To my Family

A força sem a destreza é uma simples massa. Fernando Pessoa

Abstract

The assessment of a propeller aircraft interior noise often occurs in the late stages of its development cycle. This makes it difficult to intervene to improve the resulting noise characteristics because many design parameters have been already fixed. This implies that frequently the interior noise of an aircraft is not optimized with respect to passenger comfort and reduction of annoyance. To improve these aspects, the design approach should shift to a human-centered paradigm. This is addressed by conducting a jury study to collect subjective evaluations of cabin sounds, which, combined with psychoacoustic features, are used to train and compare different traditional and Machine Learning feature-based prediction techniques. These models predict the subjective evaluation of a sound sample from its corresponding several psychoacoustic metrics, thus mimicking the human Sound Quality perception. However, the psychoacoustic feature extraction from a sound sample is still a *manually* performed step, through complex psychoacoustical algorithms, that demand the expertise of an experienced acoustic engineer. The recent breakthroughs in the field of Deep Learning allow the use of Convolutional Neural Networks to train a compact prediction model which is able to extract the psychoacoustic features from a sound sample. Sequentially combining this feature extractor prediction model with one of the feature-based models developed with the data from the jury studies enables the development of a *Virtual Passenger* model, that, from a sound sample is capable of directly predicting the subjective human response to it. Hence, the model obtained simulates the passenger subjective response to the stimuli in the different positions of the cabin and, combined with virtual prototyping and sound synthesis tools, paves the way for the inclusion of the human Sound Quality perception in the early phases of the propeller aircraft design process.

Keywords: Psychoacoustics, Sound Quality, Propeller Aircraft, Multiple Linear Regression, Machine Learning, Artificial Neural Networks, Support Vector Machines, Random Forests, Deep Learning, Convolutional Neural Networks, Virtual Passenger

Resumo

A avaliação do ruído interior de uma aeronave a hélices ocorre frequentemente nas fases finais do seu projeto e desenvolvimento, impossibilitando uma intervenção nas características do ruido resultante, uma vez que, a maioria dos parâmetros de *design* já se encontram fixos. Consequentemente, o ruído interior das aeronaves a hélices habitualmente não se encontra otimizado no que diz respeito ao conforto acústico dos passageiros e *annoyance reduction*. De forma a melhorar estes aspetos, o projeto aeronáutico deve procurar evoluir no sentido de também contemplar a perceção humana da Qualidade do Som. A realização de *Jury Testing* permite a recolha de avaliações subjetivas de sons do interior da cabina de uma aeronave a hélices que, combinados com as respetivas *features* psicoacústicas, são usados para treinar e comparar tanto modelos de previsao tradicionais como também modelos assentes em *Machine Learning*. Contudo, o processo de extração de *features* psicoacústicas continua a ser um processo *manual*, feito com base em algoritmos acústicos complexos e que requer o envolvimento de engenheiros acústicos experientes. Os recentes avanços das técnicas de *Deep Learning* permitem o uso de Redes Neuronais Convolucionais para treinar modelos de previsão compactos, capazes de extrair *features* psicoacústicas a partir de amostras de som. O combinar sequencial deste modelo de extração de *features* com um dos modelos capazes de, a partir de *features*, prever uma avaliação subjetiva da Qualidade do Som, possibilita a criação de um modelo denominado por *Passageiro Virtual*. Este é capaz de, a partir de uma amostra de som, modelar diretamente a perceção humana da qualidade sonora desta, sendo que, quando combinado com ferramentas de prototipagem virtual e de síntese de sons, abre o caminho para a inclusão da dimensão humana da Qualidade do Som no projeto aeronáutico.

Palavras-Chave: Psicoacústica, Qualidade do Som, Aeronave a Hélices, Regressão Linear Múltipla, *Machine Learning*, Redes Neuronais, *Support Vector Machines*, *Random Forests*, *Deep Learning*, Redes Neuronais Convolucionais, Passageiro Virtual

Acknowledgements

The work here present represents a build-up of knowledge and experiences over many years, achieved with the help of several individuals to whom I would like to express my acknowledgement.

First of all, to Faculdade de Engenharia da Universidade do Porto, my second home during the past 5 years but also a place where I grew as a person and as a future engineer. I would like to adress all the professors and staff that everyday contribute to keeping FEUP as one of the most prestigious and respected schools in the country. Also to professor José Dias Rodrigues, who was my internal supervisor and awoke my curiosity to the field of mechanical vibrations.

To Siemens PLM Software, in the person of my supervisor Eng. Claudio Colangeli who provided the means for this work and since day one guided into this new home. To Eng. Giampiero Accardo, who relentlessly supervised my work and helped me to discover new fields of knowledge. Last but not least, to all the new friends I met in the Test Division office, who taught me even the most ordinary things can escalate into greatness.

To the friends I made over the last 5 years at FEUP I would like to leave a special acknowledgement. Indeed a masters degree at FEUP was undoubtedly an excellent academic opportunity, but the evolution I had as an individual and as a part of something bigger is perhaps what I will remember the most over the next years and that would not have happened without the incredible people I had the opportunity to meet. Hence, among many others, I would like to acknowledge William Milner, Luís Alves, Ricardo Alves, João Moreira, João Varela, Rafael Vieira, Kevin Cardoso, Diogo Costa, João Costa and Pedro Silva.

To Francisco Guimarães and Luís Costa for more than two decades of friendship.

To my parents, Bernardo and Filipa, for the unconditional support, the affection, comprehension and love. I owe them this opportunity and the person I am today. Also to my grandparents and the rest of my family, for being there all these years and helping me grow.

To those who could not see me achieving this goal, taught me to walk and inspired me to run.

Contents

List of Figures

List of Tables

List of Acronyms

ML Machine Learning

NVH Noise, Vibration and Harshness

SQ Sound Quality

MLR Multiple Linear Regression

ANN Artificial Neural Network

SVM Support Vector Machine

RF Random Forest

CNN Convolutional Neural Network

VP Virtual Passenger

CASTLE CAbin Systems design Toward passenger welLbEing

LM Levenberg–Marquardt

BR Bayesian Regularization

GPU Graphics Processing Unit

SPL Sound Pressure Level

MAE Mean Absolute Error

RMSE Root Mean Square Error

Chapter 1

Introduction

1.1 Sound Quality and Aircraft Design

Sound is one of the main dimensions of human communication. With the advance of technology, people often take for granted the human sound quality perception, even thought acoustic communication represents a cornerstone of modern society.

The successful design of new products mainly relies on the capability of assessing the performance of alternative concepts already in an early stage [Oliveira et al., 2009]. In the latest decades, due to the shortening of flight distances and simultaneous increase of passenger flight frequency, the available virtual prototyping tool for the aeronautic industry have seen major improvements. However, when dealing with passenger acoustic comfort, these tools mainly focus on reducing the Sound Pressure Level (SPL) inside the aircraft cabin and fail in understanding and quantifying the human perception of cabin noise. Thus, there is a need for a greater emphasis on improving the passenger acoustic comfort instead of merely suppressing cabin noise [d'Ischia et al., 2001; Duvigneau et al., 2016].

Cabin Noise and Vibration in a Propeller Aircraft

In an aircraft, the cabin noise is mainly created by its propulsion system, and turbulent boundary layer pressure fluctuations, which excite the fuselage, causing it to vibrate thus creating a sound field in the interior of the cabin. In a propeller aircraft, the main source for the inner cabin noise is the airborne noise produced by the propellers, due to their blades in the fuselage. Therefore, periodic pressure fluctuations are produced on the external part of the fuselage, which in turn causes vibrations in the internal cabin walls and an excitation of the interior cabin sound field [Wilby, 2008].

The assessment of a turboprop aircraft interior noise frequently occurs late in its design cycle. This makes difficult to intervene to improve the resulting noise characteristics since several design parameters have already been fixed. This implies that the cabin interior in propeller aircraft's is not properly optimized with respect to passenger comfort and reduction of annoyance [Janssens et al., 2008].

Traditionally, the target vibroacoustic performances are defined in terms of maximal noise levels and design targets are propagated to maximal noise contributions from the main noise sources. Although the design process considerably benefits from several advanced numerical, experimental and hybrid Noise, Vibration and Harshness (NVH) modeling tools, the sound quality aspect is still missing in the whole engineering process [Janssens et al., 2008]. To improve these aspects, the design approach should shift to a human-centered paradigm. By using virtual aircraft prototyping and virtual sound synthesis, it

is possible to conduct a multi-attribute optimized design process, through which the assessment of the subjective resulting from a design variant can be fed back to the virtual prototype for further optimization.

Sound Quality Prediction

Sound Quality (SQ) refers to the subjective perception of a product based on emitted sound in terms of the functionality (perceived build quality) or preference (annoyance or pleasantness), being that SQ analysis consists on predicting subjective preferences based on objective measurements [Pietila and Lim, 2012]. The sound quality of a vehicle interior noise is influenced by three variables: sound field, auditory perception and auditory evaluation, which causes the evaluation of sound quality to be a multidimensional task [Otto et al., 2001].

Even though A-weighted SPL is still a simple and popular acoustic metric, the perception of a sound is also dependent on its psychoacoustic characteristics such as loudness, roughness, fluctuation strength and other extended metrics. This is due to the fact that, for example, low frequency waves can be transmitted over a longer distance than high frequency sound waves on the human hear receptors. As a result, high frequency noise can be masked more easily by low-frequency noise, so the perceived properties of a sound are not identical to the respective emitted sound [Otto et al., 2001]. Therefore, SQ is used to relate human perception (psychoacoustics) with the physics of sound generation and transmission (vibroacoustics) [Oliveira, 2009].

Currently, the most common procedure to address this issue is to use a jury study to rank product sounds on a numerical preference scale and develop statistical models based on MLR to compute the subjective preferences using objective measurements, hence obtaining a SQ prediction model. However, the traditional use of a linear model has a few shortcomings due to the fact that the relationships between the human hearing process and acoustic performance are nonlinear [Pietila and Lim, 2012].

Alternate methods with higher computational cost such as ANNs, SVMs or RFs allow to replace the linear models with more adaptive models that, besides addressing the linearity concern, can model sub-groups inside the jury testing group. Hence, conducting jury studies allows to obtain subjective evaluation of sounds that, combined with objective metrics, can be the output of a SQ prediction model [Pietila and Lim, 2012].

Machine Learning and Testing

Currently, the features used in SQ prediction models correspond to psychoacoustic metrics (such as loudness or sharpness) that are obtained using complex algorithms based on international psychoacoustic standards. This computation is usually done using testing software that apply these algorithms to the SPL measurements from a microphone. Hence, the SQ prediction models predictor variables correspond to features *manually* extracted from raw time signals, using commercial software. To properly extract these psychoacoustic features, it is often required the expertise of an experienced acoustical engineer and commercial software with low capabilites of integration in an automatized model, which from a raw time signal can predict the human SQ subjective evaluation.

As stated by LeCun et al. [2015], conventional Machine Learning (ML) techniques were limited in their ability to process natural data in their raw form, requiring careful engineering and considerable domain expertise to design a feature extractor that transformed the raw data (such as the pixel values of an image or the pressure levels of a sound

sample) into a suitable internal representation or feature vector from which the learning subsystem, often a classifier, could detect or classify patterns in the input.

Deep Learning, a field contained in the ML domain, is making major advances in solving problems that have resisted the best attempts of the artificial intelligence community for many years. Since the early 2000s, Convolutional Neural Networks (CNNs) have been used with success and in 2012, due to progress derived from new techniques and the efficient use of Graphics Processing Units (GPUs), they started revolutionizing the computer vision and speech recognition industries. In opposition with the conventional Machine Learning techniques, CNNs are able to process data that comes in the form of multiple arrays, such as 1D arrays that correspond to sound samples or 2D arrays for images, where the pixel intensities are in three colour channels [LeCun et al., 2015].

The possibility of using CNNs for performing feature extraction in sound samples paves the way for entirely new approaches in predicting Sound Quality, being this still a notably uncharted domain, waiting for innovative methods to arise.

1.2 Motivations

The assessment of a propeller aircraft interior noise occurs often in the late stages of its development cycle. This makes it difficult to intervene to improve the resulting noise characteristics because many design parameters have been already fixed. This implies that very often the interior noise of an aircraft is not optimized with respect to passenger comfort and reduction of annoyance. To improve these aspects, the design approach should shift to a human-centered paradigm.

The development of a feature-based prediction model, a common approach in the literature, that is able to predict the human subjective assessment (output) of inner cabin sounds based on psychoacoustic metrics (input), allows to avoid recollecting human evaluations, thus being possible to combine it with virtual prototyping and sound synthesis tools. This approach enables the engineer to change a design parameter, synthesize the new resulting cabin sounds and after extracting their psychoacoustic features, to use the prediction model to predict subjective evaluations for the new synthesized sounds *annoyance*.

However this process still relies on a *manual* features extraction from the sound samples, that are the input of the prediction model. This extraction has to be done using commercial software that requires an experienced acoustic engineer to perform it and is hard to *automatize* in the design process. To develop another prediction model, using CNNs, that from the interior cabin sound samples compute the psychoacoustic metrics (features) consists in a new approach in this field, being this usually done using complex acoustic algorithms or commercial software.

To combine both the mentioned prediction models, as sequential blocks, allows to obtain an easily usable, fast and compact prediction model, that directly from a sound sample predicts the human subjective evaluation of a sound, thus being developed a VP model that mimicks the human SQ perception, opening a wide range of applications in the aircraft industry. With this in mind, some objectives were defined for this thesis which are enumerated as follows.

1.3 Objectives

- Understand the fundamentals of psychoacoustics and their connection with vibroacoustics, through different metrics, contacting with the state of the art in commercial

software.

- Conduct a campaign to collect subjective evaluations of sound samples corresponding to different seats of a propeller aircraft.
- Review the knowledge of feature-based prediction models, that perform regression-based predictions from features, namely MLR, ANNs, SVMs and RFs.
- Review the fundamentals for extracting features from time signals using CNNs.
- Train and compare 4 prediction models able to predict a subjective evaluation of sound samples based on their objective metrics. These are feature-based models, namely MLR, ANNs, SVMs and RFs. they are trained using the data obtained in the campaign destined to collect the subjective evaluations.
- Using CNNs, train models able to extract several different psychoacoustic feature directly from sound samples, namely loudness, fluctuation strength, tonality, sharpness and roughness.
- Assess the applicability of sequentially combining the best performing feature-based model with the models that are able to extract features from sound samples, obtaining the VP model that simulates the passenger of an aircraft, being then able to predict a subjective evaluation of a sound sample directly from a time signal.
- Study which features are more relevant to the performance of the VP model and understand the causes for its prediction errors.

1.4 Layout

This text is structured in eight chapters. The work contained in this thesis is inserted in an international research project and is a sequence of another dissertation, being the second chapter destined to cover this. The third and fifth chapters clear the fundamental theory on psychoacoustics and prediction models. On the fourth chapter the guidelines followed for collecting subjective evaluations of sound samples are presented. Then, on chapter six, both the results of collecting subjective sound samples evaluations and of the feature-based models are shown. The 4 models are compared, being evidenced the methods used for training and testing the models. On chapter seven, after presenting the results and method for extracting features (psychoacoustic metrics) using CNNs, the VP model performance is assessed, also including detailed information regarding the methods used. Finally, the conclusions of this work and further possible improvements of the created models are mentioned.

Chapter 2

Context and Prior Work

2.1 Introduction

This thesis is inserted in the framework of an international research project. Funded by the European Union Horizon 2020 programme, Clean Sky 2 is one of the largest European research programmes, present in 24 countries, developing innovative, cutting edge technology aimed at reducing CO2, gas emissions and noise levels produced by aircraft. In this programme, one of the developed projects is CAbin Systems design Toward passenger welLbEing (CASTLE), in which the work contained in this thesis is inserted [Sky, 2018].

CASTLE is devoted to achieve an improved and optimized passenger cabin environment in both regional aircrafts and business jets, increasing passenger's well being. This is achieved by using engineering models and virtual reality to develop and design cabin interiors with respect to human perception [CIRA, 2018].

This thesis, inserted in the framework of the CASTLE project, has as a goal the improvement of acoustic passenger comfort in a propeller aircraft, more specifically, a Dornier 228, shown in Fig. 2.1.

Figure 2.1: Dornier 228, a propeller aicraft intended for regional flights [Air, 2018].

Over the next paragraphs, it will be given a brief overview of the work of Angeloni [2018], which was the starting point for this thesis, being that the data mentioned here was used for the prediction models and was provided by this author.

About the aircraft it should be kept in mind that, throughout this thesis, two flying conditions will be specified. The case when both propellers have coincident frequencies it is denominated by *synchronous* and when the speed rotations are different, it corresponds

to the *asynchronous* case. Also, as noted by Angeloni [2018], one should note there are two main contributors for the cabin noise:

- The engines during their rotation generate vibration at particular frequencies denominated by *tones*, being that each propeller generates six tonal components that depend on the frequency rotation. In an asynchronous case, a *Beating* effect is generated during a complete flight cycle.
- Around the fuselage panels, small pressure fluctuations due to the *Turbulent Boundary Layer*, are responsible for the *Broadband Noise* in the interior of the aircraft.

2.1.1 Synthesis of interior noise sounds at 85 different positions

Siemens PLM Software, with the goal of developing sound synthesis tools, recorded the interior noise in a propeller aircraft, during normal cruising conditions. Both for synchronous and asynchronous conditions, the noise was recorded in 85 positions of the aircraft. The experimental set-up in the cabin used for recording the noises can be observed on Fig. 2.2.

Figure 2.2: Experimental set-up in the propeller aircraft [Janssens et al., 2008].

The recording positions are numbered and shown in Fig. 2.3. One should note that not every position corresponds to a real seat, being that, for example, microphones were set also in the middle of the corridor. Either way, for facilitating communication, throughout this thesis these positions will be referred to as seats, with the numbering scheme represented in Fig. 2.3.

Using the experimental data described in the previous paragraph, an algorithm was developed to, from a virtual model of the aircraft with changeable parameters, synthesize the sound sample corresponding to any position in the interior of the cabin, with the possibility of changing several design parameters of the virtual model. Therefore, it is possible to reproduce and study the interior noise in each typical propeller aircraft, without having to re-record sound samples in a flight, which would be not financially possible and extremely time consuming. Consequently all the sound samples used throughout this thesis, were synthesized by Angeloni, L. from a virtual model of a propeller aircraft, using the algorithm described in this paragraph. These samples were synthesized for both synchronous and asynchronous flying conditions, in 85 positions for each case, hence a total of 170 samples.

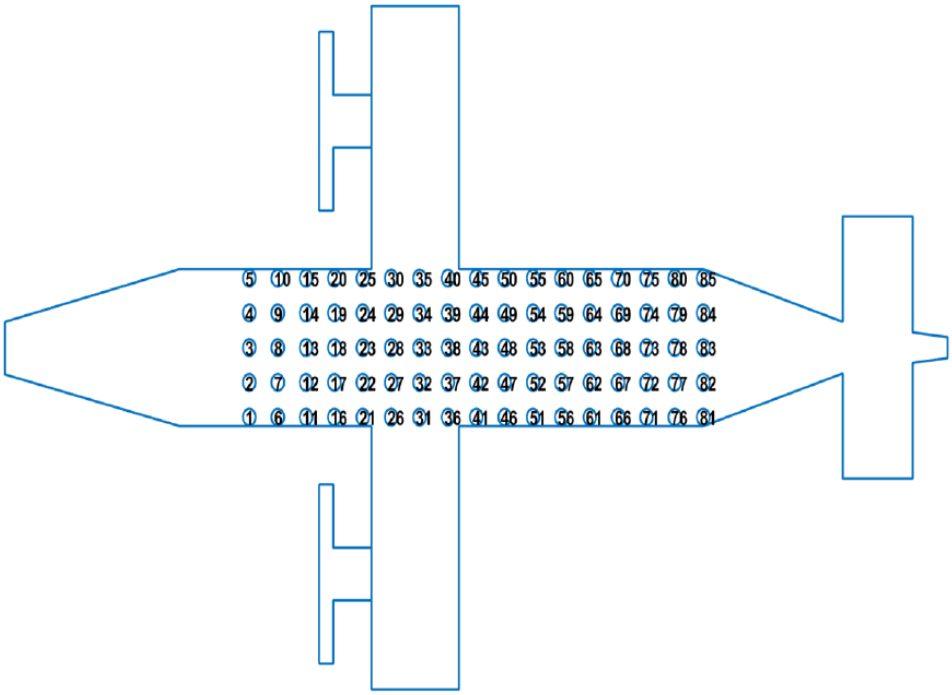

Figure 2.3: Position of each recording in the aircraft [Angeloni, 2018].

2.1.2 Psychoacoustic metrics

The psychoacoustic metrics correspondent to each sound sample that were used throughout this thesis were the ones computed in Angeloni [2018], using software provided by Siemens PLM Software, namely LMS Test.Lab. Using a colored map representation, where from the psychoacoustic metrics in each seat their values are interpolated for every point of the cabin, Angeloni L. obtained the distribution of each feature in the cabin, as shown in Fig. 2.4.

Finally, these metrics were the subject of a cluster analysis conducted in Angeloni [2018], that allowed in a high dimensional group of data find groups (i.e. clusters) with similar features. The information resultant from this procedure was fundamental in selecting the sound samples to use in the jury testing campaign, described in Ch. 4. Four clusters of seats were obtained for each flying case, being possible to find their representation in Fig. 2.5.

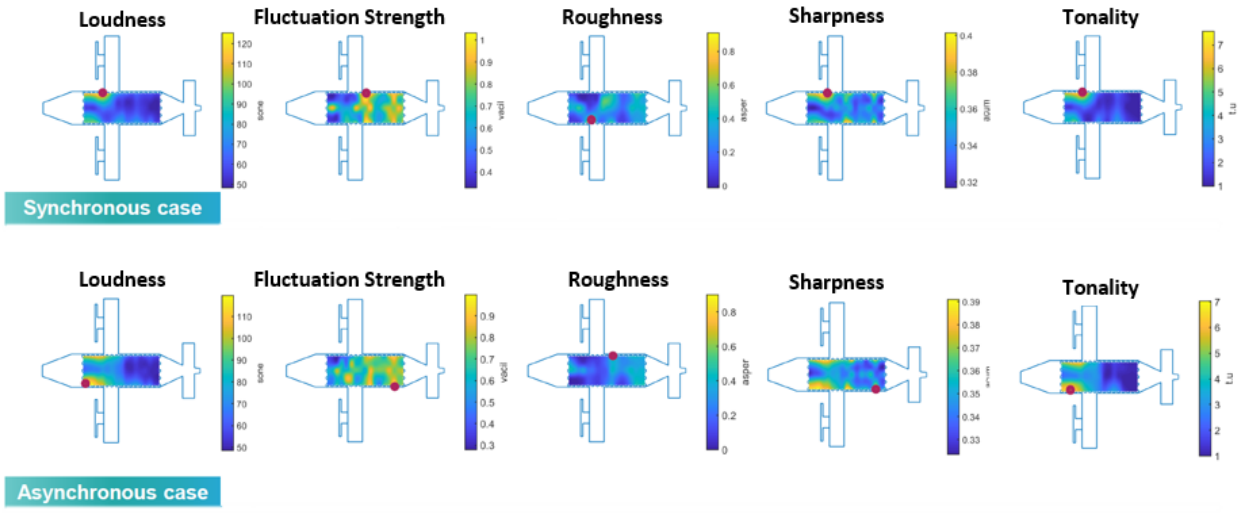

Figure 2.4: Distribution os the psychoaocutic metrics in the cabin of the aircraft. The red point corresponds to the maximum value of each metric [Angeloni, 2018].

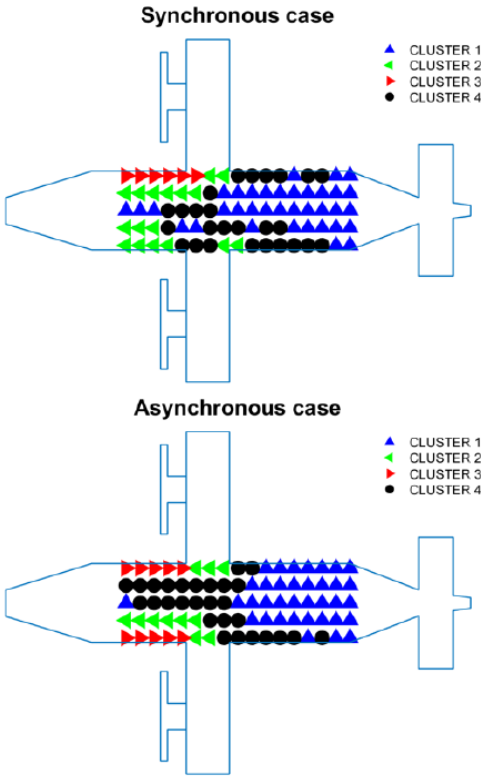

Figure 2.5: Distribution of the clusters in the fuselage [Angeloni, 2018].

Chapter 3

Psychoacoustics Fundamentals

3.1 Introduction

The characteristics of a sound, as it is perceived, are not exactly the same as the characteristics of sound being emitted. However, the understanding of auditory comfort requires first an understanding of sound.

An oscillatory motion or vibration of an object results in a sound. This motion is impressed upon the surrounding medium (such as air, solid, liquid or other gas), as a pattern of fluctuations in pressure (acoustical oscillations). It is observable that sound consists in a mechanical disturbance of the steady pressure in an elastic medium (usually air) which is propagated in all directions at a velocity of about 335 m/s [Quehl, 2001].

Sounds are easily described by means of the time-varying sound pressure, $p(t)$, and sound levels can be quantified either by their intensity (where the flow of energy is described through a unit of area) or amplitude (pressure). The temporal variations in sound pressure caused by sound sources, when compared with the magnitude of the atmospheric pressure, are extremely small. In psychoacoustics, values of the sound pressure between $10^{-5}Pa$ and $10^{2}Pa$ are relevant. Due to the wide range of sound pressures, the SPL is normally used. Eq. (3.1) relates sound pressure and SPL:

$$SPL = 20log(p/p_0) \quad dB \tag{3.1}$$

where the reference value of the sound pressure, p_0, is standardized to $p_0 = 20\mu Pa$, which is often considered as the threshold of human hearing [Fastl and Zwicker, 2007].

Additionally, other relevant physical quantities in psychoacoustics, besides sound pressure and SPL, are the sound intensity, I, and sound intensity level. In plane traveling waves, the SPL and sound intensity level are related by

$$SPL = 20log(p/p_0)dB = 10log(I/I_0) \quad dB \tag{3.2}$$

considering that the reference value, I_0, is defined as $10^{-12}W/m^2$. It is possible to realize that sound intensity is proportional to the square of sound pressure [Fastl and Zwicker, 2007].

The perception of sounds by the human ear

A key factor in explaining why two sounds with an equal dB level may have a totally different subjective quality is related to the physics of the human hearing process. The human ear is a complex, nonlinear device, with specific frequency dependent transition characteristics. In addition, the fact that hearing usually involves two ears (binaural) has

a considerable influence on sound perception. Before reaching the eardrum, an incident acoustic signal is considerably modified by the spectral and spatial filtering characteristics of the human body and ear. The human torso itself acts as an directional filter through diffraction, resulting in the fact that very significant interaural differences in SPL occur depending on the direction of the source [LMS, 2016b].

Consequently, the body, head and outer ear effects consist mainly of a spatial and spectral filtering that are applied to the acoustic stimulus. As a result, the analysis of, for instance, the frequency spectrum of a free positioned microphone does not necessarily lead to a correct assessment of the human response. In other words, there is no simple relationship between the measured physical sound pressure level and the human perception of the sound [LMS, 2016b].

Considering the dynamic range of sound intensities as the difference between the absolute threshold of hearing and the threshold of discomfort (more specifically pain), the hearing range is immense, as can be seen in Fig. 3.1, which justifies the use of a logarithmic scale for the sound pressure level and sound intensity level. In fact, this range is a function of frequency. At about 4000 Hz, it is approximately 125 to 135 dB but at lower and higher frequencies it is considerably less, for instance, 80 to 90 dB at 100 Hz. After a careful analysis of Fig. 3.1, it can be said that the best sensitivity of the human auditory system lies between 500 Hz and 5 kHz, being this area of the utmost importance for understanding human speech. Also, the thresholds of discomfort and pain represent estimates of the upper limit of sound level that humans can tolerate. In addition, the threshold of discomfort is around 110 to 120 dB SPL and the threshold of pain is about 120 to 140 dB SPL. Also, both remain relatively unchanged as a function of the frequency content of the stimulus.

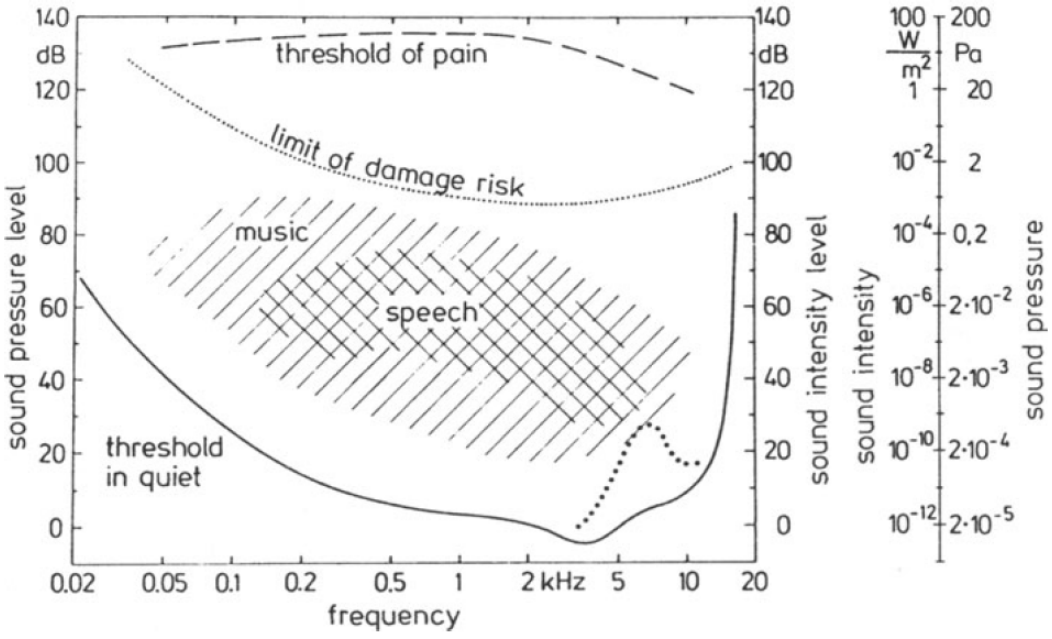

Figure 3.1: The threesholds of hearing, discomfort and pain [Fastl and Zwicker, 2007].

