Politecnico di Torino

Master of Science Course in
Cinema and Media Engineering
A.a. 2023/2024
Sessione di Laurea Aprile 2024

Sound Quality Evaluation of the Interior Noise of a Tractor HVAC System Based on Prediction Model

Supervisors:
Arianna Astolfi
Louena Shtrepi

Candidate:
Marco Favaretto

Index

1. Introduction

In the dynamic world of agricultural machinery, tractors are indispensable tools, working long hours in diverse and challenging environments. The well-being and productivity of the operator is closely linked to the comfort and functionality of the cab. As the demand for ergonomic and technologically advanced tractor cabs continues to grow, there is an urgent need to address the sound quality of these environments. In order to improve overall comfort, tractor cabs typically incorporate measures such as sound insulation to minimise vibration and noise caused by the engine. However, existing techniques primarily address engine-related issues and there is currently a lack of research and established methods to effectively mitigate interior noise generated by HVAC systems in tractor cabs. It should be noted that during the summer months, when temperatures are higher, operators are forced to run the HVAC system at maximum speed, resulting in increased noise levels. The expected increase in HVAC system noise levels on hot days is a critical consideration, given the potential impact on operator experience and comfort. This thesis focuses on a predictive modelling approach to evaluate the sound quality parameters associated with the interior noise of a tractor HVAC system, combining objective measurements with subjective evaluations. The methodology begins with an extensive review of the relevant literature on vehicle interior noise prediction models, focusing on which psychoacoustic parameters are used, what type of subjective evaluation is used, and the best performing prediction models. This foundation serves as a critical framework for the subsequent stages of the study, ensuring a comprehensive understanding of the current state of research in this area. Signal acquisition tests are an essential part of the methodology and involve the use of a 19-channel microphone, an artificial head with two types of binaural microphone headsets and an omnidirectional microphone. The recorded cabin noise, captured at the operator's ear level, serves as the empirical basis for the analysis of the psychoacoustic parameters. Recognising that the sole calculation of sound pressure level (SPL) may not be optimal for assessing noise discomfort, the study goes beyond SPL by meticulously calculating additional psychoacoustic parameters. Parameters such as loudness, A-weighted sound pressure level (A-SPL), sharpness and roughness are included in the analysis to provide a more comprehensive understanding of the perceptual aspects of tractor HVAC interior noise. The evaluation of sound quality is fundamentally based on the subjective perception of the individual. Through meticulous subjective testing, objective parameters are harmoniously integrated with human perceptual experience. This synthesis of empirical data and subjective ratings forms the basis for the construction of predictive models aimed at understanding and quantifying sound quality. Two subjective rating methods are proposed and compared: the first method uses sound judgement with a 1-10 rating scale measuring annoyance, while the second method, called Semantic Differential, uses 7-point scales with paired bipolar adjectives related to loudness, A-SPL, roughness and sharpness. The subjective tests are carried out using both binaural listening with a computer and spatialised audio listening (3rd order ambisonics) with a virtual reality (VR) headset simulating the tractor cab environment. Furthermore, the research endeavors to develop a sound quality prediction model through the application of multiple linear regression. The predictive capabilities of the model are rigorously assessed, with particular attention to its accuracy in forecasting psychoacoustic parameters and subjective noise ratings. Preliminary findings suggest promising outcomes, with the 1-10 rating scale demonstrating exceptional efficacy in predicting noise annoyance, achieving an impressive R-squared value of 0.97. Additionally, the Semantic Differential Method (SDM) showcases its utility in predicting psychoacoustic parameters, particularly excelling in roughness (R-squared of 0.88) and loudness (R-squared of 0.97). However, challenges are encountered in reliably predicting sharpness, attributed to significant errors and discrepancies in subjective ratings. The findings of this study form the basis for the development of advanced predictive models, potentially using neural networks. These models will make a significant contribution to the improvement of tractor cabin, promoting a more comfortable and operator-friendly environment in agricultural machinery.