Auditory sensations can be described in terms of the so-called psychoacoustic parameters and are usually separately distinguishable. This quantities have in common an underlying concept of a correct representation of perceptual properties of the human auditory

system. Therefore, psychoacoustics deal mainly with the relationships between physical features of an acoustical stimulus and the auditory sensations evoked. The most relevant psychoacoustical parameters are: loudness, sharpness, tonality, fluctuation strength and roughness [Quehl, 2001].

3.2 Loudness

Loudness belongs to the category of intensity sensations, being not only a sensation value, but belonging somewhere between a sensation and physical value. The sound pressure level is not linearly related to the auditory impression of sound strength (or loudness). Along with frequency dependencies, this means that the loudness sensation cannot be accurately described by the acoustic level or its spectrum [LMS, 2016b].

The Loudness Level, L_N, is expressed in *Phons*. 1 kHz-tones are used as the reference, which means that for a 1 kHz tone, the *Phon* value corresponds to the dB sound pressure level. The *Phon* is a unit which derives from equal loudness contours (isophones), represented in Fig. 3.2. In this figure are shown several curves representing levels of *perceived* equal loudness (for sinusoidal tones) across a frequency range as a function of acoustic pressure level. In an isophone, the similar loudness of a reference (a pure 1000 Hz tone at a different sound level) will be experienced at all points (tones at different frequencies) along each contour.

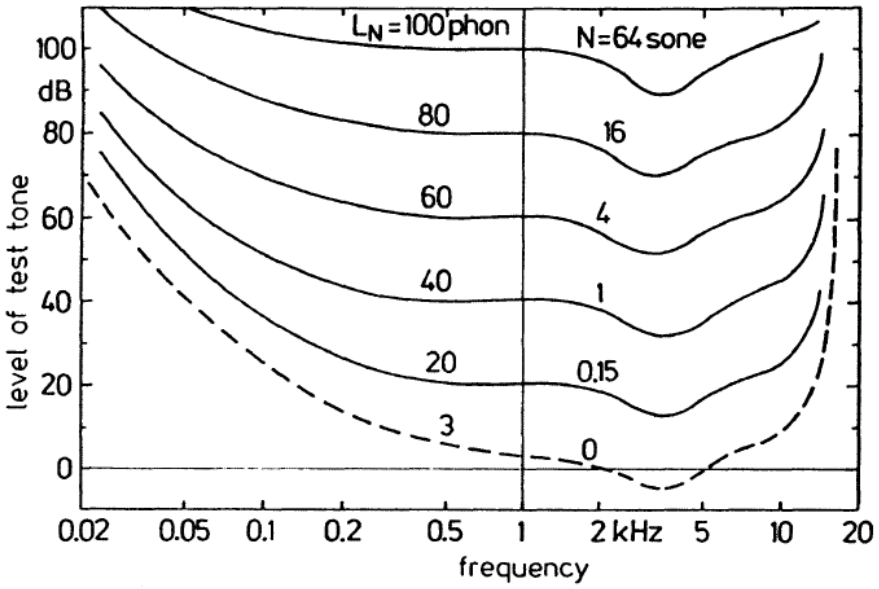

Figure 3.2: Equal-Loudness contours for pure tones in a free sound field [Fastl and Zwicker, 2007].

The *Sone*, S, is a linear unit derived from the logarithmic *Phon* values. Eq. (3.3) allows to relate both units,

$$S = 2^{(P-40)/10}. \tag{3.3}$$

The use of the *Sone* scale helps interpreting the experience loudness, due to the fact that it is a linear scale. A loudness level of 40 *Phons* corresponds to a loudness of 1 *Sone*.

A tone which is twice as loud, will have double the loudness in (*Sone*) value, and a loudness level which is 10 *Phons* higher [LMS, 2016a].

In various psychoacoustic problems, it is often useful to subdivide the frequency range over which the human ear is able to perceive tones and noises, instead of a linear or geometrically (logarithmically) division [Fastl and Zwicker, 2007]. Therefore, the audible spectrum is divided in frequency ranges, defined as *Critical Bands*. These bands have been directly measured in experiments on the threshold for complex sounds, on masking, on the perception of phase, and most often on the loudness of complex sounds. The frequency range used for this division corresponds to a Bark, being that different Barks have different frequency ranges [Fastl and Zwicker, 2007].

This scale is based on the fact that our hearing system analyses a broad spectrum into parts that correspond to critical bands. Adding one critical band to the next in such a way that the upper limit of the lower critical band corresponds to the lower limit of the next higher critical band, leads to the scale of critical-band rate. In total we have 24 bark and each of them correspond to a different frequency range. The center frequency and bandwidth for each critical band can be found on Tab. 3.1.

Table 3.1: Critical band rate (Bark), center frequency and bandwith for each critical band [Fastl and Zwicker, 2007]

Bark	Center Frequency (Hz)	Bandwidth (Hz)
1	50	100
2	150	100
3	250	100
4	350	100
5	450	110
6	570	120
7	700	140
8	840	150
9	1000	160
10	1170	190
11	1370	210
12	1600	240
13	1850	280
14	2150	320
15	2500	380
16	2900	450
17	3400	550
18	4000	700
19	4800	900
20	5800	1100
21	7000	1300
22	8500	1800
23	10500	2500
24	13500	3500

In psychoacoustic research, there are several definitions of loudness for complex sounds, however for this thesis framework, only Zwicker's Loudness was considered. This type of Loudness assessment was chosen due to its ability to deal with complex broadband noises, such as the ones in the interior of a propeller aircraft [LMS, 2016a].

Zwicker Loudness

As it was stated in the previous section, this method is capable of dealing with complex broadband noises, including pure tones. It is also relevant to point out that this method takes *masking* effects into account. These, are important for sounds composed of multiple components.

In the case of two sounds close in frequency, a high level sound component may mask another lower level sound which is too close in frequency. This effect may be found recurrently in music, where one instrument may be masked by another if one of them produces high levels while the other remains faint. If the loud instrument pauses, the faint one becomes audible again [Fastl and Zwicker, 2007]. Therefore, masking may be seen as the variation of the hearing threshold curve to a test sound in the presence of a masker, i.e., if the test sound spectrum lies below the masked threshold it will be inaudible. While inaudible, it would still be computed by a SPL-meter, which raises the need of a proper way of assessing the human perception of the sound, hence the use of Zwicker's Loudness [Oliveira, 2009].

According to Oliveira [2009] ,the first step in the numerical process consists of filtering the signal with critical band filters, followed by a masking check. In this stage, if the proceeding band level falls under the masking curve of the preceding one, this value is neglected; otherwise the value is kept.

Taking into account the method described in LMS [2016a], this masking check is achieved through the application of a sloping edge filter. This way, dominant and hence masking frequency bands will show their influence over a large frequency range and prevent masked sounds contributing to the total level. The partial loudness contours are computed for each defined segment (global evaluation) or frame (tracked evaluation) using a classical Zwicker loudness calculation. This is achieved by the use of excitation level versus critical-band rate pattern as a basis from which the loudness of the complex may be constructed, where total loudness is treated as an integral of a value that we have to find, but which can be drawn as a function of critical-band rate [Fastl and Zwicker, 2007].

Therefore, the physical quantity of loudness over the critical band rate, which the computation was described over the two last paragraphs, is defined as specific loudness, N', and has the unit of *Sone/bark*. Eq. (3.4), shows the mathematical expression for the computation of loudness.

$$N = \int_0^{24Bark} N' dz \tag{3.4}$$

where, total loudness, N, is obtained through the integral of specific loudness over the critical-band rate, thus *Sone* being the unit obtained [Fastl and Zwicker, 2007].

3.3 Sharpness

Another salient feature of auditory stimuli is the perceived sharpness, which allows to classify sounds as *shrill* (sharp) or *dull*. Sounds with a great share of high frequency components in the spectrum are perceived as sharp (for example, a piece of chalk scrapping a blackboard). According to Fastl and Zwicker [2007] and LMS [2016a], the most important parameters influencing sharpness are the spectral content and the centre frequency of narrow-band sounds. It is not dependent on loudness level or the detailed spectral content of the sound.

Roughly, it corresponds to the first spectral moment of the specific loudness, with a pre-emphasis for higher frequencies. In order to give quantitative values, a reference point

and a unit have to be defined. In Latin, the expression *Acum* is used for sharp. The reference sound producing 1 *Acum* is a narrow-band noise one critical-band wide at a centre frequency of 1 kHz having a level of 60 dB.

The dependency of sharpness on the center frequency and bandwidth of the noise is shown in Fig. 3.3. The middle curve represents a noise of one critical bandwidth as a function of center frequency, the upper and lower curves representing the sharpness of noises with respect to fixed upper (10 kHz) or lower (0.2 kHz) cut-off frequency as a function of the other cut-off value. Higher frequency noises produce higher sharpness [Quehl, 2001].

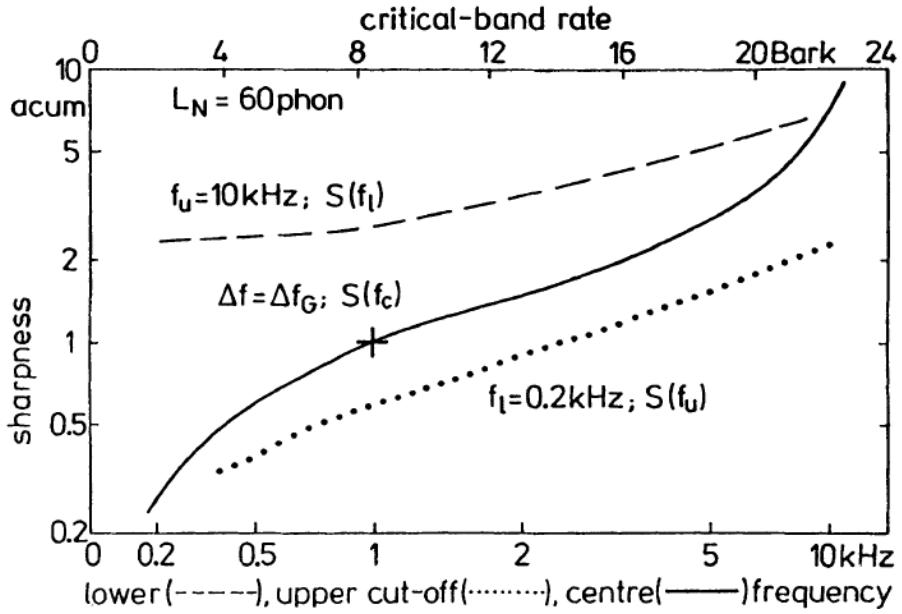

Figure 3.3: Sharpness of critical-band wide narrow-band noise as a function of centre frequency (solid), of band-pass noise with an upper cut-off frequency of 10 kHz as a function of the lower cut-off frequency (broken) and of band-pass noise with a lower cut-off frequency of 0.2 kHz as a function of the upper cut-off frequency (dotted) [Fastl and Zwicker, 2007].

In Fastl and Zwicker [2007], it is shown that the overall spectral envelope is the main factor influencing sharpness and that the spectral envelope is psychoacoustically represented in the excitation level versus critical-band rate pattern, or in the specific loudness versus critical rate band pattern.

After consulting LMS [2016a], it can be seen that sharpness corresponds to the first spectral moment of the specific loudness and according to Quehl [2001] it is dependent mainly on the center of gravity of the spectral distribution of a sound. The higher the frequency of its location, the sharper the sound is perceived.

Adopting an approach similar to the one used for loudness, a mathematical model for sharpness can be defined by computing the sharpness over the critical band, *i,e*, the specific sharpness, $S'(z)$. The unit corresponding to this physical quantity is *Acum/bark* and it is mathematically defined by Eq. (3.5) [LMS, 2016a].

$$S'(z) = \frac{0.11 N'(z) g(z) z}{\sum_{0Bark}^{24Bark} N'(z) \Delta z} \tag{3.5}$$

In the previous equation, $N'(z)$ represents the specific Zwicker Loudness and $g(z)$ a weighting function that pre-stresses higher frequency components [LMS, 2016a].

The total sharpness, S, expressed in *Acum* is obtained by integrating the specific sharpness, as defined in Eq. (3.6) [LMS, 2016a].

$$S = \int_0^{24Bark} S' dz \tag{3.6}$$

3.4 Fluctuation Strength

As can be read in Quehl [2001], fluctuation strength describes the degree of perceived fluctuation of the sound level, or irregularity versus even character of the sound that may arise due to the frequency and amplitude modulation from 1 to 20 Hz. When the sound functions have modulation frequencies below 20 Hz, they are perceived as changes in the sound volume over time. Typically, fluctuation signal sounds are louder (and more annoying) than steady state signals of the same amplitude. The unit to address the intensity of the sensation of fluctuation is referred to as *Vacil*. A reference sound of 1 *Vacil* corresponds to a 1 KHz tone of 60 dB with a 100% amplitude modulation of 4 Hz. Additionally, the ear is most sensitive to fluctuations at 4 Hz.

In the proposed quantitative models for fluctuation strength, F, it is necessary to take into account the temporal masking effects due to the sound fluctuation. Defining modulation frequency as f_{mod} and masking depth as ΔL, the dependency of fluctuation strength on these two variables is expressed in Eq. (3.7) [LMS, 2016a].

$$F \approx \frac{\Delta L}{(f_{mod}/4Hz) + (4Hz/f_{mod})} \tag{3.7}$$

In Fig. 3.4, the hatched part corresponds to the modulated signal in dB (level L_T). The black curve is a sinusoid-like curve on which the masking depth and the modulation frequency can be measured to calculate the fluctuation strength, according to Eq. (3.7) [LMS, 2016a].

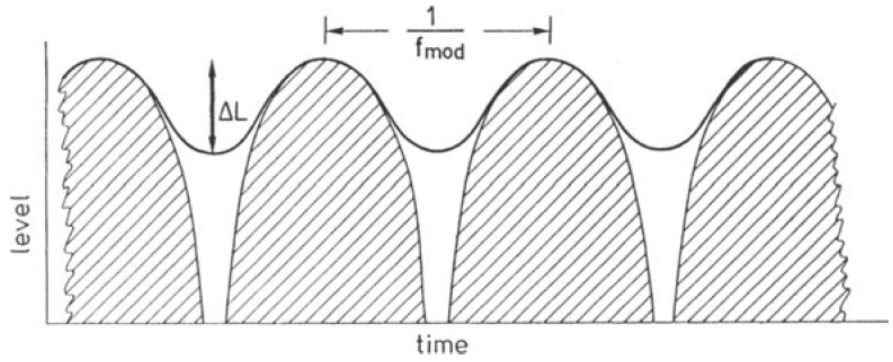

Figure 3.4: Ilustration of Fluctuation Strength [Fastl and Zwicker, 2007].

3.5 Roughness

The roughness or harshness of a sound is a consequence of the amplitude modulation of tones. As stated in Fastl and Zwicker [2007], frequency resolution and temporal resolution

of our hearing system are the two main factors that influence roughness, where frequency resolution is modeled by the excitation pattern or by specific-loudness versus critical-band rate pattern.

Low modulation frequencies (15 Hz), allow the perception of the time varying loudness fluctuations. This corresponds to fluctuation strength and has been analyzed in section 3.4.

For high modulation frequencies (above 150-300 Hz), three separate tones can heard, and in an intermediate frequency range (15-300 Hz), the sensation is of a stationary, but rough tone, which renders it quite unpleasant. For example, in the case of engine noise, this modulation effects can be caused by fractional orders [LMS, 2016a].

The increase of the degree of modulation and modulation frequency results in the increase of the perception of roughness, being less sensitive to the base or carrier frequency. However the dependency relationship between modulation depth and frequency is not straightforward. As shown in [LMS, 2016a], the roughness, R, of an amplitude modulated sound can be approximated as

$$R \cong f_{mod} \Delta L \tag{3.8}$$

The unit used to describe roughness is the *Asper*, being that 1 *Asper* is produced by a 100%, 70 Hz modulated 1 kHz tone of 60 dB.

The model used by LMS Test.Lab for quantifying roughness is quite complex, due to consideration of the masking effects. The algorithm involves the calculation of *partial or specific roughness* in each critical band, based on modulation and depth, including masking effects and integrating them to obtain total roughness [LMS, 2016a].

3.6 Tonality

Tonality is a metric which quantifies the tonal prominence of a sound. Its objective is to evaluate the presence or not of tones in the spectrum of a noisy sound, being a frequency dependent function. In a noise spectrum, a maximum tonality impression corresponds to a tone contribution at around 700 Hz. The smaller the bandwidth, the more tonal the noise seems [LMS, 2016a]. Sounds with single prominent tones are usually very annoying, even though these tonal contributes do not represent a significant contribute to the overall loudness [Quehl, 2001].

The quantification of tonality used uses the method developed by Terhardt for pitch extraction [LMS, 2016a].

First,the spectral lines that are at least 7 dB higher than their two lower and higher neighbors are isolated. A new spectrum, free of tonal components, is built by removing the detected sequences of five spectral lines, considered as pure tones. From both spectra, the fraction of the total loudness due to tonal components is calculated. This is denoted by W_n [LMS, 2016a].

Also, an extra weighting function, W_T, is derived from the pitch weights of the tonal components relevant to the pitch perception. At at 700 Hz the perception of tonality is maximal, being this a frequency dependent function [LMS, 2016a].

Additionally, a constant value, C is added to scale the tonality results to standardize the result, i.e, such that a 1 kHz sine tone at 60 dB gives a tonality of 1 $T.u.$ (tonality unit). Considering the combination of the three functions described in the previous paragraphs, Eq. (3.9) is used for finally computing the tonality, K [LMS, 2016a].

$$K = C \cdot W_T^{0.29} \cdot W_N^{0.79} \tag{3.9}$$

Chapter 4

Subjective Evaluations of Annoyance in a Propeller Aircraft

4.1 Introduction

According to Quehl [2001], both subjective (psychological) and objective (acoustic and psychoacoustic) factors determine the perception of sound events to cognitive and affective processes influencing the perception, interpretation, evaluation and reaction to auditory stimuli need to be considered in addition to acoustic and psychoacoustic parameters.

Annoyance, caused by sounds or noise, can be defined as displeasure due to sound exposure that affects health and well-being by its physical presence. Thus, it results from unwanting, interfering or disturbing acoustic waves and represents a subjective evaluation of a sound [Rainer, 1997].

Annoyance can be measured by means of questionnaire items with appropriate response scales. Generally, subjects are requested to listen to sounds or remember sounds and to use a numeric and/or verbal frame of reference which has been tested for its numeric properties as well as for the applicability, reliability and validity. It should be noted that individual human subjects in psychoacoustic laboratories still have their individual history and may use even common language in a slightly different way than the experimenters intend. Therefore, when conducting jury testing, one should make sure that the individual subjects share an understanding of the intended meaning [Rainer, 1997].

The activity of collecting subjects responses is designated by *jury testing*. The development of the several sound quality prediction models in the following chapters requires both the objective psychoacoustic metrics and the subjective data obtained through jury testing. In this chapter, first are presented some guidelines followed for performing the jury evaluations (section 4.2). Secondly, in section 4.3, it is possible to find information regarding the jury evaluation method/interface used. Finally, Ch. 6 contains the results of the subjective evaluations performed.

4.2 Jury Testing Guidelines

Subjective testing and analysis involves: presenting sounds to listeners, request judgment of those sounds from the listeners and perform statistical analysis on the responses. Jury testing is simply subjective testing done with a group of persons, rather than one person at a time. For achieving valid results while conducting jury testing, the following details should be taken into account, as mentioned in Otto et al. [2001]:

- Listening Environment
- Subjects

- Sound Samples Preparation
- Jury Evaluation Method
- Test Preparation and Delivery

The jury testing campaign was conducted at Siemens PLM Software premises in Leuven, Belgium. For the campaign, a meeting room at the company building was used. The subjects used headphones (model Sennheiser HD600, connected to an amplifier regulated by a computer), being also necessary to take the room acoustics into account. Due to several constraints related with materials and the available sound synthesis tools, only monaural sounds were considered for this activity. Each jury testing session had no more than 6 juror at a time. The outside ambient noise was kept to a minimum. Also, an effort was made in making the space as comfortable and inviting as possible, keeping in mind that, the more clinical the room looks, the more apprehension and anxiety the subjects will experience. Comfortable chairs and moderate lighting were ensured in order for the subjects to focus on the task at hand [Otto et al., 2001]. In Fig. 4.1, the adopted set-up for the jury testing room can be observed.

Figure 4.1: Jury testing room set-up.

The subjects who took part in the evaluation of sounds, also known by *jurors*, were internally selected in the company. The jurors were either foreign master students doing research work for their thesis or research engineers, working at the company. Due to the fact that the great majority of subjects were young, they were not subjected to audiometric tests before participating in the evaluation of sounds. Considering that, as stated by Otto et al. [2001], as a general rule, it is desired that the listening experience level of subjects is appropriate to the task at hand as well as representative of the target customer, all the subjects had already flew on a aircraft more than once. In Fig. 4.2 it is possible to find further information on the flying frequency of the jurors.

Additionally, the number of jurors chosen is relevant to the results. This decision is greatly influenced by whether extensive subject training is required as well as by the difficulty of the task. Taking into account that the task at hand, i.e, to classify the annoyance

of aircraft sounds, requires almost no training and also that knowing the distribution of the subject responses *a priori* is, in an industrial time-constrained context, not possible, the option taken was to follow the recomendation found in Otto et al. [2001]. Therefore, 25 to 50 was considered as an appropriate number of subjects for listening studies which use company employees as subjects, considering also that a small percentage will be automatically removed due to poor performance. After contacting several colleagues in the company, a total of 40 jurors were raised for the jury testing campaign.

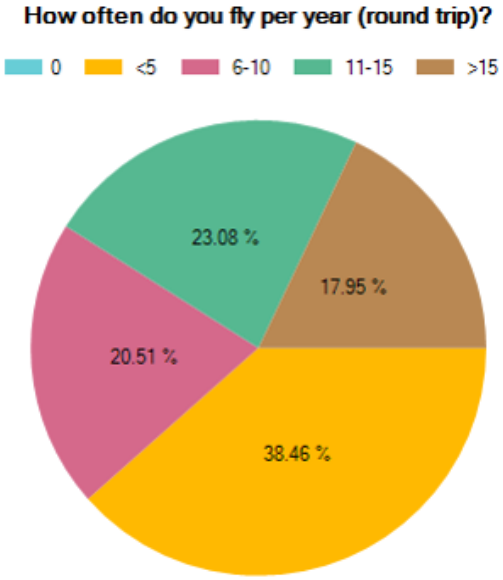

Figure 4.2: Flying frequency of the jurors.

4.2.1 Sound Sample Selection

Due to time constraints, the use of sound samples for the 85 seats in synchronous and asynchronous flying conditions is not possible. Therefore, a representative set of sound samples had to be chosen for the jury testing. Considering that, according to Otto et al. [2001], the maximum length of the test should be limited to 30-45 minutes (for avoiding juror fatigue), 35 sound samples were selected, with lengths of about 6.5 seconds.

For selecting the time samples, three different conditions were used as criteria:

(i) Ensure maximum difference between sound samples;

(ii) Provide an efficient sampling of the features domain;

(iii) Guarantee presence of audible difference in the sound samples included;

Satisfaction of condition (i) is achieved by selecting a sound sample from each one of the clusters, mentioned in Ch. 2. Because, there is a degree of correlation between the psychoacoustic metrics and annoyance, as it will be shown in Ch. 6, it is reasonable to use data clustered with the psychoacoustic metrics for obtaining annoyance.

Regarding condition (ii), by including the extreme cases for each psychoacoustic metric, it is possible to ensure an efficient sampling of the metrics domain. This is an extremely important condition in order to obtain a representative set of data for building the sound quality prediction models that are described in Ch. 6. Condition (iii) is satisfied by doing

a careful screening of the chosen sounds, prior to the jury testing, as described in Otto et al. [2001]. Screening is accomplished by simply auditing the sounds as they are to be presented.

Additionally, it is important to add that, for checking the juror consistency, 5 repeated sounds were added. After this process, a total of 35 sound samples were selected (including the repeated ones). These are represented in Tab. 4.1.

Table 4.1: Seat numbers of the sound samples used for the jury testing; (r) indicates the sound sample was repeated

Synchronous Case	Asynchronous Case
68	37
35	20
15	48 (r)
1	22
40	6
5 (r)	64
59	50
20	59
25	71 (r)
26	73 (r)
64	82
72	30
79	1
85	5
81	-
84 (r)	-

It should also be noted that all the sound samples were equalized to improve the sound quality during the test. This was done while in time domain, dividing each sample for the maximum value of both synchronous and asynchronous time histories. Also, fade in and fade out effects were included, for softening the start of the sound samples [Otto et al., 2001].

4.3 Jury Evaluation Method and Interface Design

This section is destined for defining both the presentation and the evaluation format used in the test. The method selected for performing the jury evaluations was the so-called *Semantic Differential*, with the use of an *anchor sound*. In a semantic differential test, sound samples are presented one by one to the jurors, who have to rate them based on a pair of two opposing adjetives or expressions. These are designated by *bipolar adjectives* or *pairs*, being that they lie in opposite ends of a scale with different gradations. The gradations are labeled with appropriate adverbs that allow the subject to rate the magnitude of their impressions. As recommended by Otto et al. [2001], a seven point scaled was used. This allows to break down the sound into impressions and feelings by the jurors.

However, a modification in the classical semantic differential test was used. Before, the juror listens to the sound sample he is assessing, he will first hear an anchor sound, which will be always the same sound and corresponds to the seat number 38 in synchronous conditions. The criteria for choosing the anchor was to use a seat where the corresponding

sound sample psychoacoustic metrics are an average of the other sounds. Also physically this seat is a reasonable choice, keeping in mind that seat 38 is located in the middle of the fuselage. By including the anchor sound, the juror, for each comparison, always has the same reference sound, making it easier to be consistent throughout the test. In Fig. 4.3, a schematic helps to better understand the use of the sound sample preceded by the anchor sound.

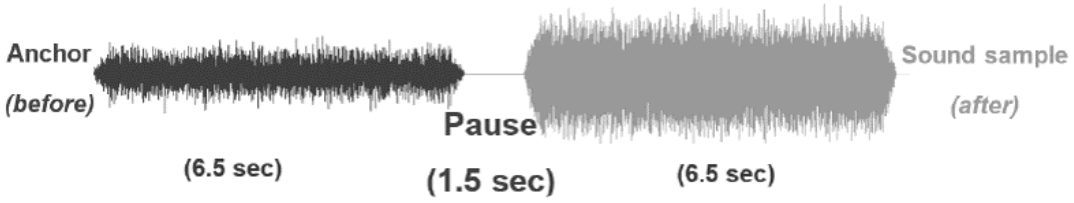

Figure 4.3: Track sequence used in the jury test.

The software used to conduct the jury testing was the *Jury Testing* component of LMS Test.Lab 17, which was developed specifically for this purpose. Fig. 4.4 represents the interface that each juror used.

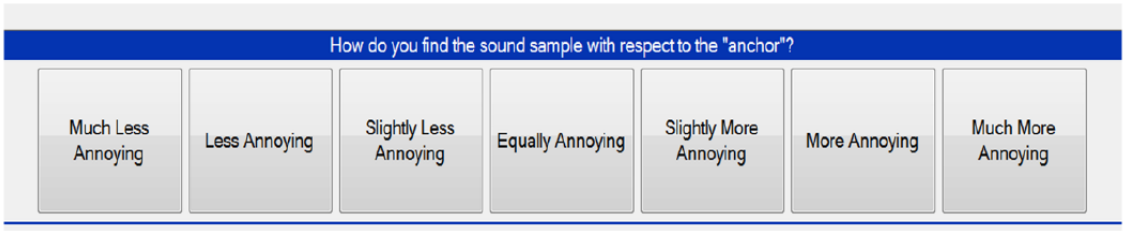

Figure 4.4: Interface created for the jury test.

It should also be added, that prior to the start of the jury test itself, the jurors received an overview of the project in which this jury testing activity is inserted, in order for them to understand the reasons that led to conduct jury testing and what can be done with the results. This was done recurring to two posters, created specifically for this purpose. Also, prior do the jury test itself, the jurors had the opportunity to experience a training test, where they can hear the anchor and a sound sample and insert their response, whithout this answer being accounted for. The goal is to acclimate subjects to both the sounds and the evaluation task, as advised by Otto et al. [2001].

Chapter 5

Sound Quality Prediction Models

5.1 Introduction

There has been a resurgence of interest in ML methods over the last few years, as researchers from diverse backgrounds have produced a firm theoretical foundation and demonstrated its usefulness in numerous applications. This recent renewed interest is due, for example, in ANNs, to new training techniques for more sophisticated architectures, also the always ongoing increase of computation power available and the new possibilities brought by the updated approaches in parallel computing [Fausett, 1994]. This advancements were a key factor to the decision of applying several different prediction techniques in a sound quality context and studying the produced results.

As introduced in Ch. 1, the ability to predict subjective sound quality metrics from objective ones allows the acoustic engineer to shorten the design cycle, since the need to conduct jury testing, as described in Ch. 4, is time consuming and inhibits him to include the human dimension in an early acoustical design intervention. As a consequence, this intervention occurs late in the product development cycle, when all the relevant parameters are already fixed, and does not properly considers the passenger comfort and reduction of annoyance. The development of a VP Model and its posterior combination with virtual aircraft prototyping tools, allows to, after a design or materials alteration, synthesize new sound samples and quantify the human reaction to these, opening the path towards developing a human-inclusive multi-attribute design optimization process. Fig. 5.1 contains a flowchart of the process described in this paragraph.

Due to the complex relationships between the human hearing process and acoustic performance, an individual psychoacoustics-based sound quality metric cannot alone model the human performance. Consequently, ML techniques are more suited to describe the relationships at hand, being developed using both the psychoacoustic objective metrics and the subjective evaluations provided by jurors. The ML based pattern recognition techniques can be broadly divided into classification-based learning methods and regression-based learning methods [Huang et al., 2016]. The work developed during this thesis focused on regression-based techniques.

The development of ML models to predict subjective sound quality metrics, from objective metrics (psychoacoustic) has already been done by some researchers. For example, the prediction performance of subjective sound quality (as annoyance) from psychoacoustic metrics usually is done by comparing the performance of a MLR approach, which represents a more traditional method, with the results of ML techniques, as Xue et al. [2016] has done (comparing the linear-based model with ANNs and SVMs models, regarding SQ for vehicle HVAC). In a similar way, Liu et al. [2015] compare different ML

prediction models for sound quality in engine-radiated noise. Therefore, section 5.2 is destined to define the model used for predicting annoyance directly from psychoacoustic metrics. For this, four different prediction techniques were used, with the goal of comparing a more current/traditional prediction method, MLR, with three machine learning techniques, namely ANNs, SVMs and RFs. Ch. 6 contains the results obtained through each one of these models.

During the work performed for this thesis, while finalizing the models referred to in the paragraphs above, a new approach on the prediction of sound quality, different from the ones found in current scientific literature on the topic, was created and implemented. The prediction process described so far relies on the use of the psychoacoustic metrics obtained using LMS Test.Lab. However, the computation of these metrics still requires the acoustic engineer to select some acoustic and digital signal processing parameters, which are case dependent and vary with the type of sound sample considered. So, due to the fact that this step still demands a *human input*, a deep learning model was implemented, using CNNs, to predict the psychoacoustic metrics from the sound samples time signals. Also, this new type of approach allows to obtain a deep learning compact model that is easily insertable, in a posterior phase, in, for example, an active control block diagram on Simcenter Amesim or MATLAB Simulink. Therefore, this type of models allow to exploit with a bigger ease, for example, active sound control or multi-attribute optimization design processes.

This decision was highly motivated by the analysis of applications of deep learning in other audio processing cases, such as voice recognition or music classification. Indeed, according to the literature, although hand-crafted acoustic features are typically well designed procedures, it is still not possible to retain all useful information due to the human knowledge bias and the high compression rate. One way to overcome this limitation is to feed raw time-signals into deep CNNs, which can learn spatially or temporally invariant features from pixels or time-domain waveforms, providing a more thorough end-to-end process by completely abandoning the feature extraction step [Aytar et al., 2016; Dai et al., 2017; Sainath et al., 2015; Thickstun et al., 2017; Trigeorgis et al., 2016].

In recent years, CNNs have been used with great success in a wide range of contexts. Although computer vision is perhaps the field where its application is more popular, if a correct representation of sound is used, CNNs are also well suited for this type of tasks [Shuvaev et al., 2017]. Their use allows the possibility to train a prediction model using the sound samples from all seats from the aircraft and their corresponding values for the psychoacoustic metrics, thus being able to predict the psychoacoustic features directly from the sound samples (inputted as time signals). This model is defined in section 5.3.

Therefore, by combining the two models described above, it is possible to directly predict annoyance for new sound samples. This final model, designated by VP model, has as an input a time signal and in a first stage predicts psychoacoustic metrics, being that it was trained using data from LMS Test.Lab. In Fig. 5.2, the data pipeline in the VP model is shown. The sound samples are inputted as time signals in a trained CNN based prediction model being predicted (as output) psychoacoustic features that will serve as an input to a feature-based model (trained on the jury testing data), being the final output an annoyance evaluation.

Throughout this chapter the data flow in the final VP prediction model will become clearer, being that each one of both the *prediction blocks* will be individually analyzed (sections 5.2 and 5.3). Finally, on section 5.4, the complete model will be further described For all the models mentioned, their results and performances can be found on Ch. 6 and Ch. 7.

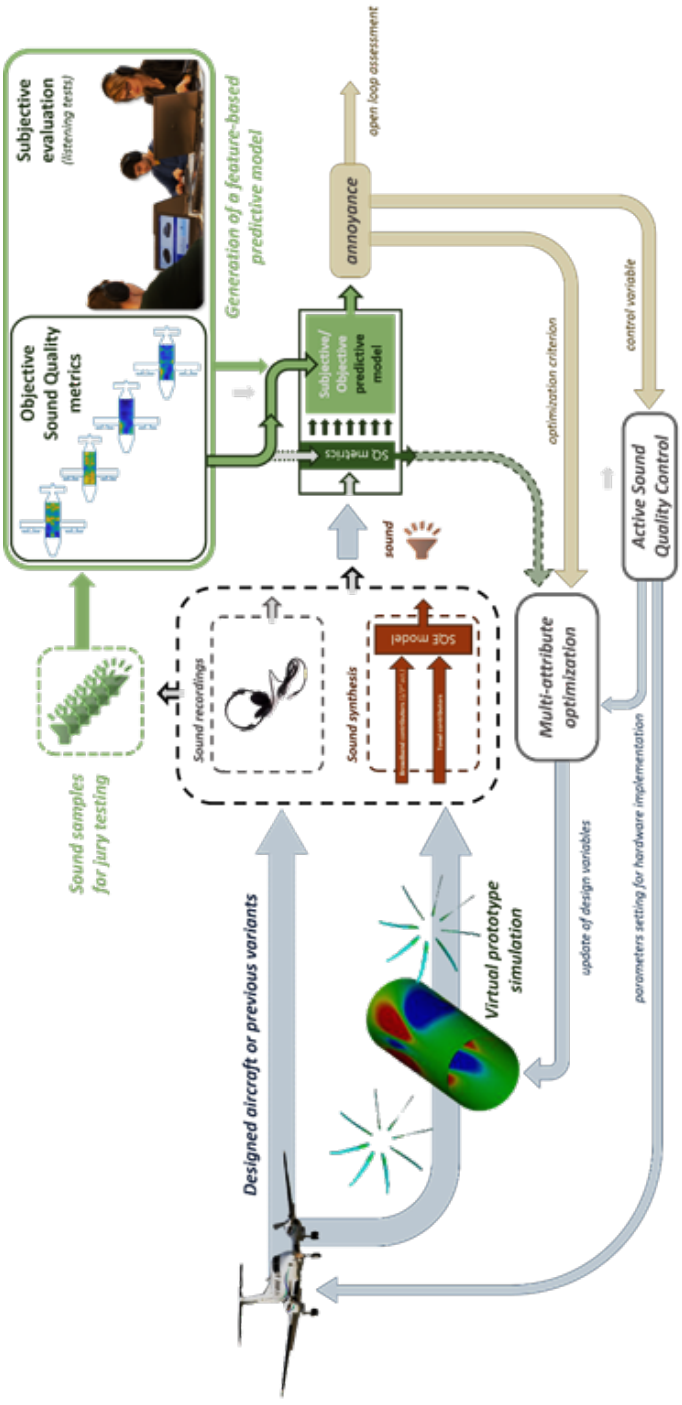

Figure 5.1: Flowchart illusttrating the role of a sound quality prediction model in the aircraft design process.

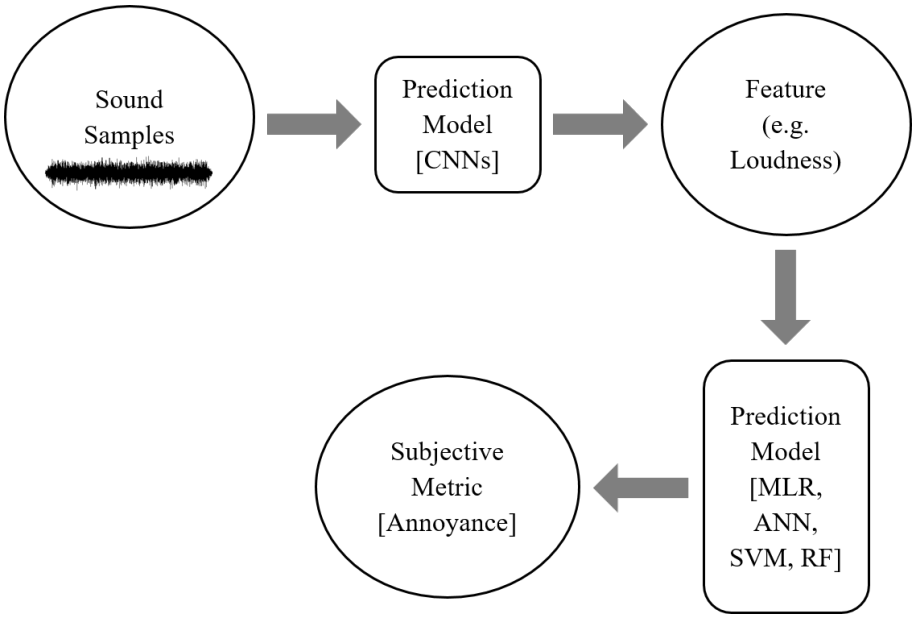

Figure 5.2: Data pipeline throughout the different blocks of the Virtual Passenger Model.

Assessing the Prediction Models Performance

After developing the prediction models, i.e, training, it is necessary to validate them and it would not be adequate to validate, i.e. test, the models with the same data that was used for training them. As an analogy with a real life example, consider a student going to an exam. If in this exam the professor would ask to solve exactly the same problems the student solved during classes, he would pass the exam not because he understood the concepts, but because he memorized the exercises. Therefore the students grade would not reflect his ability to solve new problems, but only his ability to solve only the problems taught during classes. In the prediction models the situation is analogous. For properly assessing performance, the models should predict on *fresh* and *unseen* data, left out on purpose during training [Russell et al., 2010]. Due to the time constraints associated with the work proposed for this thesis not allowing to, for example, re-conduct jury testing on more sound samples, it is necessary to divide the existing data into *training data* and *testing data*.

Let x be a set of b input predictor vectors for the model. Each vector contains p observations (or samples). For the model based on the jury testing data, each one corresponds to a objective metric (psychoacoustic) and for the CNN based model they are the time signals of the sound samples. Also, it is necessary to define y as the response variable, containing p observations or samples, being either the subjective metric (annoyance in the jury testing model) or the objective psychoacoustic metrics (CNN model). Considering training data as the data used to train the model and testing data as the data used for assessing the models performance, after dividing the data set (originally with p samples), two different data sets are obtained. Therefore, considering m samples for training and n samples for testing (in a way that $p = m + n$):

$$D = \{x_i, y_i\}_{i=1}^{p} \tag{5.1}$$

$$D_{train} = \{x_i, y_i\}_{i=1}^{m} \tag{5.2}$$

$$D_{test} = \{x_i, y_i\}_{i=1}^{n}. \tag{5.3}$$

The model, after training with the training data, D_{train}, is tested using the testing data, D_{test}. By feeding it the input testing values $x_i (i = 1, 2, .., n)$, the predicted responses, $y_{pred_i} (i = 1, 2, ..., n)$, are obtained and compared with the original response values contained in the testing data, $y_i (i = 1, 2, ..., n)$. Keeping in mind, regression-based techniques are being used, the performance is assessed with three different metrics. According to Huang et al. [2017] and Rawlings et al. [1998], the Mean Absolute Error (MAE) and RMSE are computed with the following equations:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - y_{pred_i}| \tag{5.4}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - y_{pred_i})^2}. \tag{5.5}$$

The *Coefficient of Determination*, denoted by R^2, in the context of prediction models, indicates how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. Its calculation is done using,

$$SSR = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{5.6}$$

where SSE is the sum of squared error, SSR is the sum of squared regression, SST is the sum of squared total and, as it has been explained above, n is the number of observations in the testing data. This coefficient ranges from 0 to 1 and the closer the value is to one, the better the fit, or relationship, between observed and predicted values, i.e, a value close to one most likely indicates the model is able to predict more accurately [Rawlings et al., 1998].

Regarding the data division, the diagram in Fig. 5.3 allows to gain a better comprehension on of the several steps necessary to properly train and test each one of the prediction models previously mentioned.

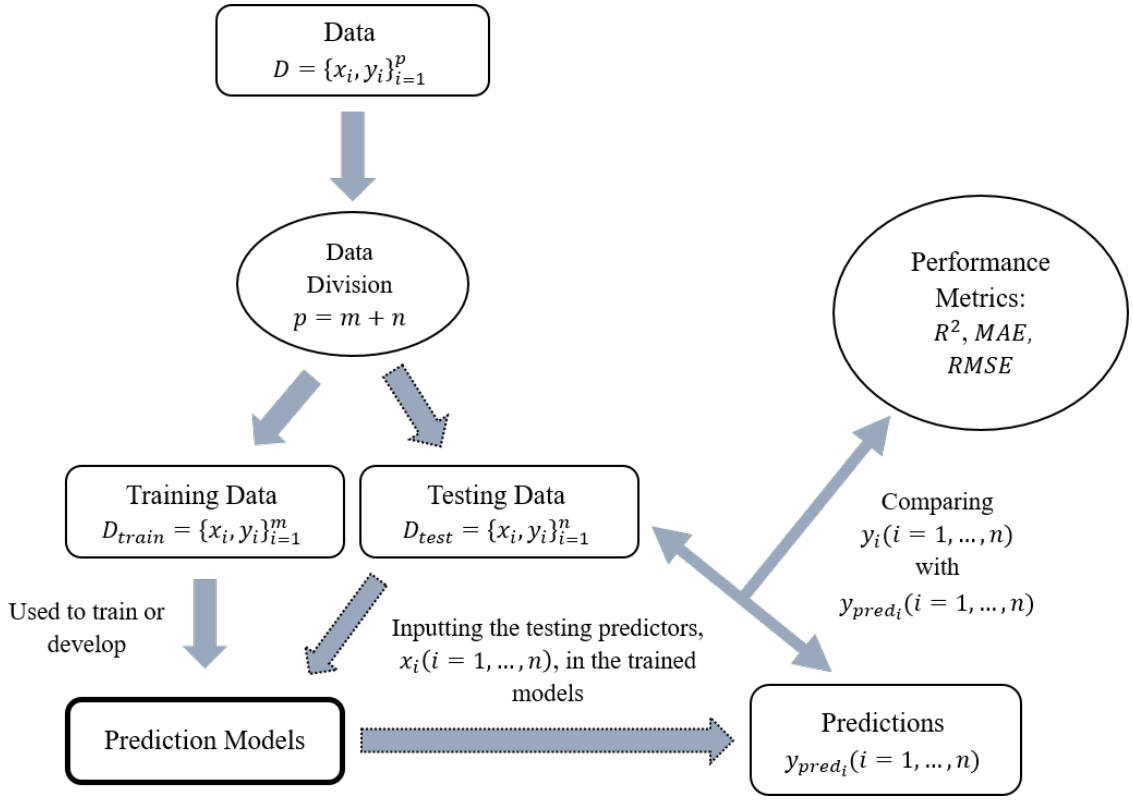

Figure 5.3: Diagram with the data flow used to train and test the prediction models.

An additional metric necessary to analyze the data at hand is the Pearson correlation coefficient. It is a measure of the linear correlation between two variables, having a value between +1 and -1, where +1 is a total positive linear correlation, 0 means that there is no linear correlation and -1 corresponds to a total negative linear correlation. For two variables, A and B with N scalar observations, the Pearson correlation coefficient, ρ, is given in terms of the covariance of A and B by:

$$\rho(A, B) = \frac{cov(A, B)}{\sigma_A \sigma_B} \tag{5.7}$$

where σ_A and σ_B are the standard deviations of A and B [Fisher, 1925].

Data Division and Monte Carlo Simulations

Even though data division is sometimes an overlooked aspect in ML, the method used and the amount of data used for training and testing may have a significant influence on

the prediction models performance. The first approach used in this work was to randomly select data for training and testing. Obviously, randomly choosing data implies that the one used for training may not be a representative group of the overall data set. This effect on performance was studied through *Monte Carlo* simulations (defined in the next paragraph).

About the Monte Carlo simulation method, as can be read in Kwak and Ingall [2007], it encompasses any technique of statistical sampling employed to approximate solutions to quantitative problems. A model or a real-life system or situation is developed, and this model contains certain variables. These variables have different possible values, represented by a probability distribution function of the values for each variable. The Monte Carlo method simulates the full system many times (hundreds or even thousands of times), each time randomly choosing a value for each variable from its probability distribution. The outcome is a probability distribution of the overall value of the system calculated through the iterations of the model.

This simulation method was applied to study several aspects of the prediction models developed for predicting annoyance from psychoacoustic metrics, where for several of the hyperparameters of the prediction models, Monte Carlo simulations were performed to study their influence on performance. Therefore, after dividing the data randomly many times, it is possible, for example, to average the obtained performance metrics, thus obtaining the overall performance through the different randomizations, for different hyperparameters of the models, and also its standard deviation.

Bayesian Optimization

ML algorithms are rarely parameter-free: parameters controlling the rate of learning or the capacity of the underlying models must often be specified and carefully tunned. Unfortunately this tunning frequently requires more experienced knowledge, rules of thumb or even *brute-force* searches [Snoek et al., 2012]. A more flexible take on this issue is to perform an automated optimization of these parameters. Bayesian optimization has been shown by Jones [2001] to outperform other state of the art global optimization algorithms on a number of challenging optimization benchmark functions.

Bayesian Optimization owes its name to the famous *Bayes* theorem, which states, in a simplified way, that the posterior probability of a model (or theory), M, given evidence (or observations), E, is proportional to the *likelihood* of E given M multiplied by the prior probability of M [Brochu et al., 2010]. The theorem is expressed as,

$$P(M|E) \propto P(E|M)P(M). \tag{5.8}$$

According to Brochu et al. [2010] and Zhang et al. [2015], first x_i is defined as the ith sample and $f(x_i)$ corresponds to the observation of the objective function, $f(x)$, at x_i. As observations are accumulated in $D_{1:t} = \{x_{1:t}, f(x_{1:t})\}$, the prior distribution is combined with the likelihood function $P(D_{1:t}|f)$. Therefore, in the framework of this optimization procedure, the objective function is assumed to be drawn from the following probabilistic model:

$$P(f|D_{1:t}) \propto P(D_{1:t}|f)P(f). \tag{5.9}$$

Rephrasing in a simpler way, as in other kinds of optimization, Bayesian optimization sets out to find the minimum of a function, $f(x)$, on some bounded set χ, which is taken as a subset of $\mathbb{R}^D$. The significant difference of this method from other optimization procedures is the fact that it constructs a probabilistic model for $f(x)$ and then

exploits this model to make decisions about where in χ to next evaluate the function, while integrating out uncertainty. The goal is to use *all* of the information available from previous evaluations of the objective function, $f(x)$, and not simply rely on local gradient and Hessian approximations. The resulting procedure is able to find the minimum of difficult non-complex functions with relatively few evaluations, at the cost of performing more computation to determine the next point to try. Even though the evaluations of $f(x)$ are expensive to perform, the ability to make better decisions justifies the extra computation cost [Snoek et al., 2012]. Considering the possibility of using parallel computing, this optimization procedure was used in some of the ML methods (SVMs and RFs) and also in the CNNs.

5.2 Annoyance Prediction from Psychoacoustic Metrics

The annoyance prediction model based on psychoacoustic features was built using the data obtained from the conducted jury study. After averaging the annoyance evaluations from all jurors for each sound sample, it is possible to obtain annoyance evaluations for the 30 stimuli. Therefore, the data set available to build the prediction model consists of 30 pairs of objective/subjective evaluated sound samples.

The main assumption involved in this approach is that there is a pattern (either linear or non-linear) behind the human estimation process, as long the context and the borders of the estimation set are kept well defined. Therefore, a curve-fitting approach is used.

The following subsections have as a goal to provide a theoretical overview of the prediction models used. However, especially for the ML methods (or the deep learning one), the framework of this thesis does not includes as an objective to extensively review this concepts. For each procedure, a global theoretical overview is presented, indicating literature where further details may be consulted.

5.2.1 Multiple Linear Regression

A MLR algorithm is a mean of determining relationships between two or more explanatory variables and a response variable by fitting a linear equation to the observed data. The *multiple* refers to multiple independent variables [Xue et al., 2016]. This procedure has as advantages the fact that for using it only a small amount of data is required and it provides reasonable outputs through simple calculations [Huang et al., 2016].

The linear model for relating a dependent variable to b independent variables is,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_b x_{ib} + \varepsilon_i \quad (i = 1, 2, \ldots, m) \tag{5.10}$$

where the subscript i denotes the observational unit from which the observations on y and the b were taken. The second subscript designates the independent or explanatory predictor variables, which correspond to the psychoacoustic metrics. The sample size (number of samples selected for training) is denoted with m, $i = 1, 2, \ldots, m$ and b denotes the number of independent variables (*i.e* the number of used psychoacoustic features). Regarding β, it is the weighting coefficient related to x and β_0 is the constant term. Finally, ε_i is the error term [Rawlings et al., 1998]. The matrix expression of Eq. 5.10 is,

$$\{y\} = [x_{train}]\{\beta\} + \{\varepsilon\} \tag{5.11}$$

where $\{y\}$ is the column vector of response observations, $[x]$ is matrix where, after a column of ones, each column represents the observations of an independent variable (psychoacoustic metric), $\{\beta\}$ is a vector of coefficients and $\{\varepsilon\}$ a vector of residual errors.

As stated by Rawlings et al. [1998], the coefficients are obtained through a least squares approximation. After this computation, usually done using statistical software, the predicted response values are calculated using the test objective metrics and the vector of coefficients:

$$\{y_{pred}\} = [x_{test}]\{\beta\}. \tag{5.12}$$

The residuals are then, as expressed by Rawlings et al. [1998], the difference between the observed test responses and the predicted (fitted) test responses,

$$\{\varepsilon\} = \{y_{test}\} - \{y_{pred}\}. \tag{5.13}$$

A MLR can effectively solve problems whose inner relationship are not very complex. However, it performs poorly when addressing strongly nonlinear issues [Huang et al., 2016].

5.2.2 Artificial Neural Networks

Another approach to modeling annoyance response of engine sounds is to use ANNs. This is an effective technique that uses a non-linear algorithm, whose methods for minimizing error are more efficient than other non-linear technique. Inspired by biological neural networks, ANNs are based on the present understanding of biological neural systems. The neural network considered for this section is the feed-forward neural network, with a back propagation algorithm [Kahn, 1998].

Firstly, before going over the neural network itself, it is necessary to understand the characteristics of each neuron. As illustrated in Fig. 5.4, on the left side neuron, a scalar input p is transmitted through a connection that multiplies its strength by the scalar weight w to form the product $w \times p$, again a scalar. Here, the weighted input is the only argument of the transfer function, F, which produces the scalar output a. Consider now the neuron on the right where a scalar bias b is considered. The bias is like a weight, except it has a constant input of 1. The transfer function net input n, again a scalar, is the sum of the weighted input $w \times p$ and the bias b. This sum is the argument of the activation function (or transfer function) f, which will be described further in the next paragraphs. Note the weights and bias are both adjustable parameters of the network. The central idea of neural networks is that these parameters can be adjusted so that the network exhibits the desired output [Kahn, 1998].

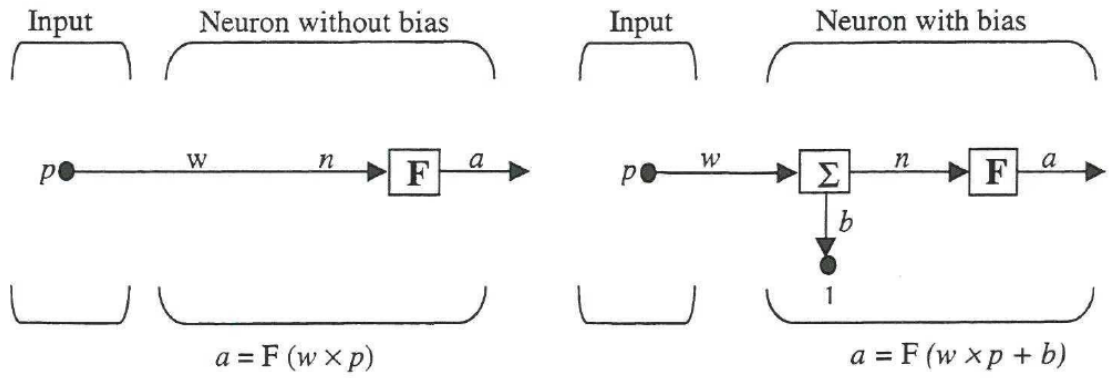

Figure 5.4: Diagram showing the characteristics of a single neuron in a back propagation neural network. The right side neuron includes the bias and the left side does not [Kahn, 1998].

A back propagation algorithm corresponds to a multilayer perceptron, consisting usually of at least three hierarchical layers of neurons, one input layer, one or more hidden layers and one output layer. The network is connected in such a way that each layer is fully connected to the next layer, i.e, every neuron in the input layer will send its output to every neuron in the input layer. The number of neurons in the input layer is equal to number of variables in the input data. The number of neurons in the output layer is the same as the number of output variables. The number of neurons in the hidden layers can be varied based on the complexity of the problem and the size of the input information [Kahn, 1998].

Fig. 5.5 illustrates the type of network described in the previous paragraph. The inputs are the b psychoacoustic metrics considered and the number of hidden neurons is k.

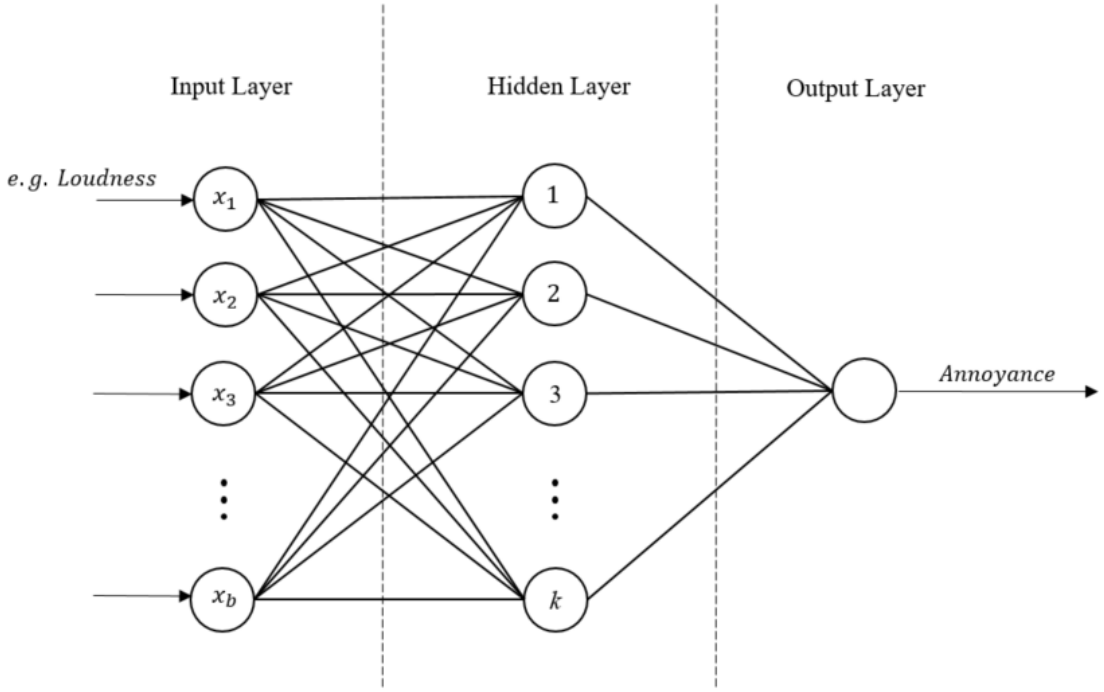

Figure 5.5: Schematic diagram of a feed-forward ANN.

The main goal of the network is to map the input, *i.e*, $x \in \mathbb{R}^b$, into the output, *i.e*, $y \in \mathbb{R}$. The mapping is performed by a network composed of processing units (neurons) and connections between them. A neuron, i, in a network accumulates input signals, x_j, in the summing block and is activated by function f to have only one output y_{pred_i},

$$y_{pred_i}(x, w) = f(a_i) = f(\sum_{j=1}^{b} w_{ij} x_i + b_i) \tag{5.14}$$

where w_{ij} are the weights of connection and b_i is the bias (threshold parameter) [Lee and Chae, 2004].

Regarding f, it corresponds to the activation function. Feedforward networks often have one or more hidden layers of non-linear neurons (for example, with sigmoid activation functions), that allow the network to learn non linear relationships between inputs and outputs, followed by an output layer. In this output layer, which, as stated by Bishop [2006], in the case of a classification-based models would be a non-linear function (for example a sigmoid) and for a regression or function fitting problem corresponds to a linear

activation function [Bishop, 2006]. A popular non linear activation function is the sigmoid function, which is expressed as,

$$f(a_i) = \frac{1}{1 + e^{-a_i}}. \tag{5.15}$$

The use of non linear activation functions in the hidden layers of the network implies that the neural network function is differentiable with respect to the network parameters. This plays a key role in the network training algorithm. Defining an error function, $E(w)$, based on predicted and observed training data, the goal is to minimize this function depicted in Eq. 5.16 [Bishop, 2006].

$$E(w) = \frac{1}{2} \sum_{n=1}^{M} \|y_{pred_i}(x_n, w) - y_n\|^2 \tag{5.16}$$

As shown by [Bishop, 2006], this problem is solved by adopting a geometrical approach. Representing the error function as a surface sitting over the weight space, illustrated in Fig. 5.6, it is clear that the goal is to chose a weight vector, w, that minimizes the error function, thus achieving the global minimum. Point w_A is a local minimum and point w_B is a local maximum. At any point w_C, the local gradient of the error surface is given by the vector ∇E.

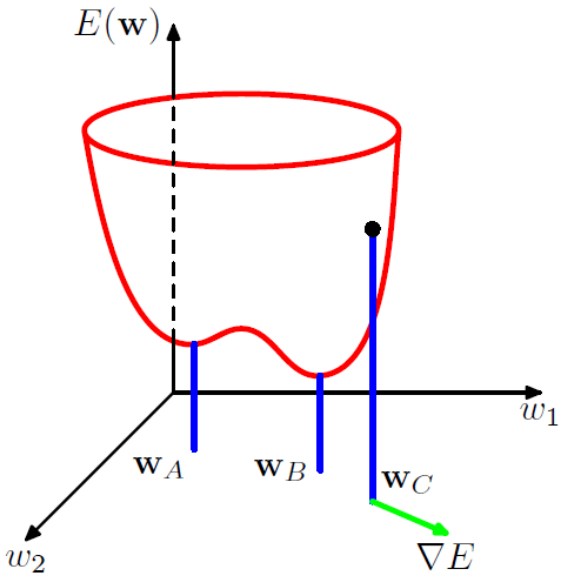

Figure 5.6: Geometrical view of the error function, $E(w)$, as a surface sitting over weight space [Bishop, 2006].

Being the error $E(w)$ a smooth continuous function of w, its smallest value will occur at a point in weight space such that the gradient of the error function vanishes, so that

$$\nabla E(w) = 0. \tag{5.17}$$

Therefore, the network *training* corresponds to the iterative process of computing the network parameters which minimize its error. There exist several training algorithms, such as the Levenberg–Marquardt (LM) algorithm (a combination of the gradient-descent and the Gauss-Newton method) which is often a first choice algorithm due to its robustness and performance [Zhang et al., 2017]. Also, in order to obtain a smaller error, some techniques

are used alongside back-propagation such as the Bayesian Regularization (BR), which is a variation of the LM algorithm. LM and BR algorithms are often able to obtain lower mean squared errors than any other algorithms for functioning approximation problems. LM was especially developed for faster convergence in backpropagation algorithms. About BR, it has an objective function that includes a residual sum of squares and the sum of squared weights to minimize estimation errors and to achieve a good generalized mode [Kayri, 2016; MacKay, 1992]. Their formulations and complete algorithms can be found in the references suggested in this paragraph. Both algorithms will be used and their results will be compared in Ch. 6.

5.2.3 Support Vector Machines

SVMs were proposed by Vapnik and co-workers in 1995 [Vapnik, 1999, 2013]. This procedure corresponds to a statistical learning approach based on a risk minimization principle. Recently, it has been successfully applied to solve classification and regression problems in numerous fields [Wu et al., 2007]. In SVMs, the input data of a low-dimensional feature space is first mapped into high-dimensional feature space using a kernel function, and then linear regression is performed in the feature space. A separating hyperplane is obtained to maximize the margin between the training examples and the class boundary in high-dimensional feature space [Liu et al., 2015].

Considering the training dataset, $D_{train} = \{x_i, y_i\}_{i=1}^{m}$, with m samples for training, according to Liu et al. [2015], it is possible to define an ε-insensitive loss function, which gives zero error if the absolute difference between the prediction and the target is less than ε, where $\varepsilon > 0$ [Bishop, 2006]. It is expressed as:

$$|y - f(x_i)|_\varepsilon = \begin{cases} 0, & \text{if } |y - f(x_i)| \leq \varepsilon \\ |y - f(x_i)| - \varepsilon, & \text{if } |y - f(x_i)| > \varepsilon \end{cases} \tag{5.18}$$

In high-dimensional feature space, support vector regression is used to find the linear relation as follows:

$$f(x_i) = \omega \cdot \phi(x_i) + b \tag{5.19}$$

where ω is a vector for regression coefficients, ϕ is a nonlinear mapping from low-dimensional feature space to high dimensional feature space, $\omega \cdot \Phi$ represents the inner product of the two vectors and b represents a bias [Liu et al., 2015]. They are estimated by minimizing the regularized risk function, $R(C)$, as:

$$min\, R(C) = \frac{1}{2} \parallel \omega \parallel^2 + C \sum_{i=1}^{m} (\xi_i - \xi_i^*) \tag{5.20}$$

In the previous equation, $\frac{1}{2} \parallel \omega \parallel^2$ is used as a flatness measurement of function, C is a penalty factor which determines the trade off between the training error and the model smoothness and ξ and ξ_i^* are positive slack variables. Considering the constraints for 5.20 as:

$$\begin{cases} y_i - \omega \cdot \phi(x_i) - b \leq \varepsilon + \xi_i \\ -y_i - \omega \cdot \phi(x_i) + b \leq \varepsilon + \xi_i^* \end{cases} \tag{5.21}$$

The Lagrange equation is built as:

$$max\,\omega(a_i, a_i^*) = \sum_{i=1}^{m}(a_i - a_i^*)y_i - \sum_{i=1}^{m}(a_i - a_i^*)\varepsilon - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(a_i - a_i^*)(a_j - a_j^*)K(x, x_i) \quad (5.22)$$

where $K(x, x_i) = \phi(x) \cdot \phi(x_i)$ is a kernel function, a_i and a_i^* are Lagrange multipliers to be solved. Only the nonzero values of Lagrange multipliers are useful in predicting the regression line, and their corresponding samples are known as support vectors [Liu et al., 2015]. The constraints of Eq. 5.22 are:

$$\sum_{i=1}^{m}(a_i - a_i^*) = 0 \quad (5.23)$$

where $a_i, a_i^* \in [0, C]$.

Therefore, as stated in Liu et al. [2015], the function regression problem on SVMs may be reduced to a quadratic programming problem. The array ω can be written in terms of the Lagrange multipliers and training samples as:

$$\omega = \sum_{i=1}^{N}(a_i - a_i^*)\phi(x_i) \quad (5.24)$$

The choice of different kernel functions can generate different support regression models. Common kernel functions types of SVMs can be found in Ding et al. [2008]. For example:

$$\begin{cases} \text{Radial Basis Kernel:} & K(x, x_i) = exp(-\gamma \parallel x - x_i \parallel^2) \\ \text{Polynomial Kerne:l} & K(x, x_i) = exp(x^T x_i + r)^q) \end{cases} \quad (5.25)$$

where γ, r and d are constants.

Finally, as shown in Liu et al. [2015], the linear Eq. 5.19, has the following explicit form:

$$f(x_i) = \omega \cdot \phi(x_i) + b = \sum_{i=1}^{k}(a_i - a_i^*)K(x, x_i) + b \quad (5.26)$$

The generalization ability of support vector regression depends entirely on the penalty constant C and on the constants used in the kernel functions. However, the SVMs technique is limited in feature subset selection and parameter optimization [Liu et al., 2015]. The Bayesian optimization procedure was used for obtaining the optimal parameter subset.

5.2.4 Random Forests

In the latest years, among the several existing ML techniques available, decision trees have stood out as a popular procedure due to their simplicity, ease of use and interpretability. Instead of averaging the predictions of a set of models, an alternate form of model combination is to select one of the models to make the prediction, in which the choice of the model is a function of the input variables. Thus, different models become responsible for making predictions in different regions of the input space. In a decision tree, the process can be described as a sequence of binary selections corresponding to the traversal of a tree structure. In this procedure, the individual models are generally chosen to be very simple, and the overall model flexibility arises from the input-dependent selection process [Bishop, 2006].

As mentioned above, decision trees build a regression or classification model in the form of a tree structure. They divide the data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing values for the attribute tested. Leaf nodes represent a decision on the numerical target. The topmost decision node in a tree corresponds to the best predictor, called root node. Fig. 5.7 shows an illustration of a recursive binary partitioning of the input space, along with the corresponding tree structure [Bishop, 2006].

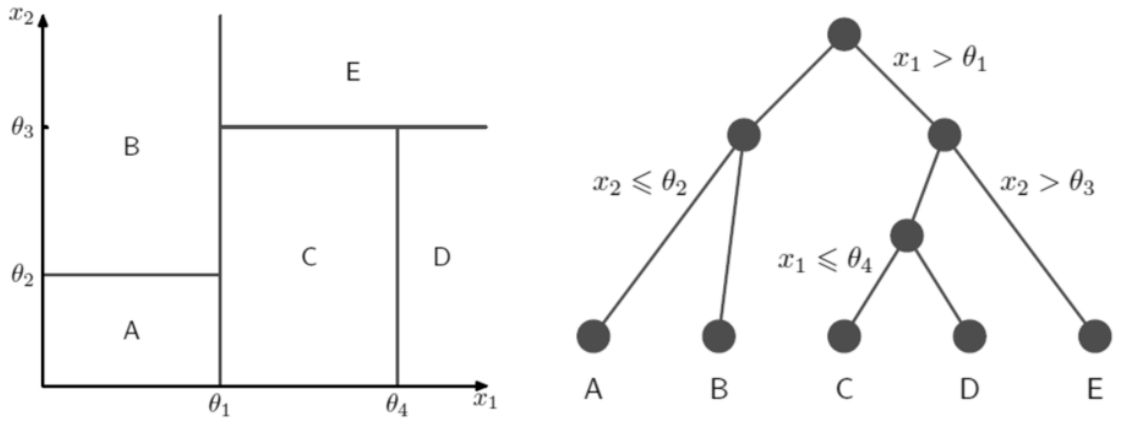

Figure 5.7: On the left, a representation of a two dimensional input space that has been partioned into five regions usig axis-aligned boundaries. The right side diagram, illustrates a binary decision tree corresponding to the input space [Bishop, 2006].

For any new input x, the region which falls into is determined by starting at the top of the tree at the root node and following a path down to a specific leaf node according to the decision criteria at each node. Within each region, there is a separate model to predict the target variable. Note that these are not probabilistic graphical models [Bishop, 2006].

Considering the training data set, $D_{train} = \{x_i, y_i\}_{i=1}^{m}$, on which the model will be built, if the input space partitioning is given, by minimizing the sum-of-squares error function the optimal value of the predictive variable within any given region is just given by the average of the values of y_m for those data points that fall in that region [Bishop, 2006].

Regarding the determination of the structure of a tree, due to computational power constraints, it is done generally through a greedy optimization, starting with a single root node, corresponding to the whole input space, and then growing the tree by adding nodes one at a time. At each step there will be some number of candidate regions in input space that can be split, corresponding to the addition of a pair of leaf nodes to the existing tree. For each of these, there is a choice of which of the input variables to split, as well as the value of the threshold [Bishop, 2006].

The joint optimization of the choice of region to split, and the choice of input variable and threshold, can be done efficiently by exhaustive search noting that, for a given choice of split variable and threshold, the optimal choice of predictive variable is given by the local average of the data, as noted earlier. This is repeated for all possible choices of variable to be split, and the one that gives the smallest residual sum-of-squares error is retained. Regarding the issue of when to stop adding nodes, it is common practice to grow a large tree, using a stopping criterion based on the number of data points associated with the leaf nodes, and then prune back the resulting tree [Bishop, 2006].

However, as stated in Bishop [2006], this procedure has shortcomings. For instance, the division of the input space is based on hard splits in which only one model is responsible for making predictions for any given value of the input variables. This decision process can be softened by combining models. The use of ensembles of trees (more specifically RFs) allows to overcome this lack of robustness.

RFs are an ensemble learning methodology and like other ensemble learning techniques, the performance of a number of weak learners (which could be a single decision tree, single perceptron, etc) is boosted by a voting scheme, where, for each test instance, every model makes a prediction, i.e., votes, being the final output based on the overall voting [Ahmad et al., 2017].

A RF is an ensemble of C trees $T_1(X)$, $T_2(X)$,..., $T_C(X)$, where $X = x_1, x_2, ..., x_m$ is a m-dimension vector of inputs. The resulting ensemble produces C outputs $Y_{pred_1} = T_1(X)$, $Y_{pred_2} = T_2(X)$,..., $Y_{pred_C} = T_C(X)$. Y_{pred_C} is the prediction value by decision tree number C. The output of all these randomly generated trees is aggregated to obtain one final prediction Y_{pred}, which is the averaged value of all the trees in the forest. A RF generates a C number of decision trees from an m number of training samples [Ahmad et al., 2017].

In RF, the training algorithm applies the general technique of *bootstrap aggregating*, or bagging to tree learners. Given a training set, bagging repeatedly selects a random sample with replacement of the training set and fits trees to these samples. This decreases the variance of the model, without increasing the bias. While the predictions of a single tree are highly sensitive to noise, the average of many trees is not, as long as they are not correlated. For ensuring this, the use of bootstrap sampling shows different training sets to the trees, de-correlating them [Breiman, 2001].

Finally, RFs include another process called *feature bagging*, using a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of features. Therefore, its possible to avoid the correlation of features in a ordinary bootstrap sample [Hastie et al., 2009].

5.3 Psychoacoustic Metrics Prediction from Time Signals

As introduced in section 5.1, throughout this thesis, a prediction model is developed, on which the input, i.e., the predictor variables are the time signals of the sound samples, being this the first block of the VP prediction model. On Fig. 5.8 it is possible to observe a plot of the time signal that corresponds to the sound sample of seat number 1 of the aircraft, in synchronous flying conditions.

The goal of this model is, by using a type of neural network that is able to receive a time signal as an input (1D array), to train it with the psychoacoustic metrics associated with each sound sample (computed in LMS Test.Lab), in order for being able to, for instance, predict loudness or tonality. In the following subsection, it can be found an overview of the functioning behind CNNs, and the typical architectures that are used with this layer type. Due to performance assessment aspects, this model was developed using only the 140 sound samples that were not used for jury testing. Further explanations of this decision can be found in section 5.4.

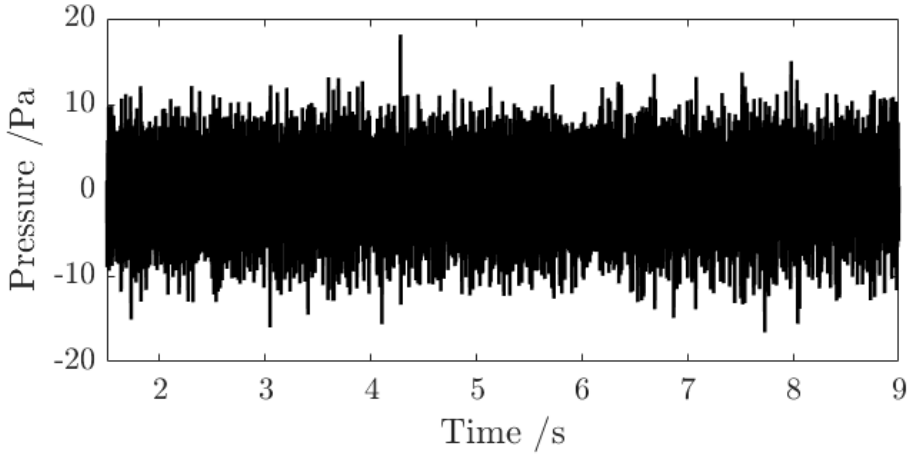

Figure 5.8: Sound sample corresponding to seat number 1, in synchronous flying conditions, as a time signal.

5.3.1 Convolutional Neural Networks

CNNs are designed to process data that comes in the form of multiple arrays, for example, a color image composed of three 2D arrays containing pixel intensities in the three color channels. Several data modalities are in the form of multiple arrays: 1D for signals and sequences, including language; 2D for images or audio spectrograms; and 3D for video or volumetric images. In CNNs, four ideas that take advantage of the properties of natural signals play a key role in their functioning: local connections, shared weights, pooling and the use of many layers [LeCun et al., 2015].

The architecture of a typical CNN is structured as a series of stages. The first few stages contain two types of stages: convolutional layers and pooling layers (or sub-sampling layers). In a convolutional layer, units are organized in feature maps, within each unit is connected to local patches in the feature maps of the previous layer through a set of weights called a filter bank. The result of this local weighted sum is then passed through a non-linearity (activation function). All the units in a feature map share the same filter bank. Different feature maps in a layer use different filter banks [LeCun et al., 2015].

Fig. 5.9 illustrates the structure of a CNN. Two motives justify the use of this structure. First, in data arrays such as images, local groups of values are often highly correlated, forming distinctive local motifs that are easily detected. Second, the local statistics of images and other signals are invariant to location, i.e, if a motive can appear in one part of an image, it could appear anywhere, hence the idea of units ate different locations sharing the same weights and detecting the same pattern in different parts of the array. Mathematically, the filtering operation is performed using a discrete convolution, thus the name [LeCun et al., 2015].

As described above, the role of the convolutional layers is to detect local conjunctions of features provenient from the previous layer. This requires the use of a pooling layer in between the convolutionals, for merging semantically different features into one. Because the relative positions of the features forming a motif can vary somewhat, the reliable detection of motifs is done by fine-graining the position of each feature. The typical pooling unit computes the maximum of a local patch of units in one feature map (or in a few feature maps). Neighbouring pooling units take input from patches that are shifted by more than one row or column, thereby reducing the dimension of the representation

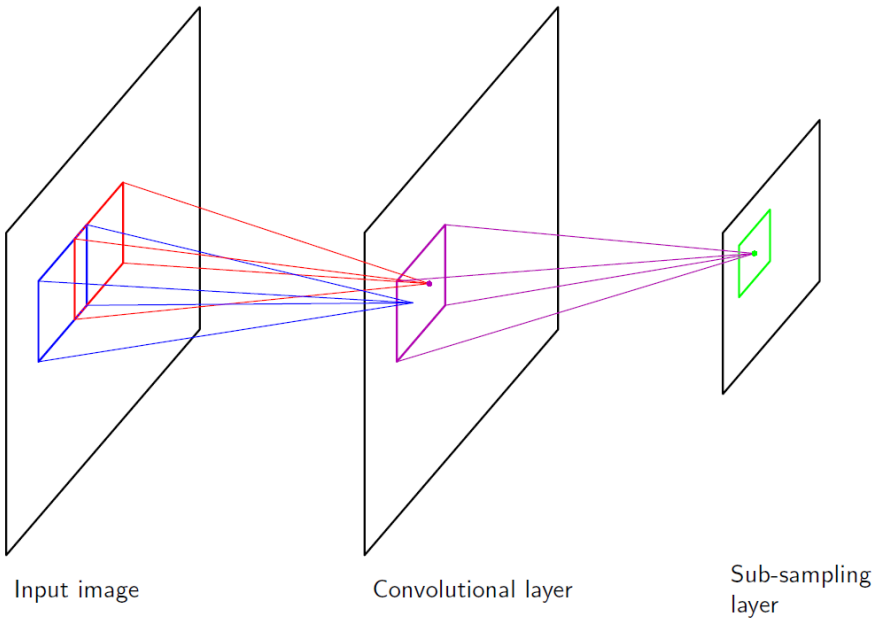

Figure 5.9: Diagram illustrating part of a convolutional neural network, showing a layer convolutional units followed by a layer of sub-sampling units. Several sucessive pairs of such layers may be used [Bishop, 2006].

and creating an invariance to small shifts and distortions. Usually, each stacked block in a CNN contains a convolution layer, followed by the non-linearity and pooling. Depending on the dataset, severall blocks similar to these are connected, forming stacks LeCun et al. [2015]. The mathematical formulation of CNNs can be found in Bishop [2006].

However, regarding the use of pooling layers when processing time signals, this should be done carefully. Due to the fact that, according to the Nyquist sampling theorem, performing temporal pooling in audio processing corresponds to a downsampling operation, thus possibly originating an aliasing effect.

Other types of layers quite popular in the architecture of CNNs are the batch normalization layers, used between the convolutional layers and the nonlinearities. These normalize each input channel, speeding up CNN training and reducing sensitivity to network initialization Le Ba et al. [2016].

Finally, a relatively recent regularization technique, called dropout, is also widely used. As LeCun et al. [2015] points out, along with the efficient use of GPUs and data augmentation procedures, this is one of the techniques responsible for the successful use of CNNs in the last decade. A dropout layer randomly *drops* units (along with their connections) from the network during training. This results in a significant reduction of overfitting [Srivastava et al., 2014].

5.4 Virtual Passenger Model: Predicting Annoyance from Time Signals

Sequentially combining the prediction models described on sections 5.2 and 5.3, it is possible to, in a first stage, predict the psychoacoustic metrics with CNNs, as seen in 5.3, and then input those feature predictions in one of the feature-based models from 5.2, thus simulating a VP in a propeller aircraft. This procedure is represented in Fig. 5.2. In

Fig. 5.10 the data used for training each block is specified.

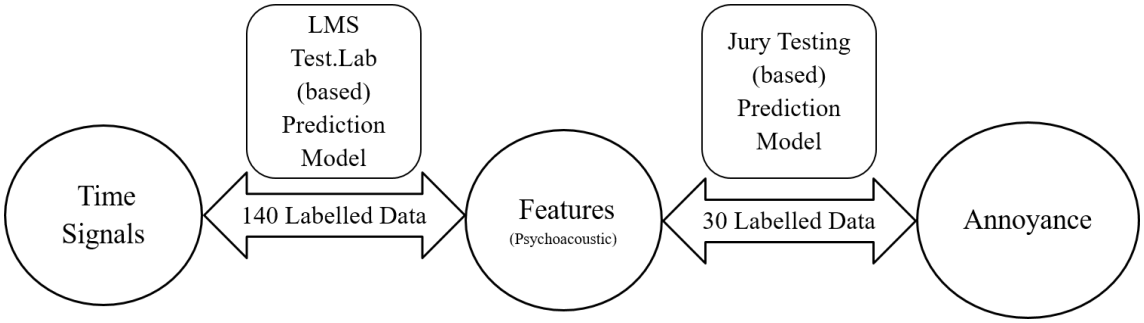

Figure 5.10: Diagram showing the data used for training each block of the VP model.

Keeping in mind that the first block is developed with the 140 sound samples that were not used for jury testing, the complete model performance is assessed by inputting the VP model with the 30 sound samples used for jury testing (thus not used for training the first block) and comparing the annoyance predictions with the original juror responses. Even though the second block of the model was built with these sounds, the features were predicted on the previous block without contact with this sounds, thus being reasonable to measure performance with the original jury testing sounds. This procedure is represented as a flow chart on Fig. 5.11.

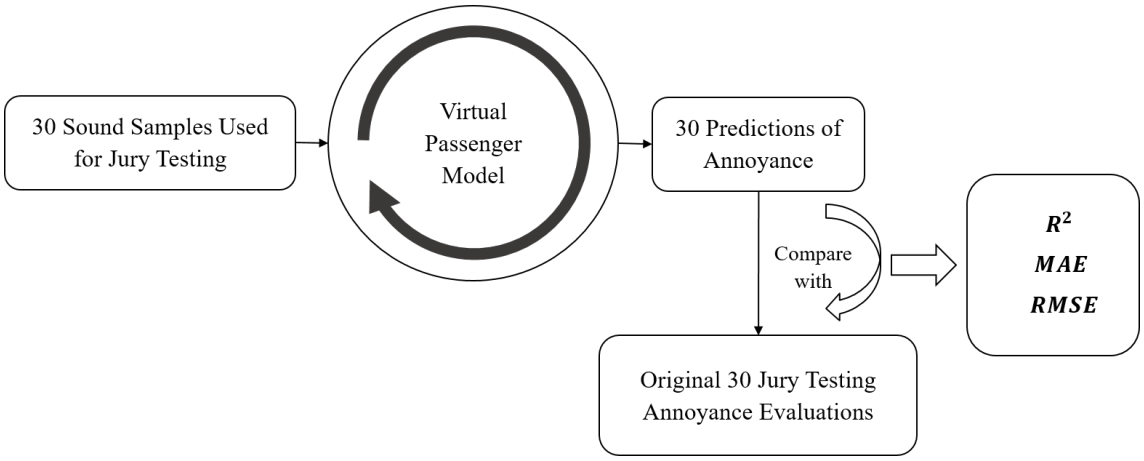

Figure 5.11: Illustration of how to assess performance of VP model.

In opposition to what was done for the second block of the VP model, for the first block of the prediction block the data is not split into testing and training, being the block trained with the entire set of the 140 sound samples. The performance in predicting, for example loudness, is assessed with the jury testing 30 sound samples, being this data yet unseen by this trained block, hence representing a valid testing sample.

Chapter 6

Predicting from Objective Metrics: Results

After defining how the objective psychoacoustic metrics are computed and the methods used to collect subjective evaluations of the sound samples, the performance of the developed prediction models is analyzed throughout this chapter. The best performing model is chosen as the second prediction block of the VP model, illustrated in Fig. 5.2. The results of the campaign for collecting subjective evaluations of sound samples are included.

For facilitating the understanding of the here presented contents, some of the results that were not deemed fundamental for displaying are included in Appendix A. These have a more individual focus on each prediction model allowing to deeper analyze their performance. All the work presented in this chapter was done using MATLAB 2018.

6.1 Subjective Evaluations of Sound Samples

As shown in Ch. 4, each juror classified a sound sample, i.e, stimulus, regarding the anchor sound, by choosing one of the adjectives presented in the software interface. Going from a discrete to a continuous annoyance scale, let 100 correspond to the extreme for maximum annoyance (*Much More Annoying*) and 0 be the extreme for minimum annoyance (*Much Less Annoying*).

Keeping in mind the scale proposed above, in Fig. 6.1 it is possible to find the complete results obtained in the jury testing. They correspond to the 1200 subjective evaluations. These result from 40 jurors, each one evaluating 30 sound samples.

In order to facilitate the interpretation of the results, each annoyance evaluation, x_{ij}, from the ith stimulus and jth juror, was standardized as,

$$Z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{6.1}$$

where μ_j represents each juror average evaluation and σ_j is the standard deviation of each juror. After this operation, each jurors evaluations has mean 0 and a standard deviation of 1.

Also, it was computed, for each stimulus, the average evaluation of all jurors. Thus, a vector of 30 annoyance evaluations was obtained. Finally, each stimulus evaluation, y'_i, was re-scaled as:

$$y_i = 100 \times \frac{y'_i - max(y)}{max(y) - min(y)} \tag{6.2}$$

Note that when converting from the classes to a numeric scale, the range of each class is 17 (in a range from 0 to 100).

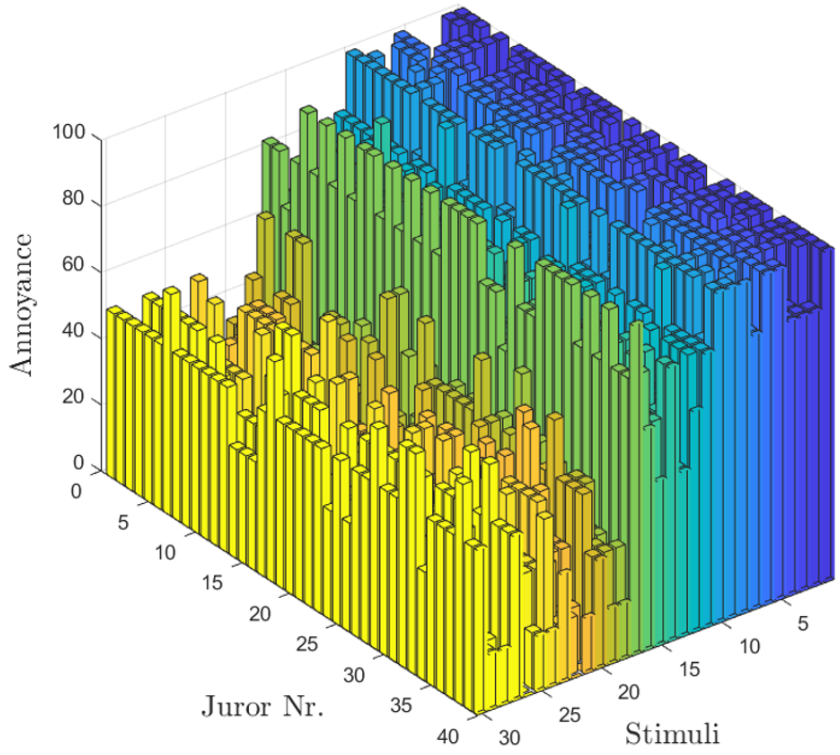

Figure 6.1: Annoyance evaluations provided by each juror for all sound samples.

For the remaining sections of this chapter, Eq. 6.2 was also used to re-scale the psychoacoustic metrics. Therefore, each metric was re-scaled in a way that 100 corresponds to its maximum value and 0 to its minimum value, according to the extreme values of both synchronous and asynchronous flying conditions. These can be found on Tab. 6.1.

Table 6.1: Maximum and minimum values for each psychoacousitc metric, in both synchronous and asynchronous flying conditions

-	Loudness [Sone]	Fluc. Strength [Vacil]	Tonality [T.u.]	Sharpness [Acum]	Roughness [Asper]
Maximum	124.07	1.03	6.96	1.14	0.88
Minimum	48.69	0.27	1.00	0.61	0.01

Tab. 6.2 contains the re-scaled values of the psychoacoustic metrics and corresponding annoyance, allowing to relate each stimuli with the corresponding seat on the aircraft.

Fig. 6.2 contains the final annoyance values for each stimuli, allowing also to see where each stimuli is located within the original (discrete) classes used in the jury testing.

Finally Fig. 6.3 corresponds to a box plot of the final results, where median values are shown in red and mean values are the mid points of the represented boxes. The edges of the boxes correspond to the 25^{th} and 75^{th} percentiles of the population, and extreme data points (outliers) are plotted individually as red signs.

Table 6.2: Subjetive evaluations of sound samples obtained through jury testing and their respective psychoacoustic metrics (s indicates synchronous)

Seat Nr.	Stimuli	Loudness	Fluc. Strength	Tonality	Sharpness	Roughness	Annoyance
1	1	97.02	34.69	77.49	93.09	1.72	97.96
5s	2	90.97	0.00	88.96	87.16	7.96	95.22
5	3	75.20	18.78	77.29	70.87	8.19	91.21
15s	4	92.04	0.06	100.00	90.15	1.93	94.99
20	5	87.01	23.96	91.97	85.86	1.93	94.03
20s	6	100.00	1.98	99.17	100.00	0.79	97.84
25s	7	99.96	6.50	99.92	99.88	0.00	97.46
30	8	64.93	24.13	63.29	63.46	2.94	81.79
1s	9	63.51	6.64	48.95	57.57	6.96	87.96
6	10	98.36	40.03	86.64	94.85	1.93	100.00
35s	11	52.85	39.47	40.12	49.98	5.47	79.12
40s	12	38.03	55.39	28.64	37.64	19.28	68.12
26s	13	17.77	66.50	3.24	17.13	100.00	54.07
37	14	34.66	33.14	29.95	32.06	9.68	71.01
50	15	24.01	84.17	9.59	33.93	62.95	63.80
71	16	24.93	100.00	2.08	37.48	48.43	59.46
22	17	60.45	72.31	54.82	58.90	6.24	92.13
48	18	7.94	82.11	6.76	13.61	34.72	22.07
59s	19	12.74	50.95	16.40	10.33	13.71	33.61
59	20	6.65	54.95	7.75	5.06	23.79	24.79
64	21	5.08	59.13	5.22	4.38	25.30	8.88
64s	22	8.93	42.41	9.39	7.26	17.05	20.24
68s	23	7.37	77.50	1.12	13.04	37.15	25.55
72s	24	11.11	79.32	2.46	14.20	33.60	21.99
73	25	3.38	57.79	1.26	3.87	37.01	6.92
79s	26	0.18	53.10	0.00	0.88	32.81	0.00
81s	27	13.92	71.95	0.00	19.69	36.87	38.87
82	28	0.00	58.44	0.00	0.00	40.70	1.33
84s	29	2.23	41.29	1.50	0.56	25.95	5.67
85s	30	12.90	68.73	0.82	17.25	31.71	42.95

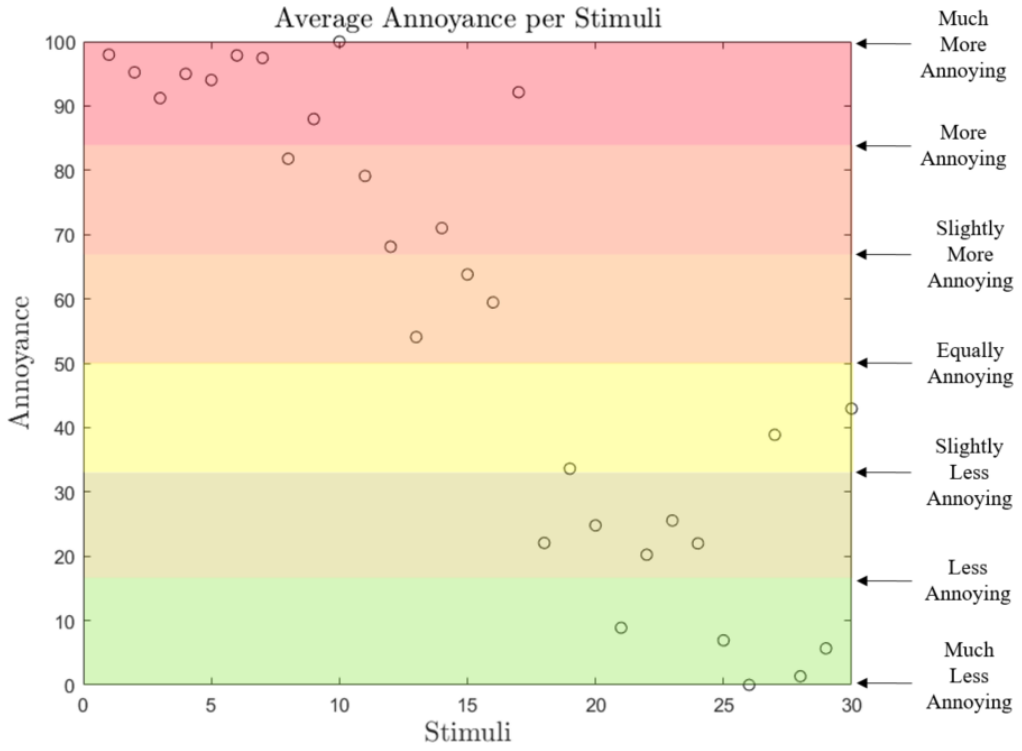

Figure 6.2: Annoyance for each stimuli, with the original classes used in the jury testing evidenced.

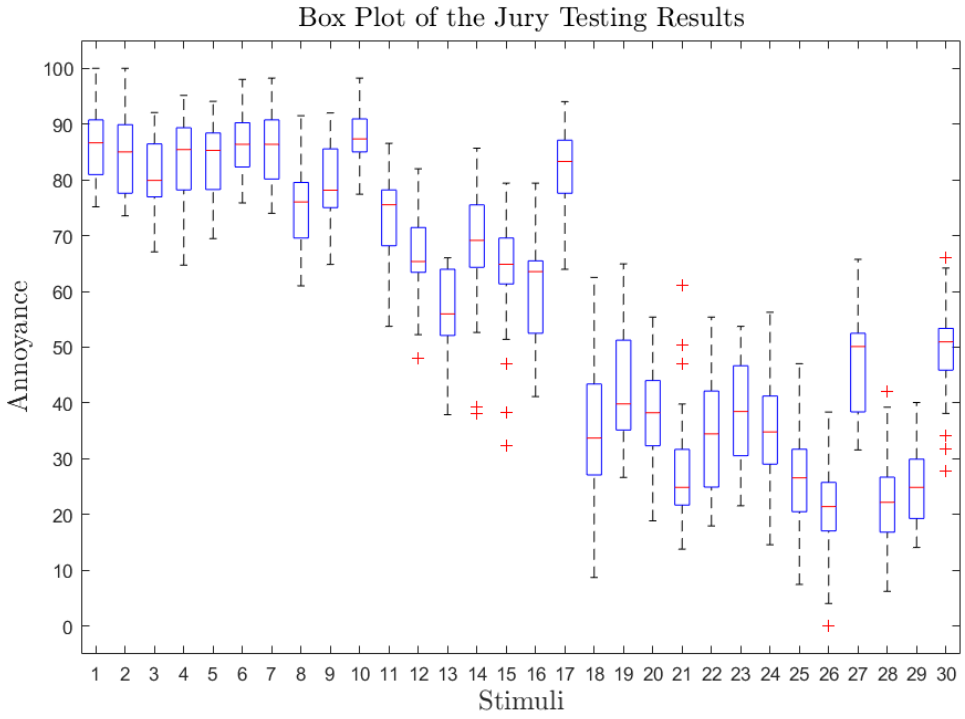

Figure 6.3: Box plot of Annoyance for each stimuli.

By observing the box plot in 6.3, it is possible to note that for low annoyance stimuli (specially the ones ranging from 18th to 30th) there is a greater dispersion of the results. In fact, when conducting the jury testing, most jurors mentioned difficulties particularly in this type of sounds evaluations, i.e, in evaluating sounds with annoyance in mid-low to low values. Before starting to develop the prediction models, it is important to study the correlation between the psychoacoustic metrics and the annoyance obtained for each sound sample, with jury testing. Thus, the Pearson correlation coefficient was computed for each psychoacoustic metric and annoyance and also between the different psychoacoustic metrics.

Table 6.3: Correlation matrix for the objective and subjective metrics used for jury testing

-	Loudness	Fluc. Strength	Tonality	Sharpness	Roughness	Annoyance
Loudness	1	-0.74	0.98	0.99	-0.66	0.93
Fluc. Strength	-	1	-0.81	-0.68	0.70	-0.58
Tonality	-	-	1	0.97	-0.72	0.87
Sharpness	-	-	-	1	-0.62	0.94
Roughness	-	-	-	-	1	-0.53
Annoyance	-	-	-	-	-	1

It is possible to observe from the matrix shown in Tab. 6.3 that the objective metric (feature) with the higher correlation with annoyance is sharpness followed by loudness (almost with an almost equivalent correlation), then tonality, fluctuation strength and finally roughness.

6.2 Prediction Models

As defined in section 5.2, from the data obtained in the jury testing, 4 different types of models will be used to establish prediction models that allow to predict annoyance from the psychoacoustic metrics of new sound samples.

Therefore, in sections 6.2.1, 6.2.2, 6.2.3, 6.2.4 some specific details regarding each model are mentioned. For each one, in order to provide a future user with their performance, Monte Carlo simulation has been run.

The effect of the training data percentage on prediction performance was studied with great emphasis. For each properly tuned prediction model, the average performance (and standard deviation) for 100 random data divisions was computed, using different percentages of training data. Therefore, in section 6.2.5 these results are presented, being shown the mean RMSE and its standard deviation for every 100 random data splits, for the 4 models. On appendix A some plots relating individually to each model are included, which will be mentioned further in this chapter, that allow to observe the models behavior with more detail.

A key aspect that is important to emphasize, is that in a small data set (30 pairs of labeled data) a random data division has a great influence on performance. Therefore, the Monte Carlo simulation method allows to draw more confident conclusions in the hyperparameter tuning process and in comparing performances.

In section 6.2.5, all the generated models are compared. Finally, the best performing model is selected and exploited in section 6.2.6.

6.2.1 Multiple Linear Regression

Keeping in mind the previously presented model formulation, it has no relevant parameters to adjust. The influence of training data on the RMSE was studied. For each different percentage of training data, the data is randomly divided 100 times, being computed the average RMSE and its standard deviation. Additional details from the results of this simulation can be found in 6.4.

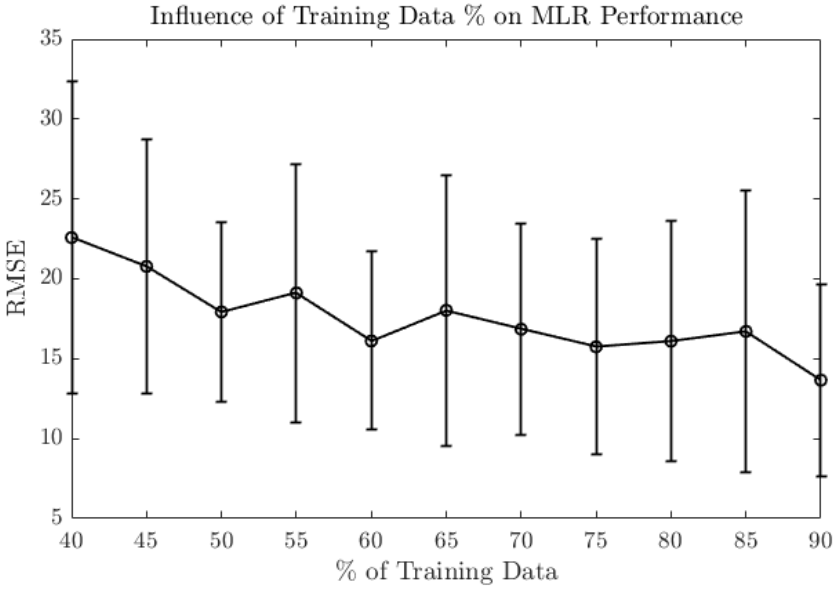

Figure 6.4: Study on the influence of percentage of training in a MLR model using the Monte Carlo method. Each point represents the average RMSE of 100 random data divisions the vertical bar corresponds to its standard deviation.

6.2.2 Artificial Neural Networks

Regarding the ANN, several options for its hyperparameters are available. After trying several types of training algorithms, the two that stood out, due to their stability and performance, were the LM and the BR. Also, the small size and high correlation of the data set imply that no more than 1 hidden layer should be necessary to model the relationship between predictor and response variable.

As previously mentioned, a neural network has a set of weights and bias. These coefficients during training are adjusted in order to find the ones that minimize the prediction error. When creating the neural network, these are randomly selected. The initially arbitrarily chosen weights and bias have a (small) effect on the network performance, hence the adopted methodology is to, for each random data division, train 20 different neural networks and chose the one that performs better, i.e., has a smaller RMSE (computed on the testing data).

In order to decide which training algorithm to chose and how many hidden neurons to use on the hidden layer, the Monte Carlo simulation method was performed, where, for two different training algorithms, the performance corresponding to different numbers of hidden neurons was computed. For each hidden neuron number, 100 random data divisions were performed and the mean RMSE was computed and displayed in Fig. 6.5.

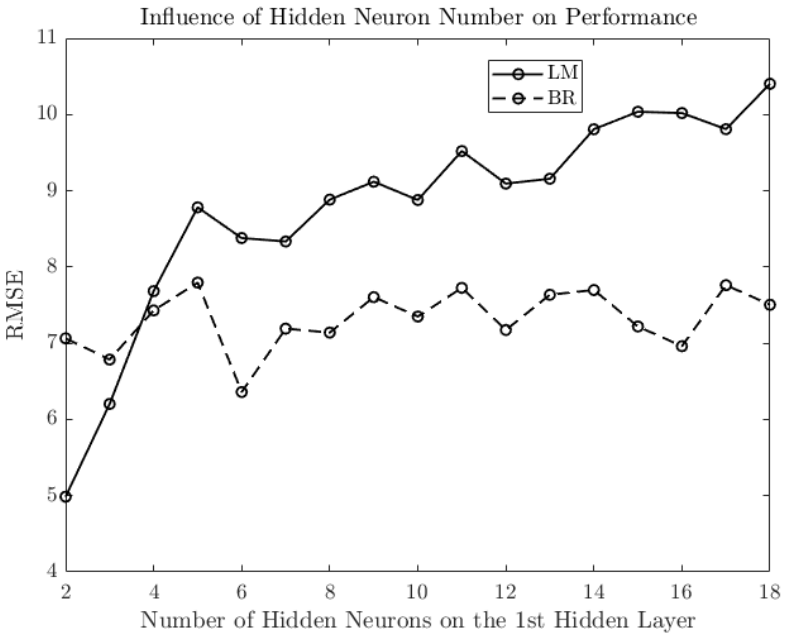

Figure 6.5: Study on the number of hidden neurons influence on performance, for two training types. Each point represents the mean RMSE of 100 random data divisions and the standard deviation of the RMSE in these 100 divisions.

Analyzing Fig. 6.5, it is observable that the best performance occurs for 2 hidden neurons, using the LM training algorithm. Therefore, during the rest of this section, the ANN is always built using 2 hidden neurons and the LM training algorithm.

Being now the ANN hyperparameters properly tuned, it is possible to conduct the same study done for MLR. Therefore, in Fig. 6.6 the individual results of applying the Monte Carlo simulation to study the influence of training data percentage on model performance are shown.

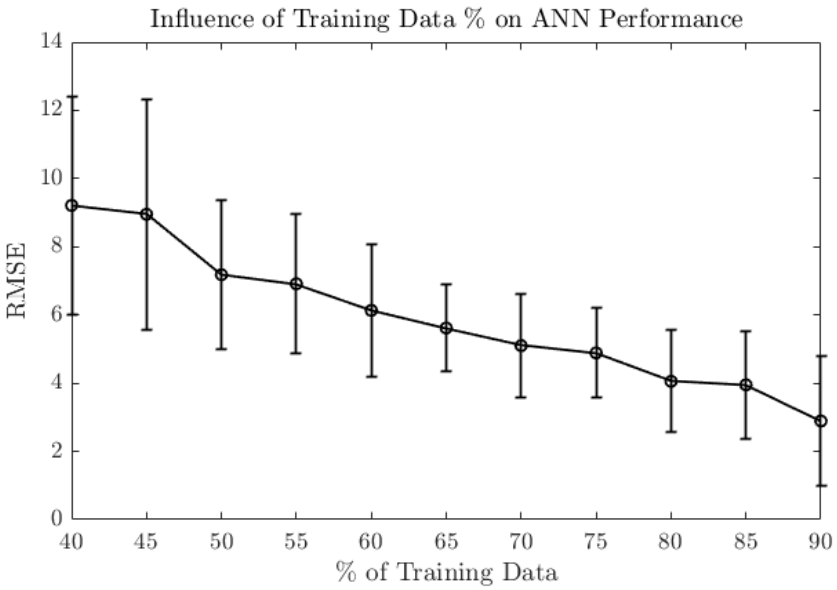

Figure 6.6: Study on the influence of percentage of training in ANN model using the Monte Carlo method. Each point represents the average RMSE of 100 random data divisions the vertical bar corresponds to its standard deviation.

6.2.3 Support Vector Machines

The prediction model developed with SVM also has several hyperparameters to tune. However, due their non-linear relationships, manually tuning an SVM model is harder than to tune ANNs. Therefore, when studying the influence of training data percentage on performance, for each random data division, in the Monte Carlo Method, a bayesian optimization is conducted, being chosen the hyperparameters that better fit each specific data set. The results are displayed in Fig. 6.7.

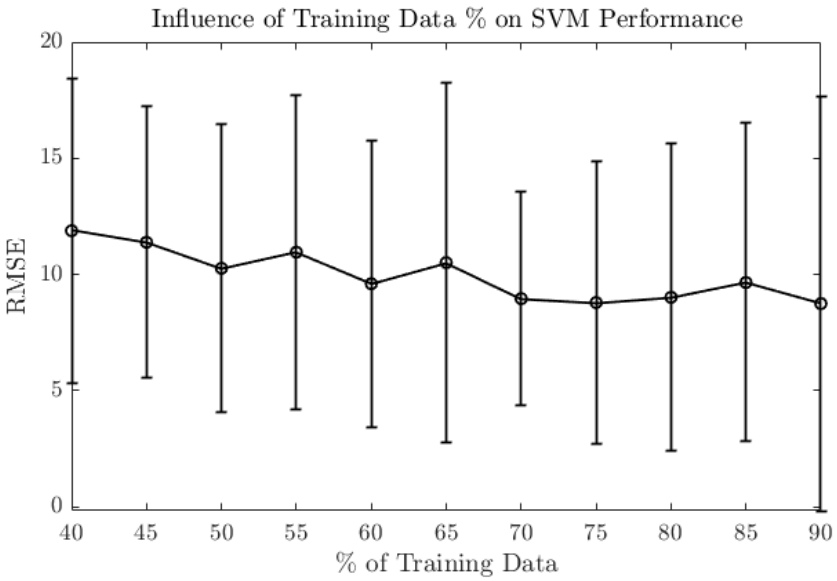

Figure 6.7: Study on the influence of percentage of training in SVM model using the Monte Carlo method. Each point represents the average RMSE of 100 random data divisions the vertical bar corresponds to its standard deviation.

6.2.4 Random Forest

Following a similar process to the one used for SVMs, the hyperparameter tunning for RF was also done using a bayesian optimization. Therefore, for the RF, the same study is presented in Fig. 6.8.

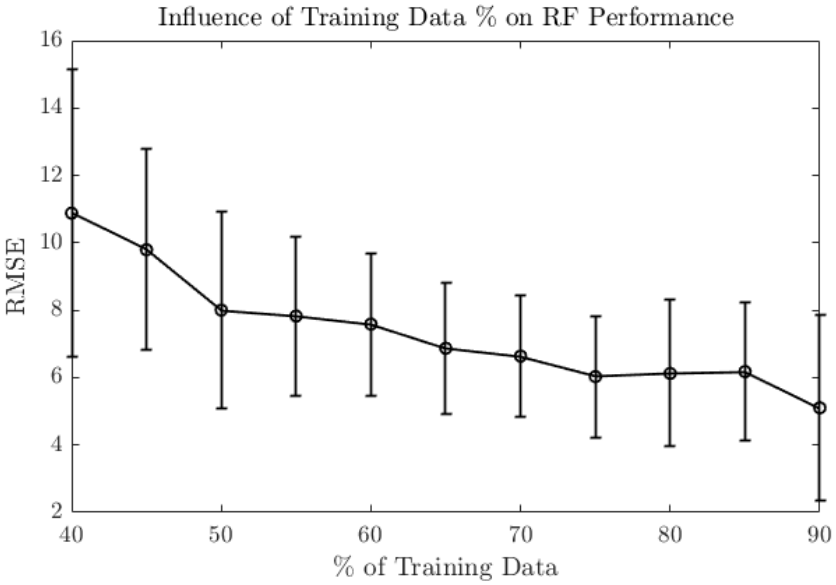

Figure 6.8: Study on the influence of percentage of training in RF model using the Monte Carlo method. Each point represents the average RMSE of 100 random data divisions the vertical bar corresponds to its standard deviation.

6.2.5 Prediction Models Comparison

Remembering that for each model, both the mean RMSE and its standard deviation over 100 random data splits were computed, the results can now be presented. Recalling the study of the influence of training data percentage on performance conducted for each prediction model, it is possible to compile all the obtained results in one plot. This can be found on Fig. 6.9.

Analyzing the obtained results, it is possible to observe that the ANN, for all different percentages of training data, has a lower RMSE, thus performing better. The second top performer model is the RF model, followed by the SVM and then the MLR. Taking into account the stability of each model performance, it is also possible to compare the standard deviation of all the 100 random data divisions done for each % of training data. This is shown in Fig. 6.10.

Considering the results from 6.10, the ANN proves to perform with more stability than the other prediction models, having a smaller RMSE standard deviation than the other prediction models. Regarding the other prediction models, the RF have a RMSE standard deviation close to the ANN, while the MLR and SVM appear to have a RMSE that is highly affected by the stochastic aspect of data divison.

Merely observing the RMSE evolution for all models it is visible that an increase of percentage of training data leads to decrease of the RMSE, thus improving performance. However, one should be careful when increasing this percentage. Keeping in mind that the data-set used contains 30 pairs of labeled data, for example using 90% of the data for training implies that merely 3 sound samples are used for testing, which doesn't qualify as a proper amount of data for measuring performance. Hence, the choice of the amount of data used for training consists in a trade-off between performance and the reliability of its assessment.

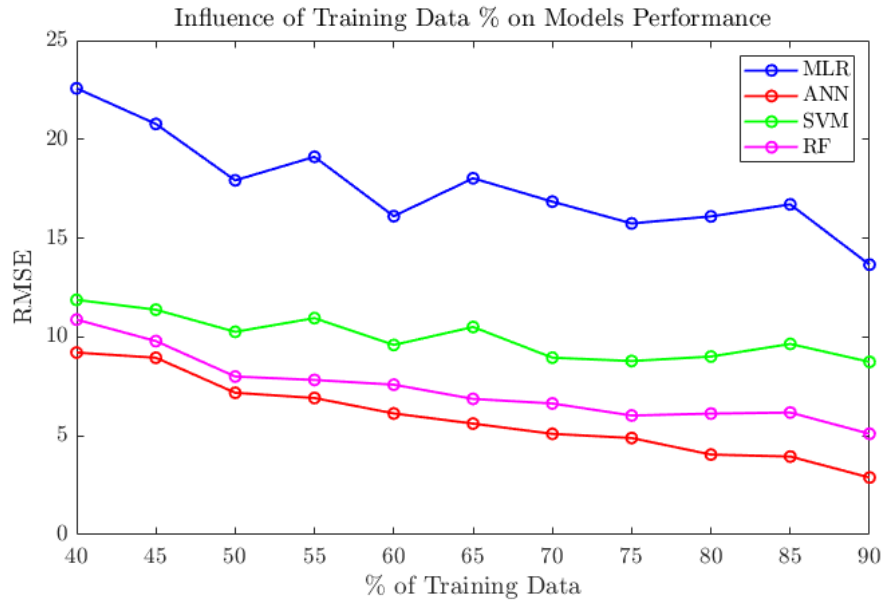

Figure 6.9: Comparison of how the 4 prediction models perform for different percentages of training data, using the Monte Carlo method. Each point represents the mean RMSE of 100 random data divisions.

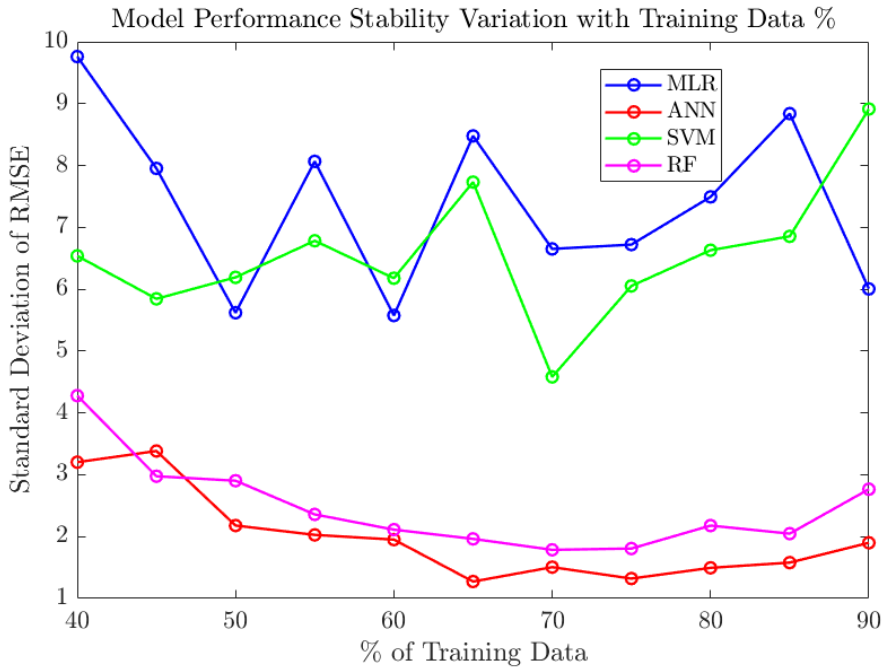

Figure 6.10: Comparison of the standard deviation of the RMSE in 100 random divisions, for different percentages of training data.

In order to continue to analyze the predictions it is necessary to chose a percentage of training data. Based on the obtained results, 70% training data is selected since it allows to obtain good performance and ensures a reasonable amount of stimuli (9) for testing. Also, it is possible to observe that the RMSE standard deviation is smaller between 65% and 75%, hence 70% being a logical choice.

Considering the 100 random data divisions that were done for 70% training data, for each prediction model, the one with the best performance was chosen. Note that, for example, for the MLR, the RMSE was lower at randomization number 22 while for the ANN, the lowest RMSE occurred at randomization number 18. So, for showing the results of the prediction models, the best performing model for each one was selected. Their performance metrics are presented in Tab. 6.4.

For each model, three plots are presented, using as comparison the jurors original averaged response. First the predictions are plotted along with the juror responses, also including the error for each stimuli (being this the predicted minus the mean juror response). Secondly, a correlation between predicted and mean juror response is presented. Figs. 6.11 and 6.12 contain the results for the MLR. Figs. 6.13 and 6.14 refer to the ANNs. Figs. 6.15 and 6.16 are related to the SVM. Figs. 6.17 and 6.18 are associated with the RF model. Fig. 6.19 contains the four models predictions compared with the original responses of the 40 jurors.

It is important to comment that the observable performance variation with the different randomizations is partially due to the particular set of testing data chosen. Indeed, analyzing the results it is noticeable that some stimuli are more *easy* to predict than others. For example, if the data used for training contains only low annoyance sounds, for the model to predict on additional low annoyance sounds can be considered as *easy*, but it will be *hard* to predict on high annoyance sounds. Therefore, the randomization with superior performance is possibly the one that uses the *easier* stimuli as testing data (or also the one that uses the most representative set of stimuli for training).

Re-using a previously used analogy, two scenarios may explain this situation: considering a student taking an exam with, for example 9, questions and having studied through 21 solved exercises. His good performance is due to either having studied (i.e. trained) on a representative set of 21 exercises or the exam only evaluating his answers on easy questions. Therefore, to evaluate the average performance of the student over 100 different exams (i.e. the prediction model) allows to draw more exact conclusions on his performance, being the Monte Carlo simulation results the ones that should be taken into account when assessing the models performance.

Table 6.4: Performance in 100 random data divisions (70% data for training)

(a) Averaged performance

-	R^2	MAE	$RMSE$
MLR	0.862	12.145	15.323
ANN	0.98185	3.851	5.018
SVM	0.92733	7.086	8.637
RF	0.972	4.783	6.216

(b) Performance with the best RMSE

-	R^2	MAE	$RMSE$
MLR	0.9695	5.966	7.263
ANN	0.9964	1.657	2.103
SVM	0.9946	3.761	4.084
RF	0.9973	2.010	2.264

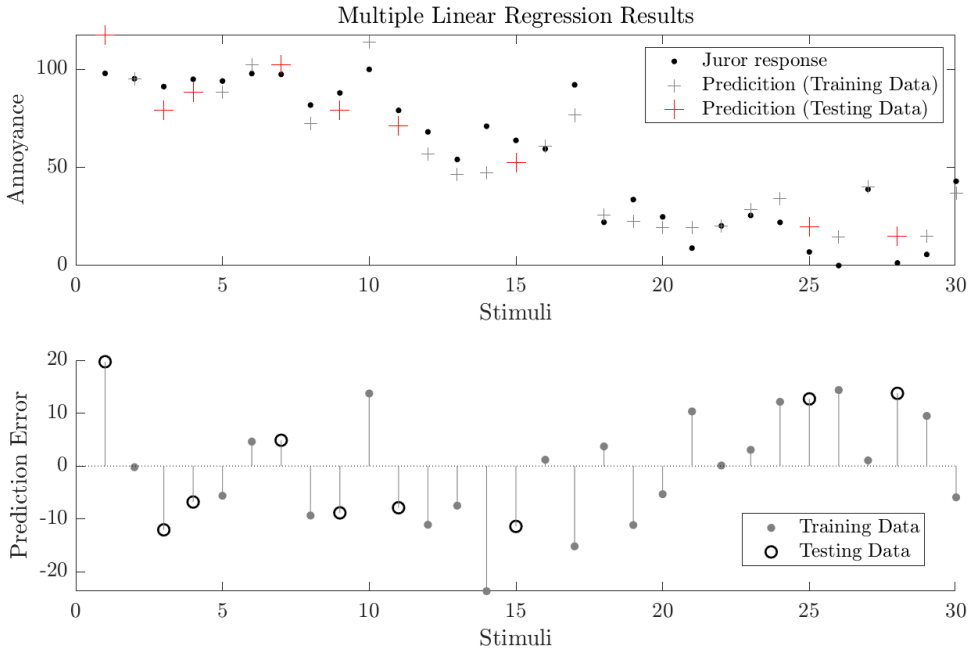

Figure 6.11: Predictions of the best performing MLR throught 100 random data divisions, with 70% data for training.

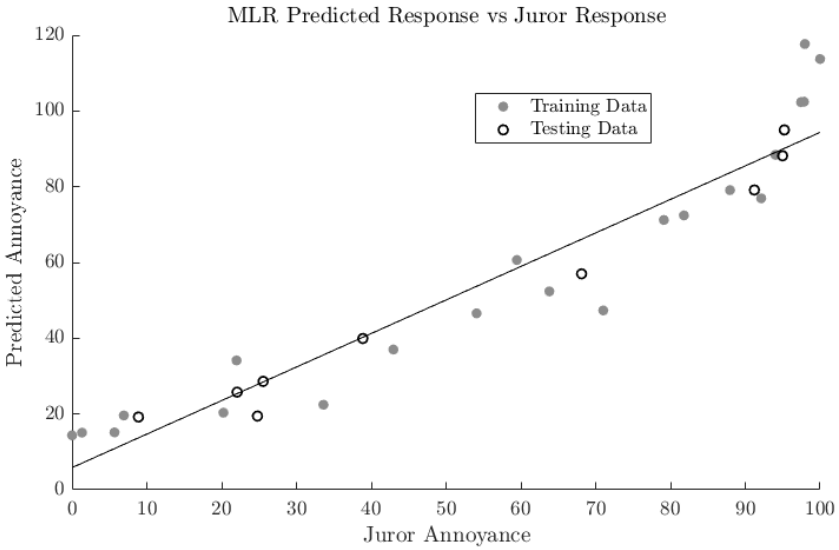

Figure 6.12: Correlation analysis of the best performing MLR throught 100 random data divisions, with 70% data for training.

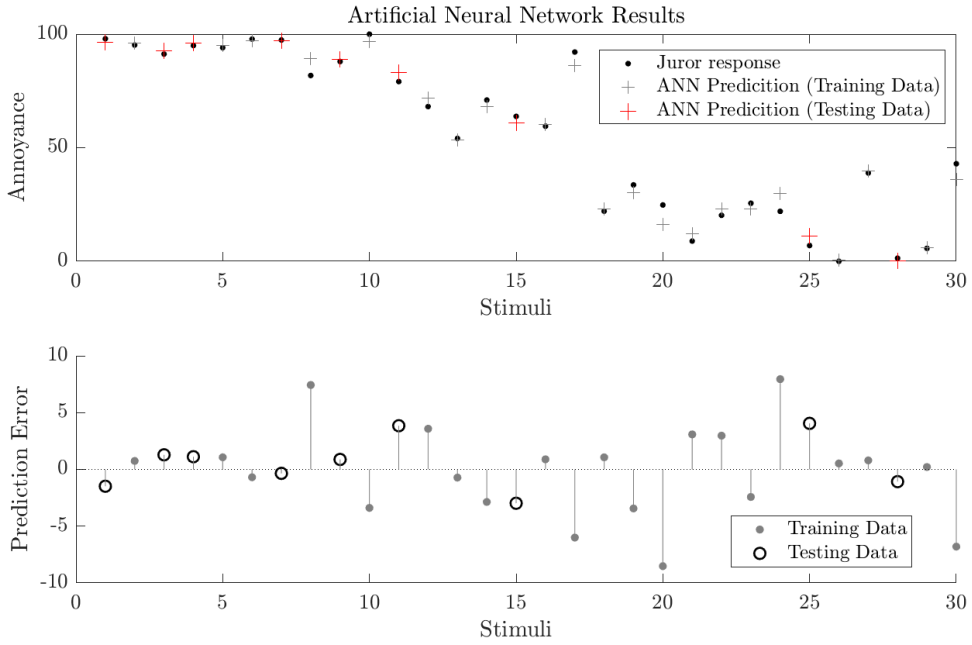

Figure 6.13: Predictions of the best performing ANN throught 100 random data divisions, with 70% data for training.

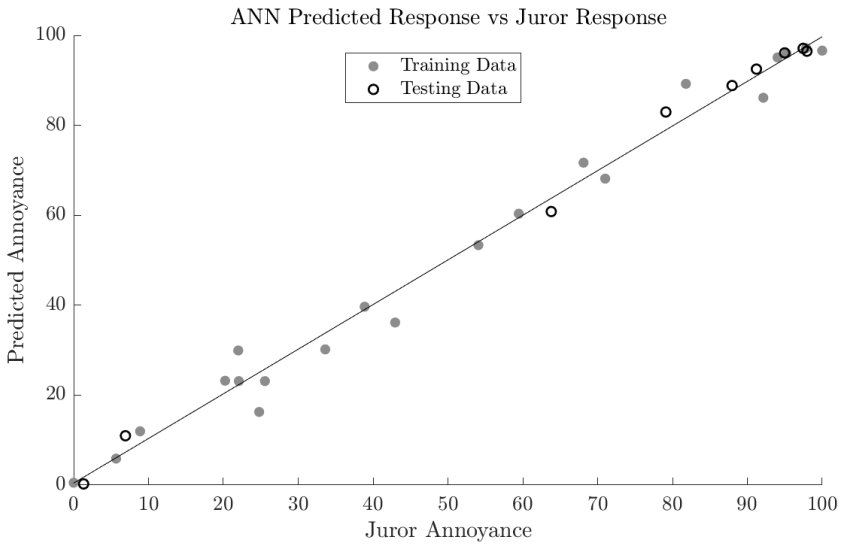

Figure 6.14: Correlation analysis of the best performing ANN throught 100 random data divisions, with 70% data for training.

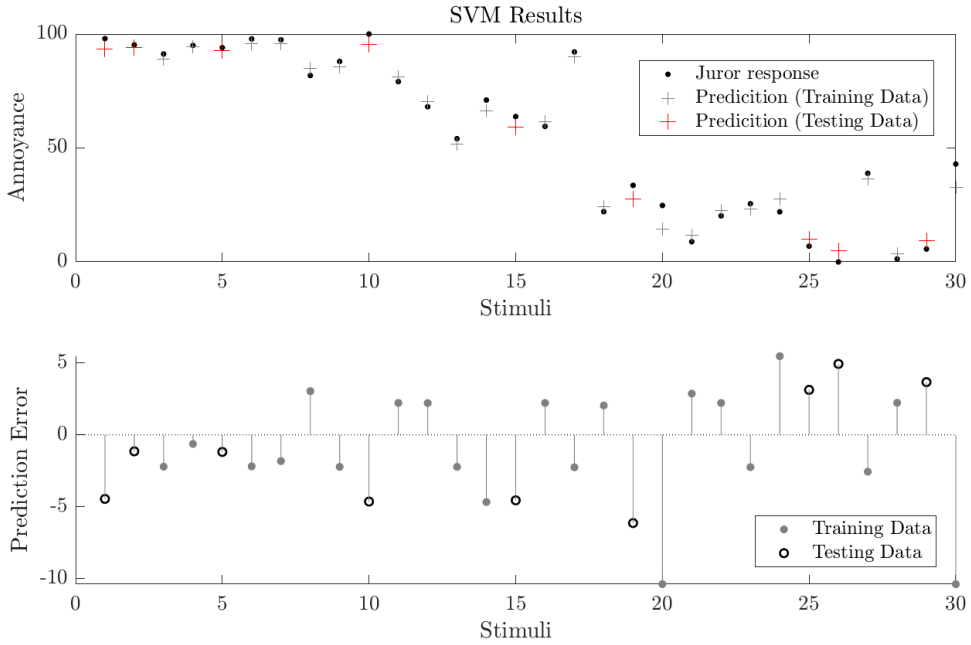

Figure 6.15: Predictions of the best performing SVM throught 100 random data divisions, with 70% data for training.

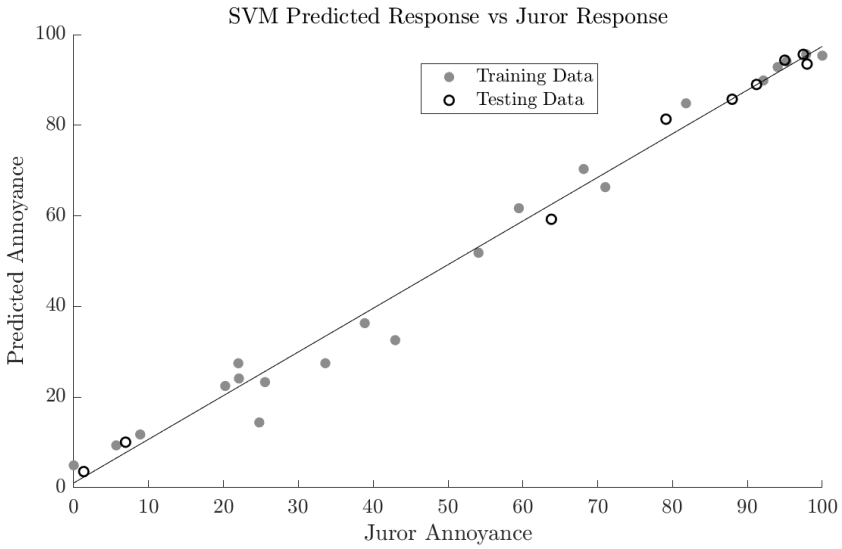

Figure 6.16: Correlation analysis of the best performing SVM throught 100 random data divisions, with 70% data for training.

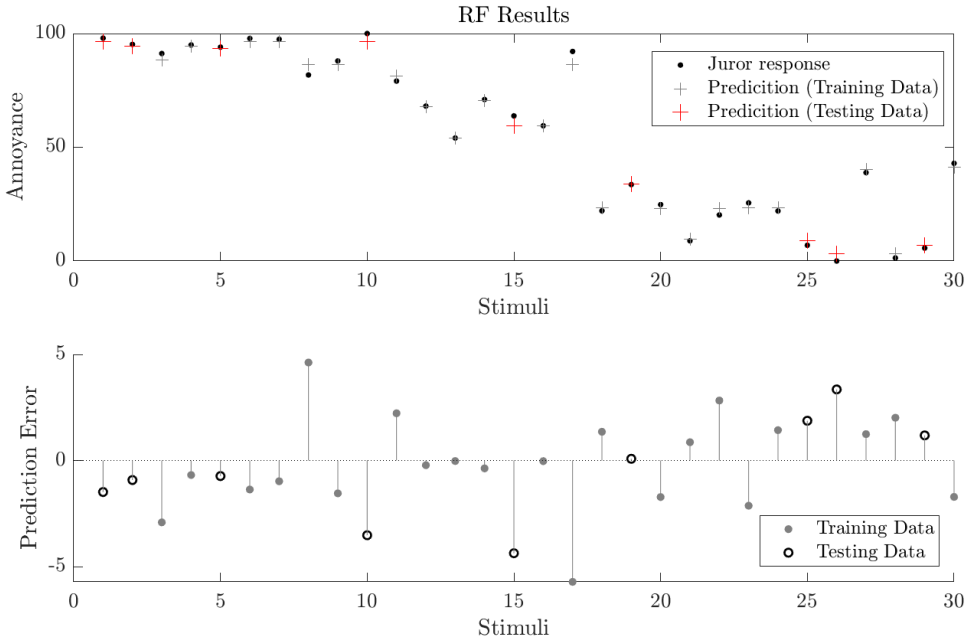

Figure 6.17: Predictions of the best performing RF throught 100 random data divisions, with 70% data for training.

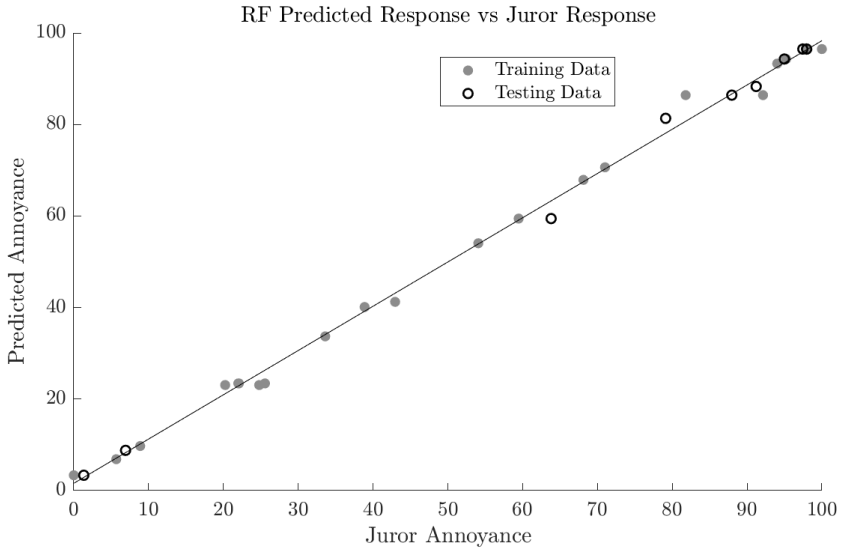

Figure 6.18: Correlation analysis of the best performing RF throught 100 random data divisions, with 70% data for training.

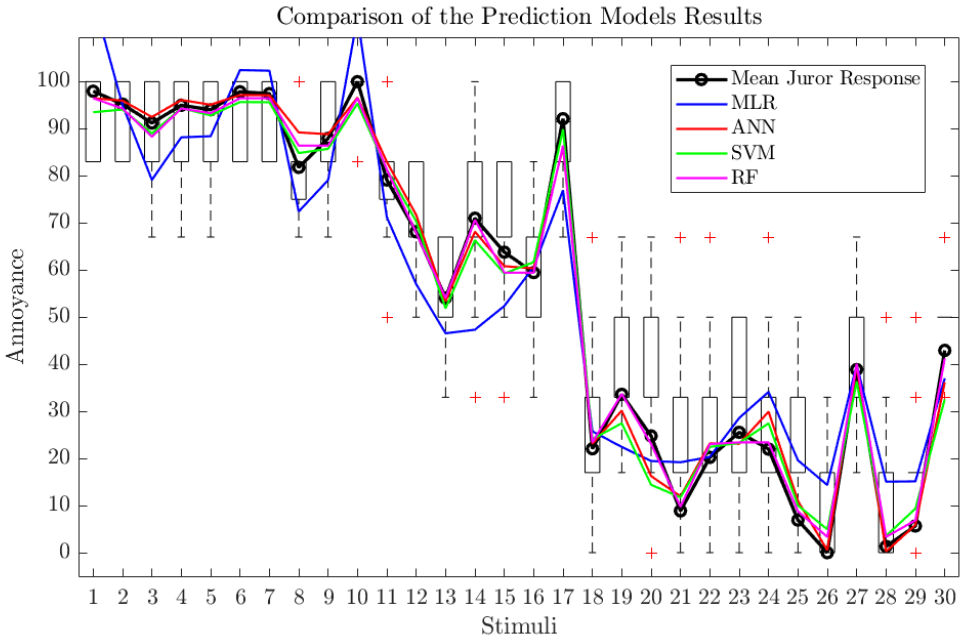

Figure 6.19: Comparison of the models predictions with the original juror responses.

Study of the Models Response

Having already the trained prediction models, it is possible to study their response to a manual alteration of the objective metrics. This study was done as an experimental approach, with the mere goal of trying to check if by just manually changing one of the metrics, the models respond in a physically logical way (according to the psychoacoustics principles). For example, a big increment in loudness should result in an increase of annoyance. Obviously, this is a very simplistic approach and is performed in the context of a *what if?* reasoning. It should be noted that to change only one metric is not a physically realistic situation for the models, therefore one should be careful in drawing conclusions from this. One of the motivations for performing this analysis is to observe how the different models respond to the same variations in the inputs.

Considering the seat with annoyance closer to the middle of the range (so stimulus 13 with an annoyance of 54.07), each one of its metrics is sequentially altered from 0 to 150, and the prediction model response is registered. The results, for each objective metric are presented in Figs. A.1, A.2, A.3, A.4 and A.5.

Recalling the correlation matrix, presented in Tab. 6.3, it would be expected that the relationships between each one of the metrics and annoyance are correspondent with the models response. For example, it is expected that for objective metrics with high positive correlation with annoyance, an increase on the objective metrics would result in an annoyance increase. Also, intuitively an extremely loud sound should be more annoyant.

Observing Figs. A.1 and 3.3, related with loudness and sharpness, the ANN is the model that appears to model more logically their relationship with annoyance. Effectively, both these metrics have a very high correlation with annoyance, and from a psychoacoustic point of view, it is reasonable that both very loud and sharp sounds are highly annoyant, as the ANN models it (where a positive increment in annoyance is leading to a continuous loudness and sharpness increase). For example, according to the SVM model, for the top range of loudness or sharpness, annoyance decreases, which is not what would be expected. About the other objective metrics, due to their complexity in a psychoacoustic point of view, it is quite hard to draw any valid conclusions from the results. As expected, due to the nature of these models, the MLR is limited to a linear response and the RFs respond in *steps* of annoyance for the input variations.

Considering the entire analysis conducted throughout this section and all the results obtained, the prediction model selected for including in one of the blocks of the complete prediction model (7.2) is the **Artificial Neural Network**, with the already specified hyperparameters. This decision was based fundamentally on its superior performance and better stability for different random data divisions and percentages of training data.

6.2.6 Annoyance Spatial Distribution in the Propeller Aircraft

Being the ANN chosen in the previous section, with 2 neurons on the first hidden layer and trained with the LM algorithm (using 70% data for training), it is possible to exploit it, predicting annoyance for the remaining seats of the aircraft in both asynchronous and synchronous flying conditions. Hence, the annoyance can be spatially mapped in the aircraft using the ANN, being that from the annoyance values in each seats, the annoyance for each point of the fuselage is interpolated. This is shown in Fig. 6.20, as color maps, for both flying conditions, where the bottom of the image corresponds to the front of the aircraft.

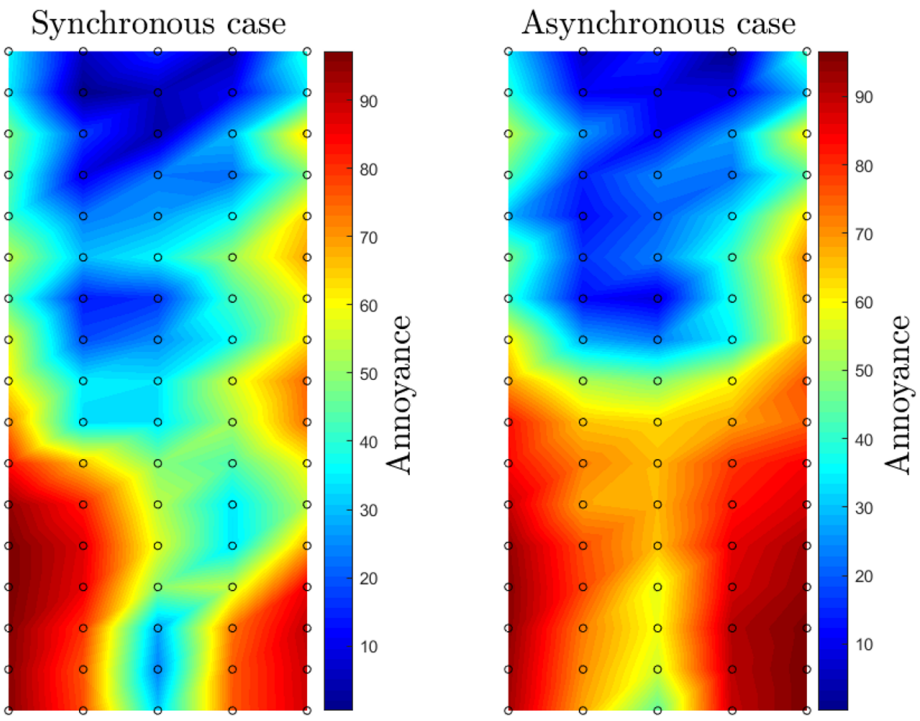

Figure 6.20: Annoyance prediction for all the seats of the aircraft, using the trained ANN.

Analyzing Fig. 6.20, it can be seen that annoyance, in the interior of the aircraft cabin, has a great degree of variation in this space and this variations occurs in specific zones, being possible to establish a correspondence between the physical components of the aircraft and the predicted annoyance. Both in the synchronous and asynchronous the higher annoyance occurs in the bottom part of the color maps near the lateral borders. This zones correspond to the seats closer to the engines of the aircraft and as expected, seats near the engines are more acoustically uncomfortable.

As stated by Wilby [2008], the blades passing on the fuselage result in airborne noise. Hence, periodic pressure fluctuations are produced on the external part of the fuselage causing vibration in the interior part of the cabin and an excitation of the interior sound field. This phenomena has a greater intensity near the engines, thus it is physically expectable that this excitation of the interior cabin sound field leads to higher annoyance values. Thus, it is then possible to state that the obtained annoyance spatial mapping is coherent with the the aircraft physics, being this an argument in favor for validating the obtained predictions.

However, it should also be held into account that the obtained spatial mapping is quite similar to the one obtained for loudness and sharpness, shown in Ch. 5. Remembering that both these metrics have the highest correlation with annoyance, it is natural that the annoyance spatial mapping resembles the mapping obtained for loudness and sharpness, which also is coherent with the propeller aircraft physics.

As a final comment, regarding the annoyance distribution in the synchronous case, it is observable that in the cabin area near the right engine there is an increase of annoyance, and the opposite happens for the left engine. A possible explanation for this is that, due to the fact that, even though the rotation frequency of both propellers is coincident,

they are rotating in opposite directions. Hence, due to the opposite relative speeds of the propellers, the creation of a lateral air flow may result in the pressurization of the fuselage area that shows the annoyance increase. This has as a result the formation of a turbulent boundary layer and the appearing of vortexes, which may be exciting a cavity mode thus resulting in the asymmetry on the annoyance distribution for the synchronous case.

Chapter 7

Predicting from Sound Samples: Results

This chapter is devoted to presenting and analyzing the results of the VP model, that from a sound sample (inputted as a time signal) outputs a subjective evaluation (annoyance).

Recalling Ch. 5, this model is composed of two blocks: a first one that receives time signals as an input and outputs features and a second one, which is a feature-based block, that from the first block features (objective psychoacoustic metrics) outputs annoyance and whose results have already been analyzed in Ch. 6. After analyzing 4 alternatives for the second block, an ANN was chosen due to its best performance, being used throughout this chapter for the complete VP model, as decided in the previous chapter.

Regarding the first block, it is necessary to develop 5 CNN based prediction model (one model for each one of the 5 features), whose results and respective analysis can be found on the following section.

All the work presented in this chapter was done using MATLAB 2018.

7.1 Predicting Objective Metrics from Sound Samples

As detailed on section 5.3, CNNs are used for predicting features from time signals in the first block of the VP model. So, in this section, 5 CNNs are trained for predicting, from raw time signals the 5 psychoacoustic metrics previously used as inputs. The sounds samples used in this section were the ones that were not used for the jury testing. From the global set of 170 sound samples (85 for synchronous and 85 for synchronous), 30 were used for jury testing, hence having 140 stimuli for training this 5 CNNs.

Remembering that the metrics considered are loudness, fluctuation strength, tonality, sharpness and roughness, the models performance after training was assessed inputting it with the sound samples used on jury testing, thus the targets for performance being the objective metrics on Tab. 6.2. One should note that these stimulus used for jury testing were selected based on a cluster analysis on the psychoacoustic metrics, thus being this a comprehensive and representative testing set for all the psychoacoustic metrics.

In order to decrease the computing time necessary for training the models, the time signals were downsampled using the Signal Processing Toolbox from MATLAB. Thus, they were resampled from a sampling frequency of 44100 Hz to 8820 Hz.

Due to the fact that 5 CNNs have to be trained, each one with a different data set, different architectures and hyperparameters have to be selected in order to have a good performance. However, unlike the ANNs used on the previous section, training CNNs on raw time signals has a higher computational cost, so the process used on the previous section, using the Monte Carlo method is not viable due to time constraints. However,

considering the 30 testing stimuli are a representative sample of the whole set of 5 psychoacoustic features, a proper assessment of performance is still ensured.

The tuning of the hyperparameters was done using bayesian optimization. Two architecture options were manually chosen, iterating between different layers until finding the stack that provides the best performance on the jury testing stimuli. The option more adequate for each metric was chosen using a bayesian optimization process. The selected architecture for each feature and the tuned hyperparameters can be found on Tabs. 7.1 and 7.2 respectively.

On Tab. 7.3 the obtained performance for each CNN based model is presented, scaled from 0 to 100 and also reconverted in each psychoacoustic metric original scale.

Observing the analyzed results, the prediction model with the superior performance is the one that predicts tonality, followed by loudness, sharpness, roughness and fluctuation strength. It is possible to state that for loudness, sharpness and roughness the performance is evidently superior than for fluctuation strength and roughness.

Similarly to what was done on Ch. 6, first the predictions are plotted along with the LMS Test.Lab results, also including the error for each stimulus (being this the predicted minus the software computed metric). Secondly, a correlation between predicted and the LMS Test.Lab results, for each feature, is presented. Hence, Figs. 7.1 and 7.2 contain the results for loudness, 7.3 and 7.4 correspond to fluctuation strength, 7.5 and 7.6 are related with tonality, 7.7 and 7.8 are associated with sharpness and finally 7.9 and 7.10 illustrate the prediction results for roughness. Note that it was decided to, in these figures, use the features scaled from 0 to 100, for allowing a better comparison between the prediction performance of the different metrics. Remember that with this type of scaling 0 and 100 correspond to the minimum and maximum value of the metrics in synchronous and asynchronous flying conditions, respectively.

Table 7.1: Architectures used for predicting psychacoustic metrics from time signals

(a) Fluctuation Strength and Sharpness

Image Input Layer
Convolutional Layer
Batch Normalization Layer
ReLU Layer
Average Pooling Layer
Convolutional Layer
Batch Normalization Layer
ReLU Layer
Convolutional Layer
Batch Normalization Layer
ReLU Layer
Dropout Layer
Fully Connected Layer
Regression Layer

(b) Loudness, Tonality and Roughness

Image Input Layer
Convolutional Layer
Batch Normalization Layer
ReLU Layer
Average Pooling Layer
Convolutional Layer
Batch Normalization Layer
ReLU Layer
Dropout Layer
Fully Connected Layer
Regression Layer

Table 7.2: Optmized hyperparameters for each prediction model

-	Initial Learning Rate	Gradient Decay Factor	L2Regularization
Loudness	0.00040170	0.87164	9.5025×10^{-07}
Fluctuation Strength	0.00042896	0.86859	2.9361×10^{-07}
Tonality	0.00067127	0.88906	3.2236×10^{-07}
Sharpness	0.00062097	0.91398	1.5002×10^{-07}
Roughness	0.00040087	0.86568	1.0705×10^{-07}

Table 7.3: Performance when predicting psychoacoustic metrics

-	R^2	MAE [0-100]	MAE	RMSE [0-100]	RMSE
Loudness	0.9094	8.1632	6.1531 Sone	10.9950	8.2879 Sone
Fluctuation Strength	0.6906	12.7650	0.0971 Vacil	15.2750	0.1162 Vacil
Tonality	0.9407	7.4027	0.4410 T.u.	10.2450	0.6102 T.u
Sharpness	0.8863	9.3928	0.0496 Acum	11.9310	0.0630 Acum
Roughness	0.6345	8.5497	0.0741 Asper	13.8610	0.1201 Asper

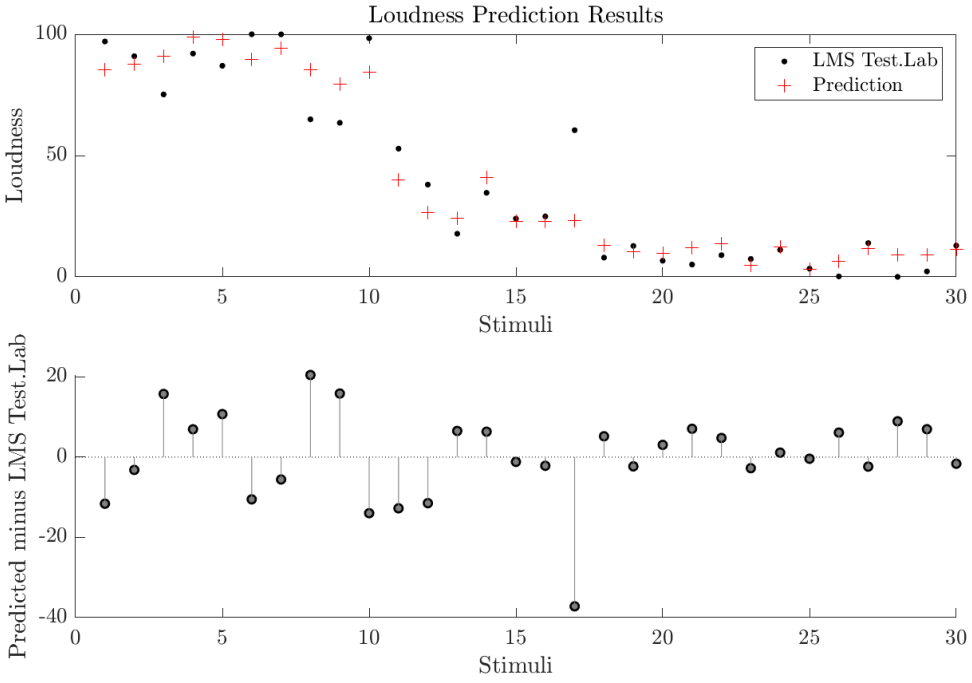

Figure 7.1: Loudness prediction results.

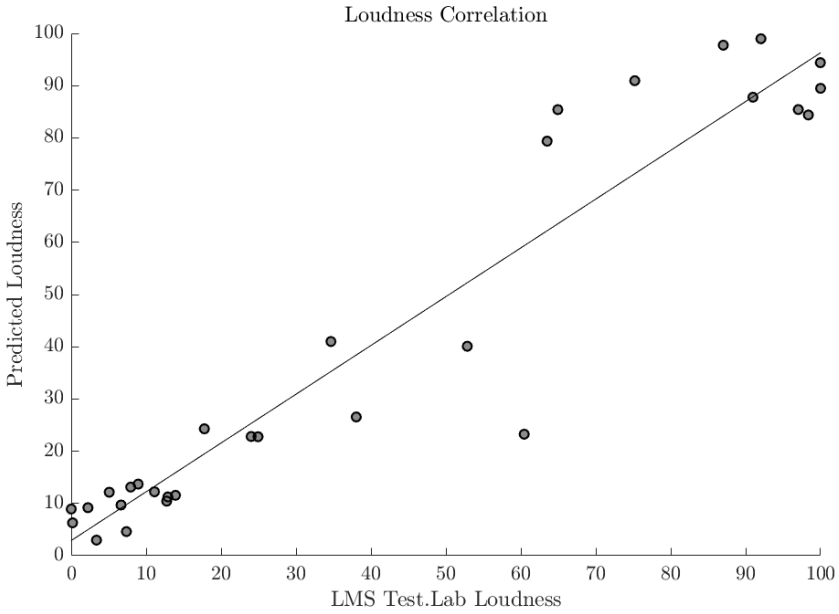

Figure 7.2: Loudness correlation analysis.

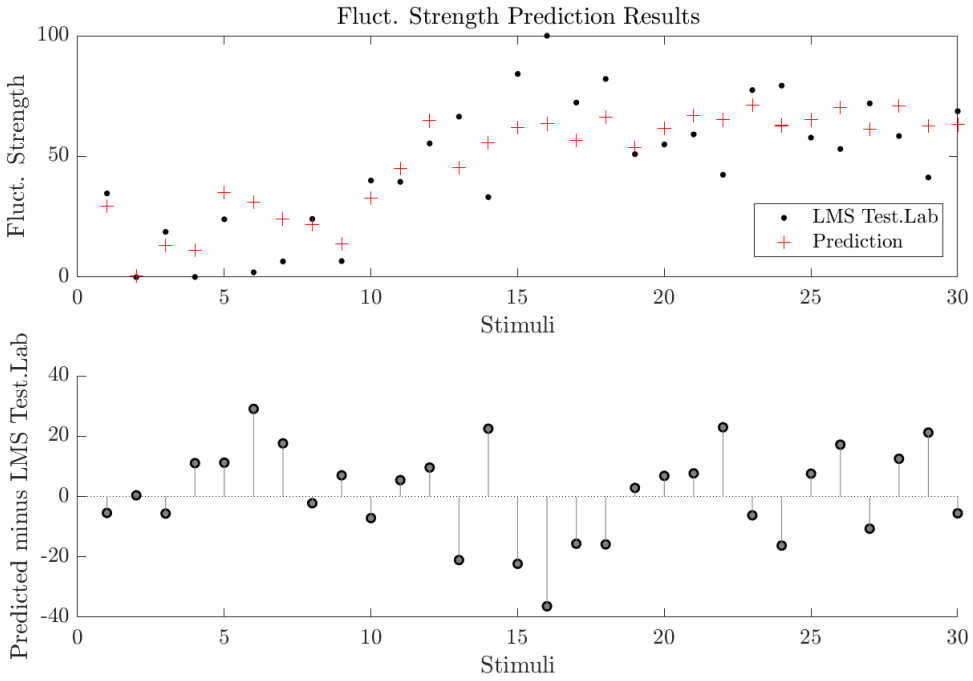

Figure 7.3: Fluctation Strength prediction results.

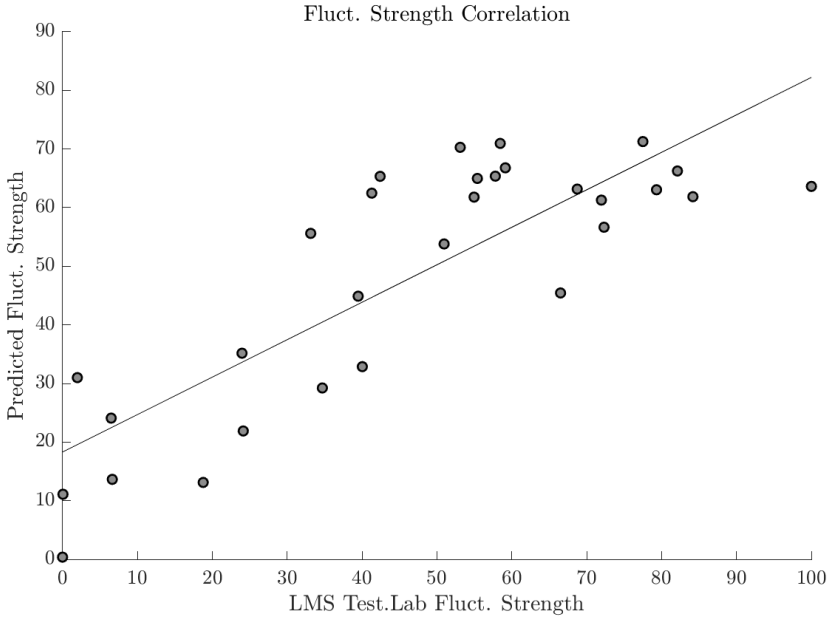

Figure 7.4: Fluctation Strength correlation analysis.

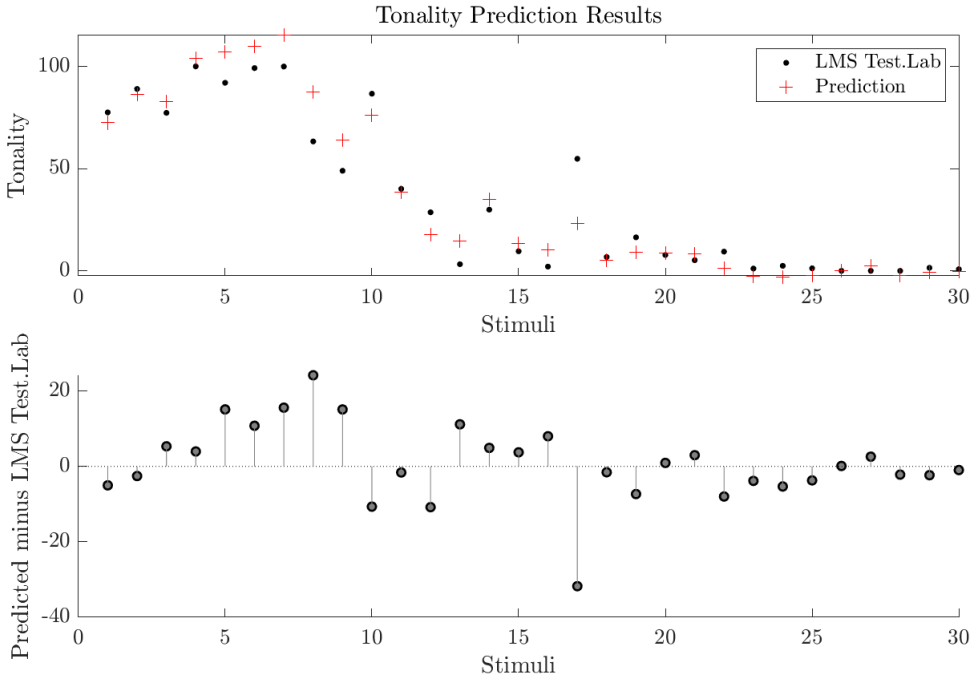

Figure 7.5: Tonality prediction results.

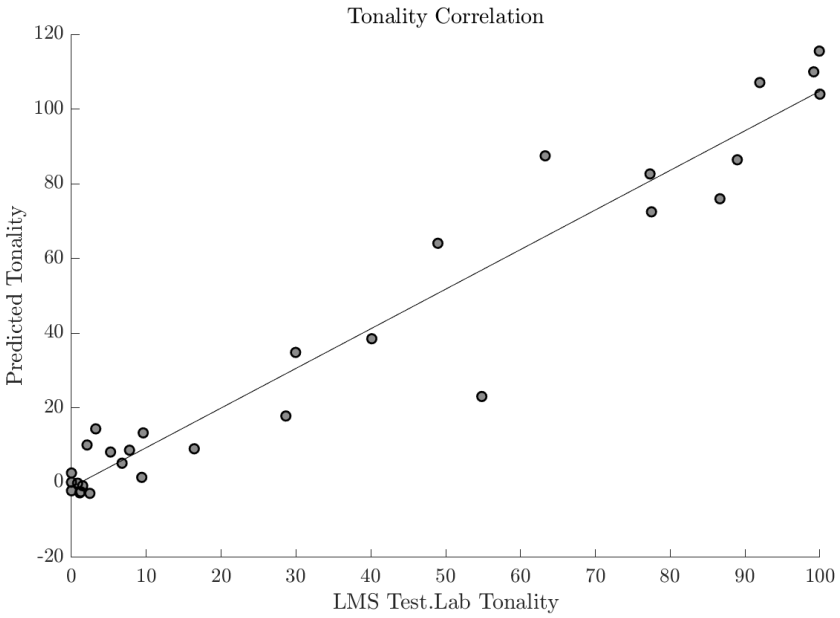

Figure 7.6: Tonality correlation analysis.

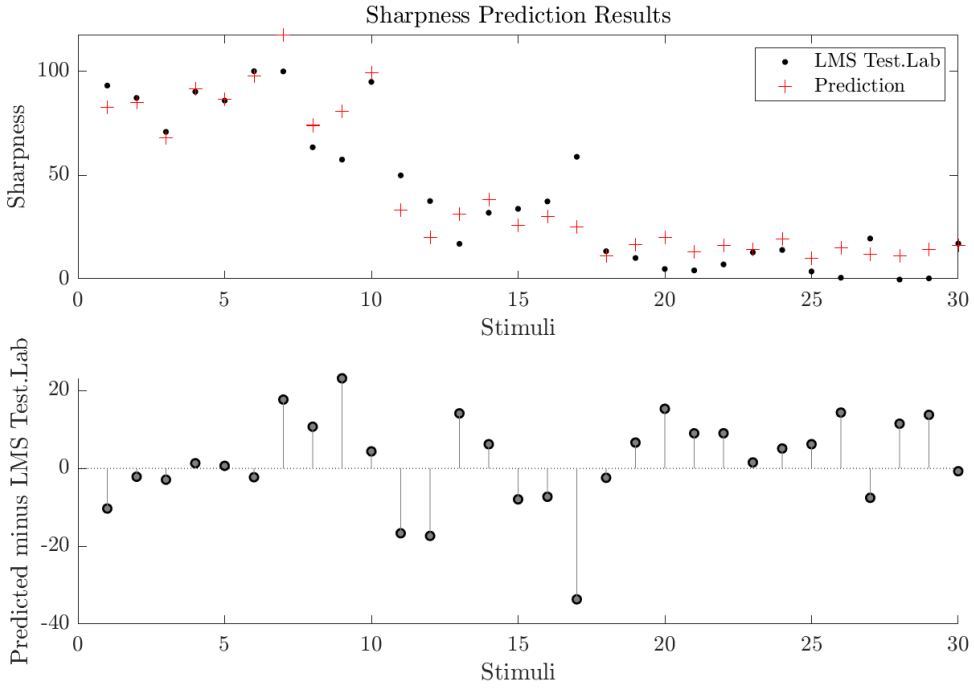

Figure 7.7: Sharpness prediction results.

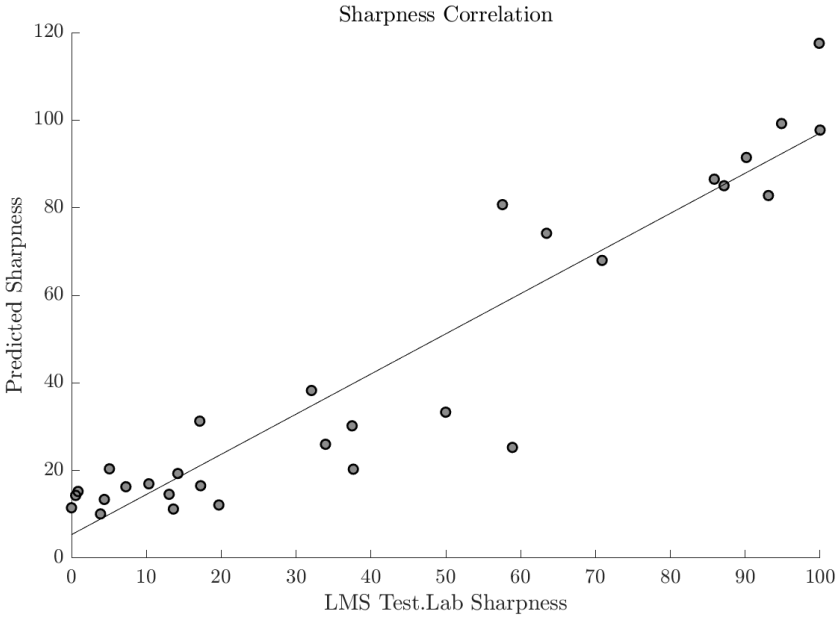

Figure 7.8: Sharpness correlation analysis.

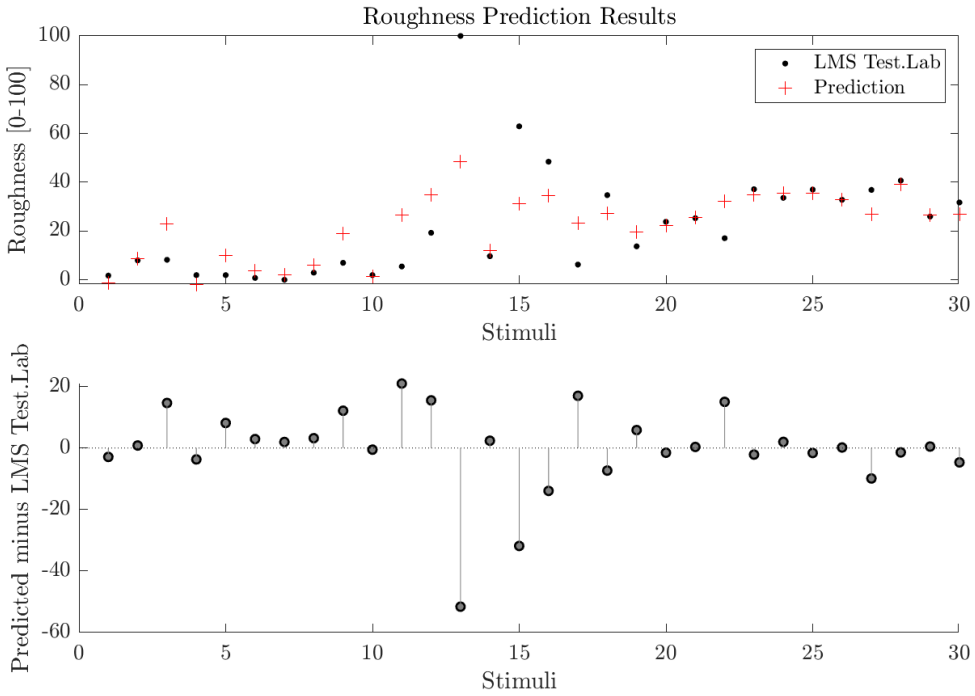

Figure 7.9: Roughness prediction results.

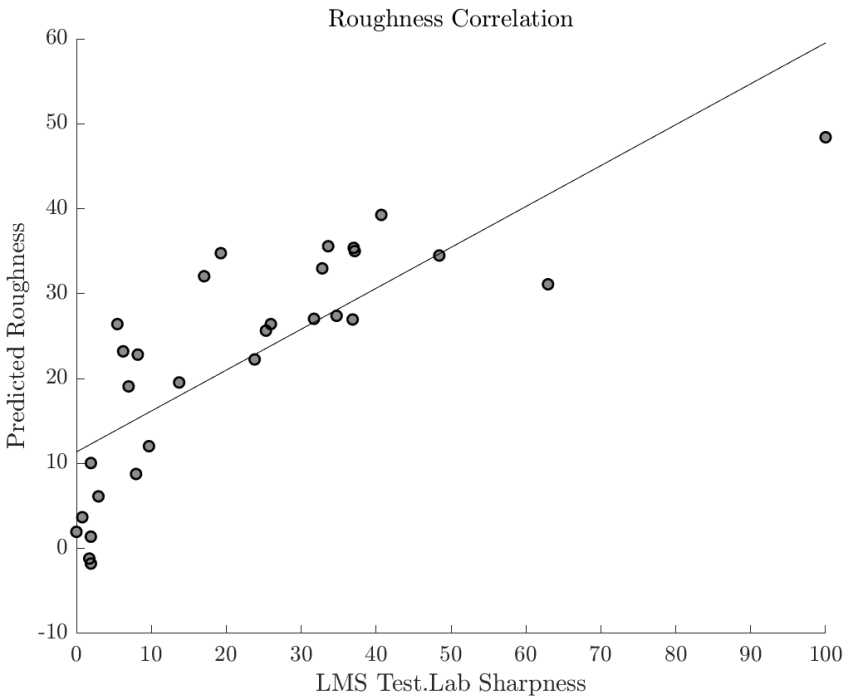

Figure 7.10: Roughness correlation analysis.

7.2 Virtual Passenger Model: Predicting Subjective Metrics from Sound Samples

Having now developed both blocks of the complete prediction model, it is possible to combine them having ready the VP model, where a sound sample can be inputted as a time signal, being the output the subjective sound evaluation (annoyance). However, as it was seen in the previous section, the 5 CNN based prediction models that predict the intermediate features (psychoacoustic metrics) from the time signals have different performances. Therefore, by doing feature selection it is possible to obtain a better overall performance of the Virtual Passenger model, i.e., it is possible to build the VP model without having to use the 5 intermediate features, being selected the combinations that allows to obtain a better performance.

7.2.1 Feature Selection

Designating Loudness by L, Fluctuation Strength by F, Tonality by T, Sharpness by S and Roughness by R, the effect of using different combinations of this features on the Virtual Passenger model performance was studied. For example, in section 6.2 the feature combination used was $LFTSR$.

In order to perform feature selection, the second block of the VP model has to be re-trained for each feature combination. Hence, it is first necessary to train ANN models very similarly to the one from 6.2 (same hyperparameters and 70% of data for training), for the different feature combinations. The procedure adopted was quite similar to the previously used, being that for each feature combination 100 random data divisions are done, being the performance averaged and, of the 100 created models, the best performing is chosen for including in the VP model.

Regarding the feature combinations used, these were selected based on the correlation between the features and annoyance, already computed and shown on Tab. 6.3. Therefore, starting with all the features ($LFTSR$), these are sequentially removed one-by-one, until obtaining the combination LS. Also, considering tonality was the feature predicted with greater performance and that loudness represents a highly relevant acoustic dimension, also the combination LT, was included in the feature selection study.

For the second block of the VP model (feature-based), randomly dividing the data into training and testing data 100 times, the average and best performances for the different ANNs are presented on Tab. 7.4, being noticeable that the use of different feature combinations does not cause a relevant effect on the performance of the second block of the prediction model, which predicts subjective from objective metrics.

Table 7.4: Feature-based ANN model performance in 100 random data divisions (70% data for training)

(a) Averaged performance

-	R^2	MAE	$RMSE$
$LFTSR$	0.9814	3.9574	5.0933
$LFTS$	0.9817	3.8713	5.1302
LTS	0.9792	3.8447	5.0426
LS	0.9806	3.8112	4.9149
LT	0.9843	3.6194	4.7294

(b) Performance with the best RMSE

-	R^2	MAE	$RMSE$
$LFTSR$	0.9976	1.7220	2.1001
$LFTS$	0.9987	1.7942	2.2318
LTS	0.9970	1.9082	2.3293
LS	0.9988	1.2341	1.5362
LT	0.99741	1.6656	2.0227

Having both blocks of the VP model trained, it is now possible to combine them and assess its performance. The 30 sound samples used for jury testing are utilized for assessing performance, being inputted in the VP model and the annoyance prediction for each stimuli compared with the respective mean juror annoyance. However one should keep in mind that, the second block of the VP model was built using only 70% for training, thus having 9 sound samples that are entirely new to the VP model. Hence, when, in this section, the expression *testing data* is used, it means that the sound samples are entirely new to the VP model. However, about the remaining 21 sound samples, the second block of the model was trained using their LMS Test.Lab computed psychoacoustic metrics, but here they are inputted with metrics predicted with the first block, therefore being also interesting to analyze the prediction performance the overall set of 30 sound samples.

In order to be able to perform a fair comparison between the different feature combinations, the Monte Carlo Method is used one last time. First the 30 sound samples are inputted in the model, being obtained 30 feature predictions. Then, for each feature combination, the predicted features are introduced into 100 trained feature-based blocks (the same ones used for Tab. 7.4). Then the 100 annoyance predictions are compared with the original mean juror annoyances, being the performances averaged, both for the entire 30 jury testing stimuli and for the 9 *unseen* stimuli. The results are shown on Tab. 7.5. Also, Tab. 7.6 contains the best performance obtained for each feature combination, based on the RMSE.

Analyzing the results on Fig. 7.5, both for the 30 whole sound samples and for the 9 unseen samples, the feature combination that only includes loudness and sharpness (LS) is the one with the best performance. Comparing the performance between the entire 30 stimuli and that ones that are *unseen* to the model, for the feature combination LS the performance is slightly worse for the *unseen* data, however, the same does not occur for some of the other feature combinations. This may mean that, due to the fact that the stimuli that were familiar to the second block of the VP model were used with predicted features, it is reasonable to assess performance on the entire set of 30 stimuli.

The feature combination ***LS*** is thus the one chosen for the VP model, due to its superior performance. In Figs. 7.11 and 7.12, for this feature combination, the predictions of the best performing VP model are presented, with an error and correlation analysis, respectively.

Table 7.5: Virtual Passenger average performances over 100 random data divisions for the second block, for each feature combination and considering both the jury testing 30 samples and the 9 used for evaluating the performance of the feature-based models

(a) 30 sound samples used for jury testing

-	R^2	MAE	$RMSE$
$LFTSR$	0.8489	11.035	14.592
$LFTS$	0.83519	10.969	14.881
LTS	0.86077	9.9895	13.758
LS	0.87609	9.6746	13.164
LT	0.85931	9.8297	13.752

(b) 9 *unseen* sound samples

-	R^2	MAE	$RMSE$
$LFTSR$	0.86007	10.652	13.795
$LFTS$	0.84325	11.014	14.424
LTS	0.85547	10.236	13.601
LS	0.88877	10.023	12.829
LT	0.87196	10.059	13.483

Table 7.6: Virtual Passenger model best performance for each feature set, based on the best RMSE over 100 random data divisons on the model second block

-	R^2	MAE	$RMSE$
$LFTSR$	0.887	9.016	11.896
$LFTS$	0.881	9.089	12.405
LTS	0.899	8.755	11.308
LS	0.906	8.610	11.372
LT	0.901	8.267	11.020

As a final note, it should be kept in mind that, as it has been seen, even though it is possible to improve performance by discarding features, each one of the metrics represents a psychoacoustic dimension, that allows the model to be more robust and complete when predicting on new sound samples. Hence, in case one wished to use the VP model on new sound samples, the feature selection step should be done taking into account the *a priori* known characteristics of the sound on which the model will predict on. For example, it would be unwise to discard fluctuation strength in a model for predicting on highly modulated sounds.

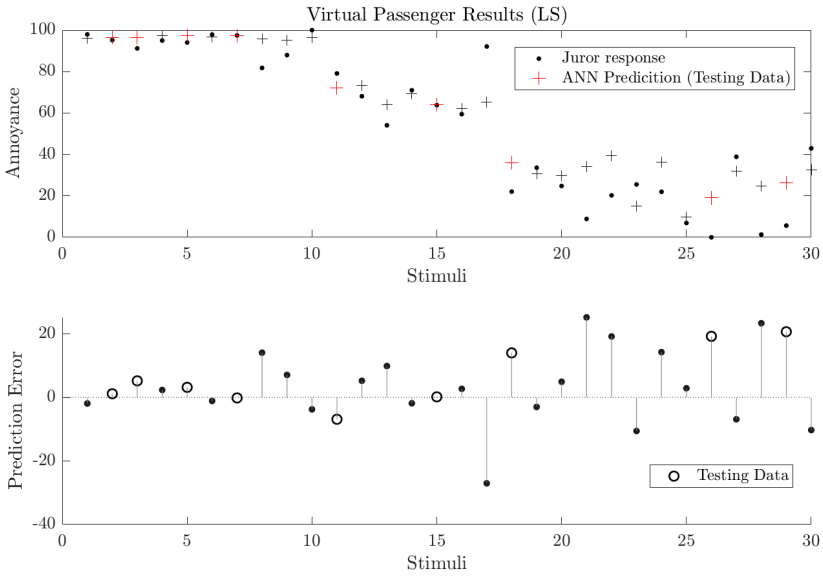

Figure 7.11: Virtual Passenger prediction error as a function of annoyance.

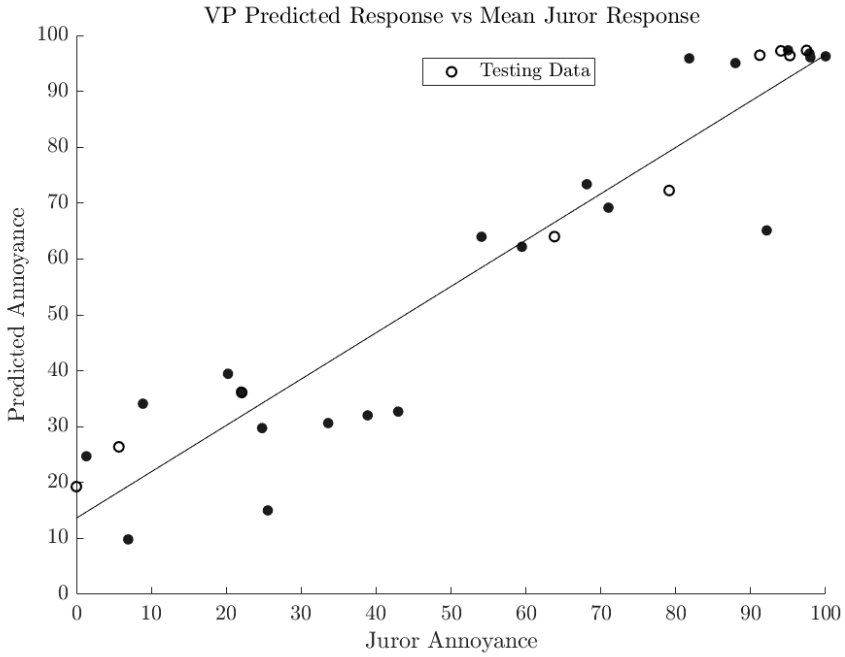

Figure 7.12: Correlation between Virtual Passenger annoyance predictions and mean juror evaluations.

7.2.2 Virtual Passenger Model Prediction Results Analysis

As mentioned on the previous section, the feature combination that provides superior performance is *LS*, being the one in analysis in this section. Reconsidering the categories initially used for jury testing (shown in Fig. 6.2), it is possible to verify if the predictions are within the range of a category. In Fig. 7.13, the annoyance prediction error for each stimulus is shown along the mean juror annoyance range, for *LS*. A line corresponding to the width of a jury testing class (17/100) is drawn. Defining accuracy as the number of times the VP model correctly predicts with an error below the range of a category (i.e., the difference between an annoyance prediction and mean juror response is smaller than 17/100) over the total number of predictions, the VP model, for *LS*, is 80% (only six stimuli have an error superior than 17/100), as it can be seen on Fig. 7.13.

Also, still from Fig. 7.13, it is notable that the prediction error is greater on certain annoyance ranges, namely stimulus with mean juror annoyance inferior than 30/100 and between 80/100 and 90/100, being that this second interval contains the stimuli with the higher annoyance prediction error. Both this cases will be analyzed.

Regarding the low mean juror annoyance stimuli, it is notable that these represent the greater error component in the overall predictions. It is notable that the dispersion of the results when collecting the jury testing evaluations may play a role on this specific type of error. In Fig. 7.14, the standard deviation for each stimuli is included, allowing to relate the error between the VP prediction and the jurors with the results dispersion.

Analyzing 7.14, it is possible to observe a greater standard deviation for the stimuli with mean juror annoyance below 30/1000, i.e., in these stimuli the juror had a bigger dispersion when evaluating the sound samples. In fact, for the 10 stimuli with lower annoyance, the mean standard deviation is 20% greater than the mean standard deviation for all the stimuli. So, the bigger prediction error in these *noisier* range of stimuli shows that to predict on more dispersed results can lead to worst performance.

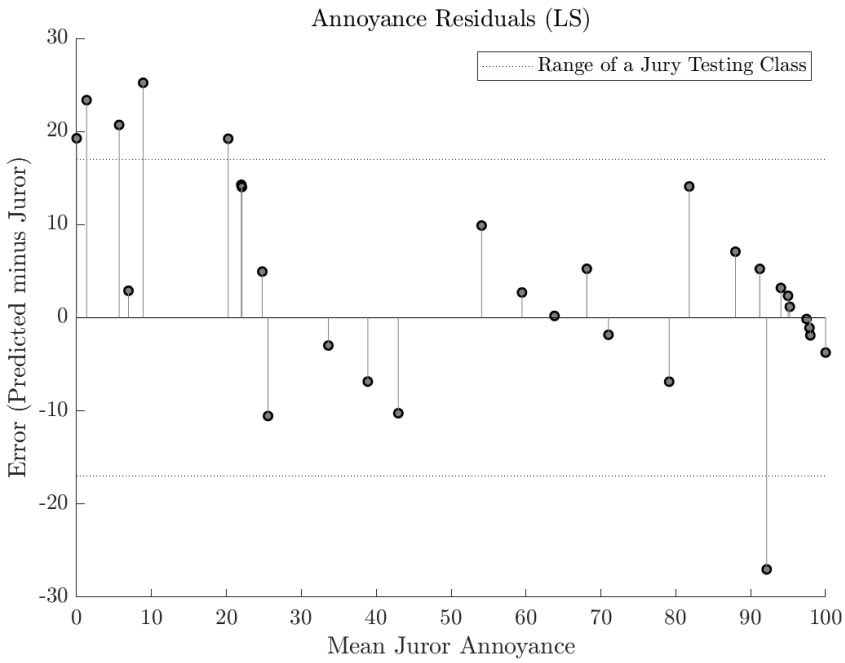

Figure 7.13: VP model predictions compared with the original mean juror annoyance.

About the prediction error in the range between 80/100 and just above 90/100, its possible to observe a connection between these results and the error in the intermediate feature prediction. Remembering that the intermediate features are loudness and sharpness, it is important to analyze the influence of the error in the CNN models on the complete VP model.

Secondly, when analyzing the loudness and sharpness prediction error, it is possible to observe that this error is bigger for intermediate ranges of this metrics, as was previously observed in section 7.1. Starting with loudness, analyzing Fig. 7.15, it is possible to observe two loudness ranges with high annoyance prediction error. The first corresponds to low loudness sounds (below 60 *Sone*), which are the low annoyance stimuli already analyzed in the previous paragraphs, that have a low loudness prediction error and high juror variablity. The other range is the one with loudness between 80 and 100 *Sone*, where both the annoyance and loudness prediction error are above average. Note that in this specific loudness range the mean absolute loudness prediction error is more than the double of the mean absolute error for all stimuli.

Regarding sharpness, the situation is quite similar. For sharpness values below 0.7 *Acum*, there is a high annoyance prediction error and a low sharpness prediction error, being this the high variability stimuli already covered. The interesting range is the one that contains the stimuli with sharpness between 0.9 and 1 *Acum*. Here, both the annoyance prediction error and the sharpness prediction error are both high. This is represented in Fig. 7.16.

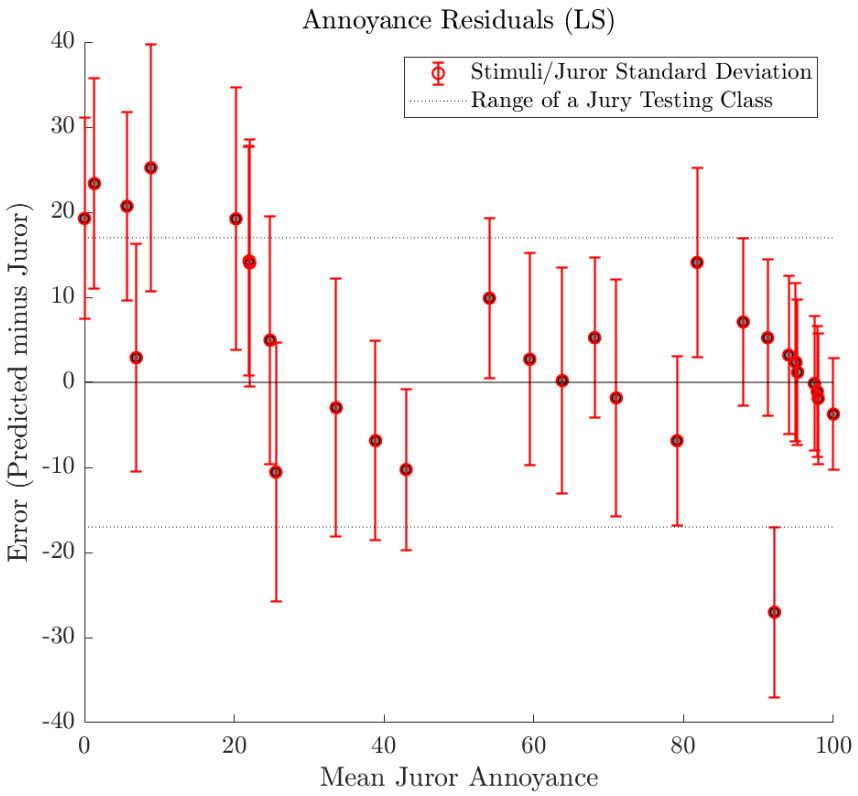

Figure 7.14: VP model predictions compared with the original mean juror annoyance, including the standard deviation of all jurors for each stimuli.

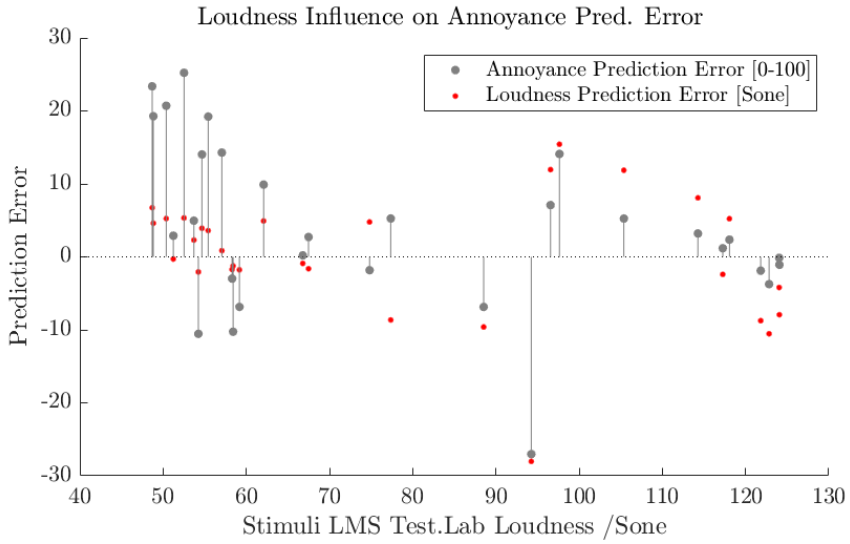

Figure 7.15: Loudness prediction error as a function of stimuli loudness in *Sone*.

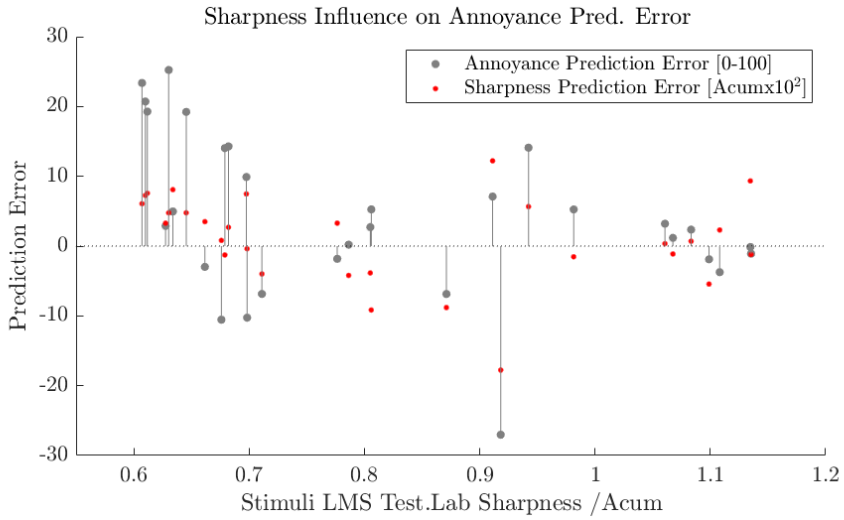

Figure 7.16: Sharpness prediction error as a function of stimuli loudness in *Acum*.

Therefore, it is important to retain that for stimuli with loudness contained in between 80 and 100 *Sone* and sharpness ranging from 0.9 to 1 *Acum*, there a notable increase in both the annoyance and the feature prediction error. Observing Fig. 7.17 it is observable that both this feature ranges correspond to stimuli with mean juror annoyance contained between 80/100 and 90/100. Remembering that on Fig. 7.13 this range was identified as more prone to prediction error (containing even the stimuli with the higher annoyance prediction error), its then possible to relate it with the poor performance in predicting intermediate features.

Taking the last paragraphs into account, when analyzing the VP results for a feature combination of loudness and sharpness (LS), it was possible to link the annoyance prediction error to high juror variability (annoyance values below 30/100) and error in predicting the intermediate features, for stimuli with loudness between 80 and 100 *Sone* and sharpness superior to 0.9 and inferior to 1 *Acum* (corresponding to stimuli with annoyance ranged between 90/100 and 100/100).

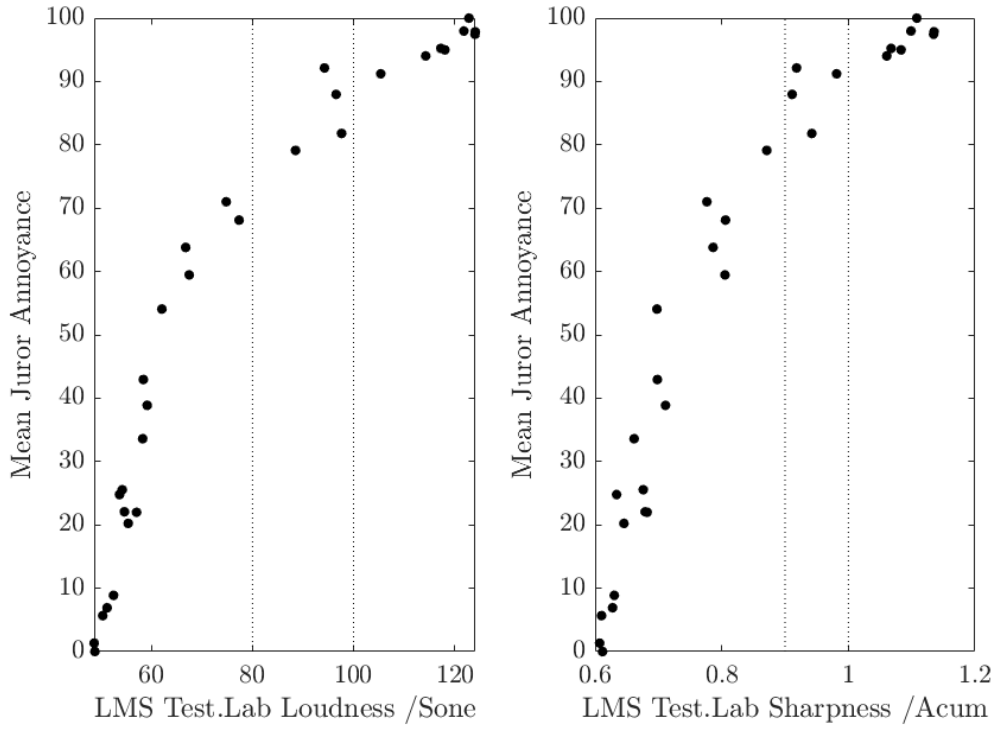

Figure 7.17: Correspondence between stimuli loudness (left) and sharpness (right) with the mean juror annoyance.

Chapter 8

Conclusion and Future Work

8.1 Conclusions

After all the research and work performed throughout this thesis, several conclusions were formed.

When designing new products, it is still required to use jury studies to quantify the human SQ perception, as a subjective metric. Several authors propose the use of feature-based linear (MLR) or non-linear prediction models (e.g. ANNs, SVMs, RFs) in order to, from psychoacoustic features, predict the subjective evaluation of a sound sample, as annoyance. According to the existing literature on the field, the use of ANNs, with a low number of hidden layers is the most widely used approach, allowing to mimic the human perception of sounds and being trained with data collected in jury studies.

From the conducted campaign to collect subjective evaluations of sound samples, it was possible to develop 4 feature-based models (MLR, ANNs, SVMs and RFs). As expected, the performance for the nonlinear models was superior than for the linear one, being verified that the human perception of sounds can be mimicked with the use of non-linear feature-based models. The ANNs stood out among the non-linear models, both for their performance and stability, having been studied the performance using Monte Carlo simulations and taking into account the effects of data division. The training and testing of prediction models should be done with care, always keeping in mind that the data division process has a great influence on the obtained performance. By exploiting the ANN trained prediction model, a spatial distribution of the subjective SQ perception was done, showing that, although it is a subjective metric, annoyance has a strong correlation with physic and psychoacoustic phenomena.

Due to both the advancements in ML techniques and available computing power in the last decade, CNNs are referred to in the literature as a powerful feature extractor, however the extraction of psychoacoustic metrics using CNNs is not a common approach. It was shown that it is possible to train CNN models that, from a time signal are able to predict psychoacoustic metrics, namely: loudness, fluctuation strength, tonality, sharpness and roughness. The prediction performance, was superior for loudness, tonality and sharpness than for fluctuation strength and roughness.

The sequential combination of a psychoacoustic feature extracting model with the feature-based ANN model allowed to develop a VP model, that from a sound sample, inputted as a time signal directly predicts its subjective metric (annoyance). A feature selection study was done, being seen loudness and sharpness are the feature combination that allows to obtain the superior performance for the VP model. The complete model, for many sound samples, is able to predict with an error inferior to the initial discrete

ranges used when conducting the jury testing campaign. The developed full model was also validated on unseen sound samples, showing that using CNNs as a feature extractor, although the prediction error increases, allows to obtain a much more compact and exportable model than with the traditional approaches.

The higher prediction errors were analyzed over the annoyance range, being identified two different cases where most of them are located: low annoyance stimuli and medium-high annoyance stimuli. For the first case, the high prediction error arises from an increase in juror response variability. For the second, medium-high annoyance stimuli are associated with certain ranges of loudness and sharpness. In the second case, high prediction error in certain ranges of loudness and sharpness was identified in the CNN feature extraction block, which is responsible for the high annoyance prediction error in the full model, in the range of medium-high annoyance.

8.2 Future Work

Considering the models developed so far, and also the results and their sub-consequent analysis, several developments and studies can be made to improve their performance and to better test their applicability to other real cases.

First, for both the Virtual Passenger model and the four feature-based prediction models trained (MLR, ANN, SVM and RF), it should be noted they were trained using data from a specific model of a propeller aircraft. It would be of interest to assess their performance when predicting on data from other models of propeller aircrafts or even a jet aircraft. This would allow to test how far can these models be generalized in an aircraft design context.

Still for both types of models, it is now possible to implement them in a design cycle. Thus, the possiblity of including them in a multi-attribute design optimization process would allow to include the human SQ perception factor in the aircraft design.

Focusing on the Virtual Passenger model and taking into account that, in opposition to feature-based models, this consisted in a new approach in predicting SQ with numerous possible ways to improve its performance. As a starting point for this process, three solutions to decrease the prediction error are suggested:

- On the psychoacoustic feature extraction block, the architecture choice and also the hyperparameter tuning are complex processes, often iterative and time consuming. Further improvements can still be done exploring diffferent architectures and hyperparameter sets, using also bayesian optimization as a tool.

- A common approach in the literature about speech recognition and audio processing consists into turning a *machine hearing* problem into a *machine vision* one, i.e., by transforming the time signals correspondent to sound samples into spectrograms, which are inputted into the CNNs as if they were images. Hence, for the spectrogram correspondent to each sound sample, the psychoacoustic metrics would be predicted.

- Recalling that the VP model had a significant prediction error associated with stimuli with low annoyance, due to the high variability in the juror evaluation of this type of stimuli, to re-conduct jury testing on a new set of sound samples similar to these would be a possible way to augment the available data. If the results in this new jury study were to have low variability, one should expect an increase in performance.

- Another important error component of the VP model is associated with the prediction error in its first block, that predicts psychoacoustic features from time signals.

For loudness and sharpness, specific ranges of this metrics were identified as prone to higher prediction error and leading to also high error in the VP model when predicting on medium-high annoyance stimuli. Hence, to synthesize additional stimuli with loudness and sharpness contained in the referred ranges would consist in a data augmentation technique, that should contribute to an improvement of the prediction performance.

It should be noted only monaural sounds were used, being possible that this may play a role in the jury testing results. For a future jury study, in order to obtain more accurate experimental results and for the jurors to have a more realistic experience, the sound samples should be re-recorded and re-synthesized with binaural properties.

It was shown that the annoyance distribution in the interior of the cabin of a propeller aircraft is not homogeneous, existing, for example, areas with higher annoyance near in the engines. When contemplating active sound control solutions, to apply them equally over the cabin, even though suppressing noise, does not assure an increase in passenger comfort. Hence, considering it was developed a compact easily deployable SQ prediction model, a final application of the VP model could be to to include it in a localized active sound control solution, where each seat would be equipped with an active sound control application and where the developed Virtual Passenger model would be able to predict the human reaction to a sound on which the sound control would act.

References

Ahmad, M., M. Mourshed, and Y. Rezgui
2017. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings*, 147:77 – 89.

Air, A.
2018. Dornier 228. `https://alkanair.com/alkanair_fleet/dornier-228-2/`. Accessed: 2018-06-20.

Angeloni, L.
2018. Sound Quality Target Settings and Enhanced Sound Equalization of a Turbopropeller Aircraft Interior Noise.

Aytar, Y., C.Vondrick, and A. Torralba
2016. SoundNet: Learning Sound Representations from Unlabeled Video. *CoRR*, abs/1610.09001.

Bishop, C.
2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, H.: Springer-Verlag.

Breiman, L.
2001. Random Forests. *Machine Learning*, 45(1):5–32.

Brochu, E., V. Cora, and N. Freitas
2010. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. *CoRR*, abs/1012.2599.

CIRA
2018. CASTLE - CAbin Systems design Toward passenger welLbEing. `https://www.cira.it/en/aeronautics/safety-e-security-air-transport/castle/CASTLE%20-%20CAbin%20Systems%20design%20Toward%20passenger%20welLbEing`. Accessed: 2018-06-20.

Dai, W., C. Dai, S. Qu, J. Li, and S. Das
2017. Very deep Convolutional Neural Networks for raw waveforms. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Pp. 421–425.

Ding, Y., X. Song, and Y. Zen
2008. Forecasting financial condition of Chinese listed companies based on Support Vector Machine. *Expert Systems with Applications*, 34(4):3081 – 3089.

d'Ischia, M., A. Concilio, and A. Sorrentino
2001. Civil - Aircraft Passenger Comfort: Guidelines for Analysis and Simulation. *Proceeding of the 4th European Conference on Noise Control, EURONOISE 2001.*

Duvigneau, F., S. Liefold, M. Höchstetter, J. Verhey, and U. Gabbert
2016. Analysis of simulated engine sounds using a psychoacoustic model. *Journal of Sound and Vibration*, 366:544 – 555.

Fastl, H. and E. Zwicker
2007. *Psychoacoustics: Facts and Models*, Springer series in information sciences. Springer Berlin Heidelberg.

Fausett, L., ed.
1994. *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications.* Prentice-Hall, Inc.

Fisher, R.
1925. *Statistical methods for research workers.* Edinburgh Oliver & Boyd.

Hastie, T., R. Tibshirani, and J. H. Friedman
2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer series in statistics. Springer.

Huang, H., X. Huang, R. Li, T. Lim, and W. Ding
2016. Sound quality prediction of vehicle interior noise using Deep Belief Networks. *Applied Acoustics*, 113:149 – 161.

Huang, H., R. Li, M. Yang, T. Lim, and W. Ding
2017. Evaluation of vehicle interior sound quality using a continuous restricted Boltzmann machine-based DBN. *Mechanical Systems and Signal Processing*, 84:245 – 267.

Janssens, K., A. Vecchio, and H. V. der Auweraer
2008. Synthesis and Sound Quality evaluation of exterior and interior aircraft noise. *Aerospace Science and Technology*, 12(1).

Jones, D.
2001. A Taxonomy of Global Optimization Methods Based on Response Surfaces. *Journal of Global Optimization*, 21(4):345–383.

Kahn, M. S.
1998. *Sound quality evaluation of heavy-duty engines in free field conditions.* PhD thesis.

Kayri, M.
2016. Predictive Abilities of Bayesian Regularization and Levenberg–Marquardt Algorithms in Artificial Neural Networks: A Comparative Empirical Study on Social Data. 21:1–11.

Kwak, Y. and L. Ingall
2007. Exploring Monte Carlo Simulation Applications for Project Management. *Risk Management*, 9(1):44–57.

Le Ba, J., J. Ryan Kiros, and G. E. Hinton
2016. Layer Normalization.

LeCun, Y., Y. Bengio, and G. Hinton
2015. Deep Learning. 521:436–44.

Lee, S. and H. Chae
2004. The application of Artificial Neural Networks to the characterization of interior noise booming in passenger cars. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 218(1):33–42.

Liu, H., J. Zhang, P. Guo, F. Bi, H. Yu, and G. Ni
2015. Sound quality prediction for engine-radiated noise. *Mechanical Systems and Signal Processing*, 56-57:277 – 287.

LMS
2016a. Sound Metrics 16A.

LMS
2016b. Sound Quality 16A.

MacKay, D.
1992. *Bayesian Interpolation*, volume 4.

Oliveira, L.
2009. *Active sound quality control: design tools and automotive applications*. PhD thesis.

Oliveira, L., K. Janssens, P. Gajdatsy, H. Van der Auweraer, P. Varoto, P. Sas, and W. Desmet
2009. Active Sound quality control of engine induced cavity noise. 23:476–488.

Otto, N., S. Amman, C. Eaton, and S. Lake
2001. Guidelines for Jury Evaluations of Automotive Sounds. 35.

Pietila, G. and T. Lim
2012. Intelligent systems approaches to product sound quality evaluations – A review. *Applied Acoustics*, 73(10):987 – 1002.

Quehl, J.
2001. *Comfort Studies on Aircraft Interior Sound and Vibration*. PhD thesis.

Rainer, G.
1997. Psychological Methods for Evaluating Sound Quality and Assessing Acoustic Information.

Rawlings, J., S. Pantula, and D. Dickey
1998. *Applied Regression Analysis*, Springer texts in statistics, 2. ed edition. New York, NY [u.a.]: Springer.

Russell, S., P. Norvig, and E. Davis
2010. *Artificial Intelligence: A Modern Approach*, Prentice Hall series in artificial intelligence. Prentice Hall.

Sainath, T., R. Weiss, A. Senior, K. Wilson, and O.Vinyals
2015. Learning the speech front-end with raw waveform CLDNNs. In *INTERSPEECH*.

Shuvaev, S., H. Giaffar, and A. Koulakov
2017. Representations of Sound in Deep Learning of Audio Features from Music.

Sky, C.
2018. `http://www.cleansky.eu/`. Accessed: 2018-06-20.

Snoek, J., H. Larochelle, and R. Adams
2012. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems 25*, Pp. 2951–2959. Curran Associates, Inc.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov
2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Thickstun, J., Z. Harchaoui, and S. Kakade
2017. Learning Features of Music from Scratch. In *International Conference on Learning Representations (ICLR)*.

Trigeorgis, G., F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou
2016. Adieu features? End-to-end speech emotion recognition using a deep Convolutional Recurrent Network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Pp. 5200–5204.

Vapnik, V.
1999. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999.

Vapnik, V.
2013. *The Nature of Statistical Learning Theory*. Springer New York.

Wilby, J.
2008. *Aircraft Cabin Noise and Vibration Prediction and Passive Control*, chapter 99, Pp. 1197–1206. Wiley-Blackwell.

Wu, C., G. Tzeng, Y. Goo, and W. Fang
2007. A real-valued genetic algorithm to optimize the parameters of Support Vector Machine for predicting bankruptcy. *Expert Systems with Applications*, 32(2):397 – 408.

Xue, F., B. Sun, and L. Li
2016. Improvement in the predicition performance of subjective evaluation of sound quality for vehicle HVAC systems by using SVM algorithm. *Inter-Noise Hamburg 2016*.

Zhang, H., L. Wang, J. Yin, P. Chen, and H. Zhang
2017. Performance of the Levenberg–Marquardt Neural Network approach in nuclear mass prediction. *Journal of Physics G: Nuclear and Particle Physics*, 44(4):045110.

Zhang, Y., K. Sohn, R. Villegas, G. Pan, and H. Lee
2015. Improving Object Detection with Deep Convolutional Networks via Bayesian Optimization and Structured Prediction. *CoRR*, abs/1504.03293.

Appendix A

Additional Results for the Feature-based models

In order to facilitate the reading of Ch. 6, some figures that contain results of the study on the feature-based prediction models response, were included in this appendix.

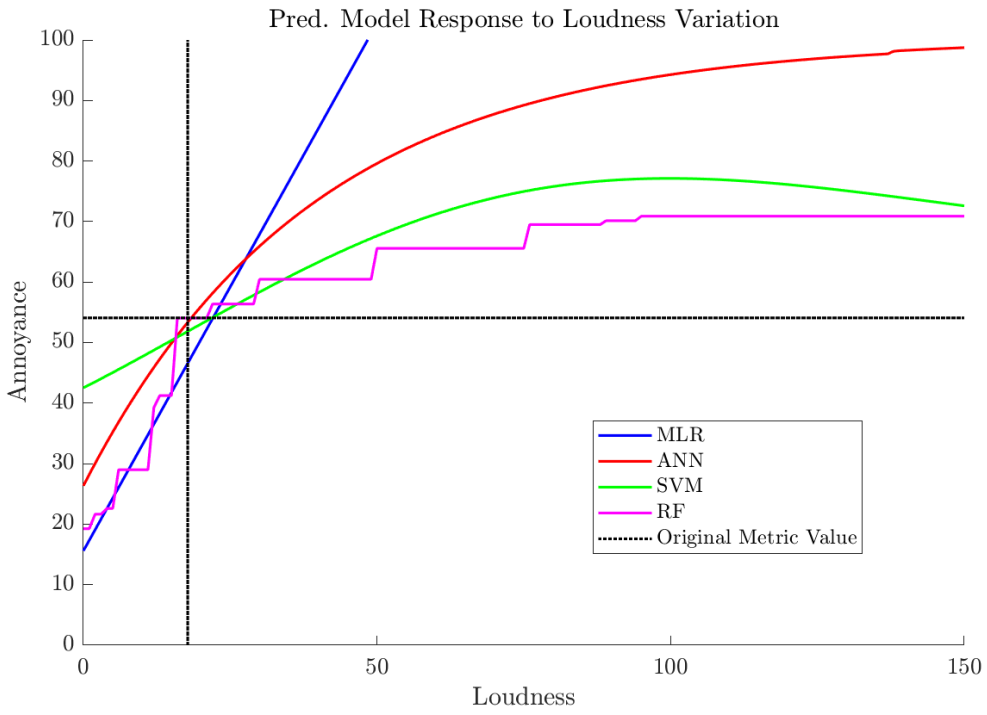

Figure A.1: Influence of a manual loudness variation on annoyance, for stimulus 13.

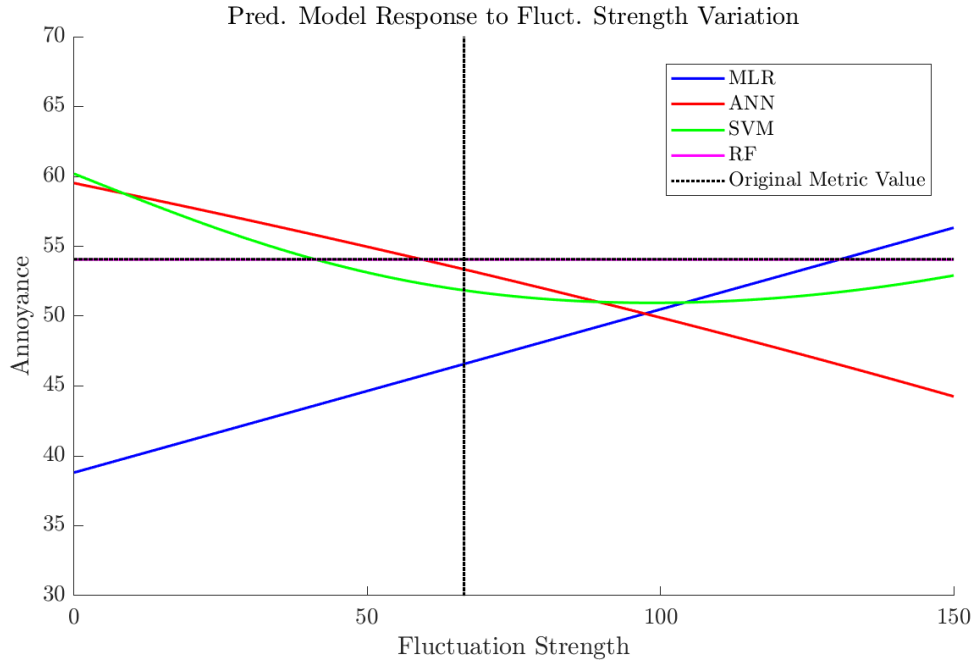

Figure A.2: Influence of a manual fluctuation strenght variation on annoyance, for stimulus 13.

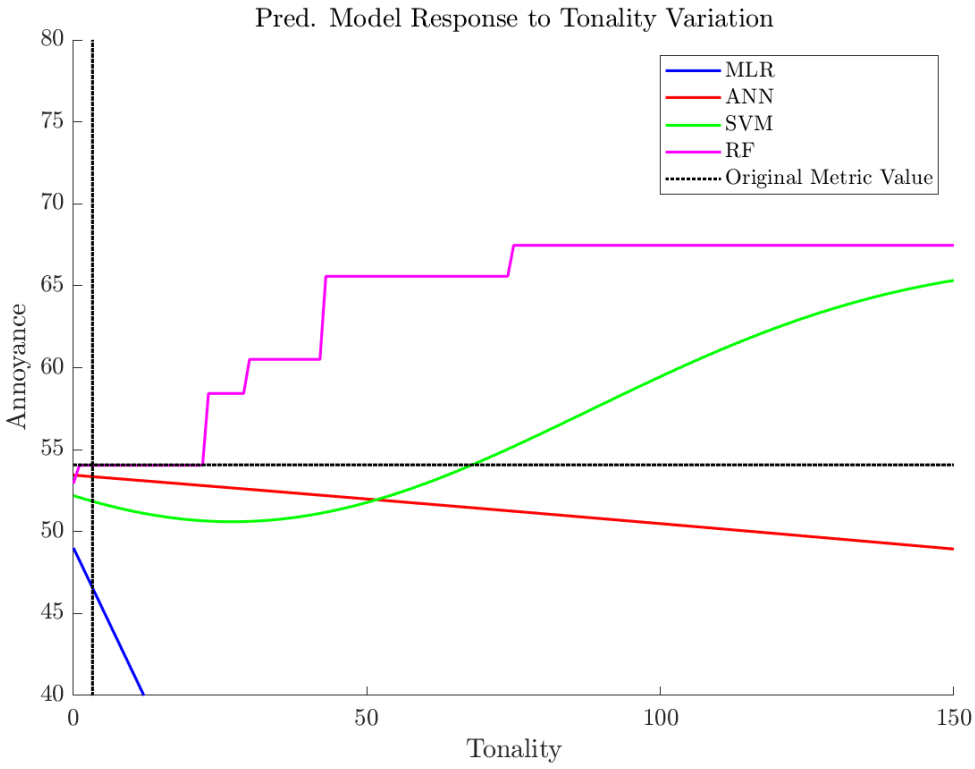

Figure A.3: Influence of a manual tonality variation on annoyance, for stimulus 13.

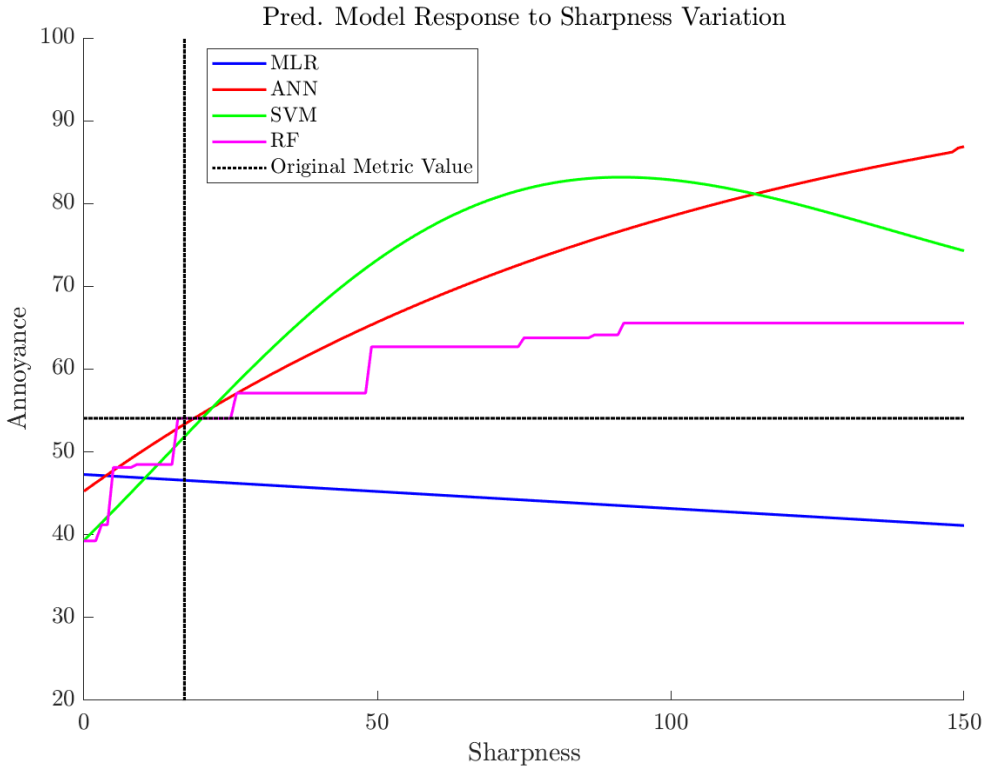

Figure A.4: Influence of a manual sharpness variation on annoyance, for stimulus 13.

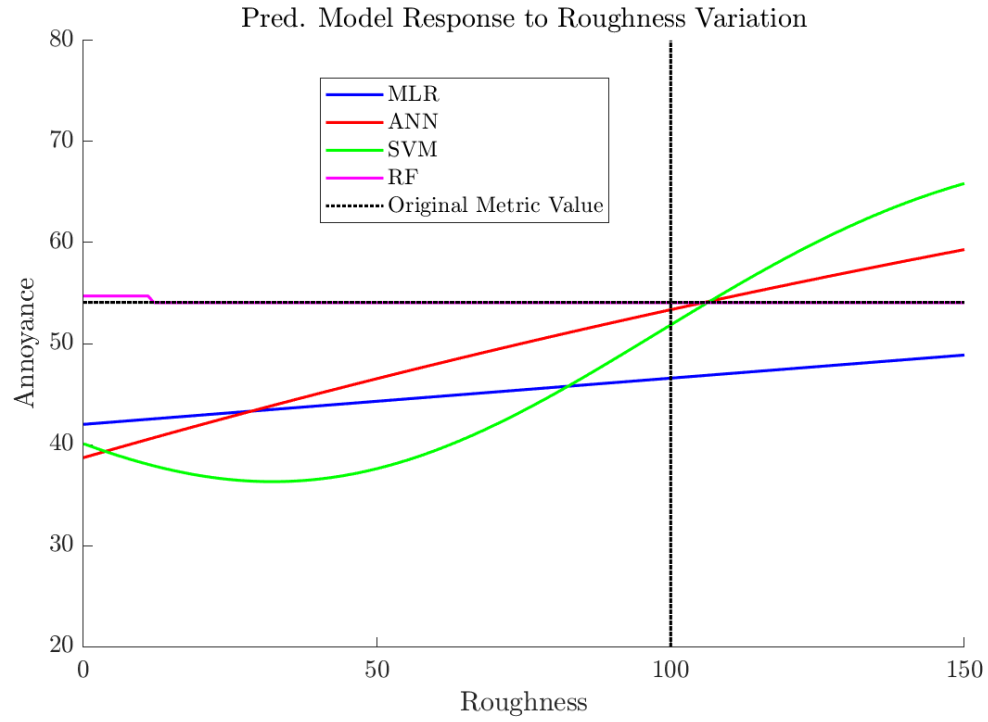

Figure A.5: Influence of a manual roughness variation on annoyance, for stimulus 13.