2. Methods

2.1. Literature search

The aim of this research is to summarise the progress made in predicting noise annoyance in tractor cab HVAC (Heating, Ventilation and Air Conditioning) systems from 2009 to the present. It aims to comprehensively review the prevailing reference methodologies and psychoacoustic parameters used in the construction of prediction models for this specific context. The methodology consisted of several steps, namely literature search, selection of papers, tabulation, and data analysis. The following sections are dedicated to the description of these steps.
Current research is based on studies that have assessed noise and vibration at operator ear level in tractors and vehicles, providing valuable subjective measures for predictive models. However, there is a clear research gap in the specific investigation of HVAC-generated noise in these environments. This highlights the need for targeted research to fully understand the impact of HVAC noise on the operator's working environment.
Specific and ecologically valid acoustic conditions were considered, with a particular focus on results and analyses from studies published between 2009 and 2023. The reason for starting with the period from 2009 is due to the scarcity of literature specifically investigating the noise prediction model of an HVAC system in a vehicle. In particular, the only available articles focusing on this aspect were published in 2009, 2012 and 2021. Other research papers primarily focus on broader evaluations of vehicle interior noise, without a specific focus on the HVAC system. Exploring studies on prediction models related to vehicle interior noise could significantly improve the understanding of tractor HVAC noise prediction models, despite the lack of explicit focus on this specific aspect in the recent literature.
The strategy for finding relevant literature was based on a **PICO** (Population Intervention Comparison and Outcome) strategy and was carried out in the Scopus database. The search insisted on terms related to sound quality prediction models studied for tractor and vehicle interiors and measurement techniques to evaluate the interior noise of an HVAC system in a tractor cabin. Figure 1 shows the clusters of terms selected for the search. These terms appeared in the title, abstract or keyword of the documents. If the full text of the papers was not available in Scopus, further searches were carried out in ResearchGate, Google and Pico, a bibliographic search engine that provides a unique and integrated access to all the bibliographic resources of the Politecnico di Torino.

2.1.1. Selection of papers

Once the papers identified through the initial keyword-based selection process were collected, they were assessed for inclusion or exclusion based on the criteria outlined in **Table 1**. This evaluation followed a two-step procedure. Firstly, titles and abstracts were screened. Secondly, papers that passed the initial screening were subjected to a more comprehensive review of the full text.
Regarding the topic of tractor noise prediction models, the tabulation process included experiments conducted inside the vehicle. This inclusion was made because there is a paucity, if not a complete absence, of studies that specifically focus on predicting sound in tractor or vehicle HVAC systems.

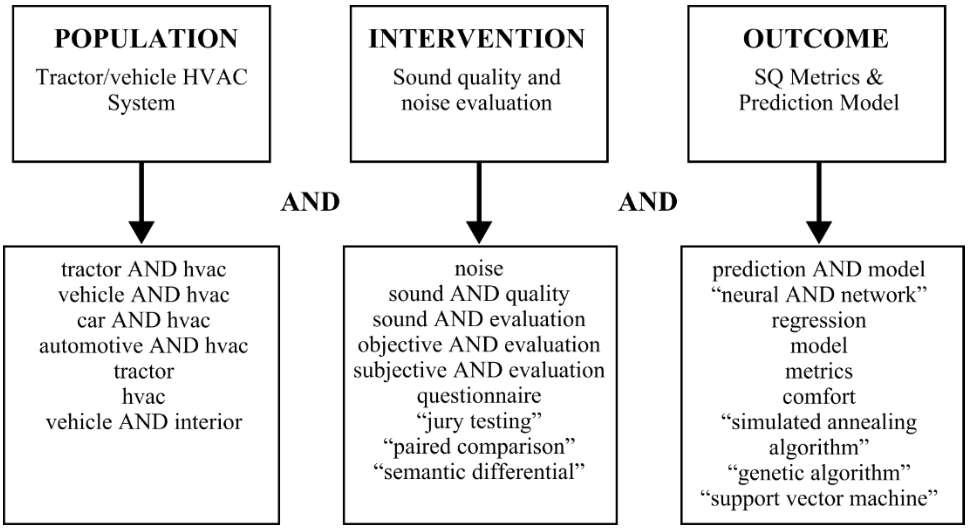

Fig. 1: terms used in the PICO strategy research

Table 1

Inclusion and exclusion criteria for the selection of papers

Inclusion	Exclusion
• Studies related to HVAC's noise prediction models inside tractor's cabin • Studies contain significant correlation ($p < 0.05$) between tractor or vehicle interior sound and the development of a noise prediction model • Experiments carried out in-field or in laboratory with ecological validity	• All papers for which the topic was unrelated to acoustics • Review papers • Proceedings of conferences • Papers not published in English • Papers not dealing with vehicle interior noise • Papers using 3D software simulations for sound prediction • Papers for which a full text is not available • Subjective tests not involving binaural or sound listening in a controlled environment • Subjective test not used

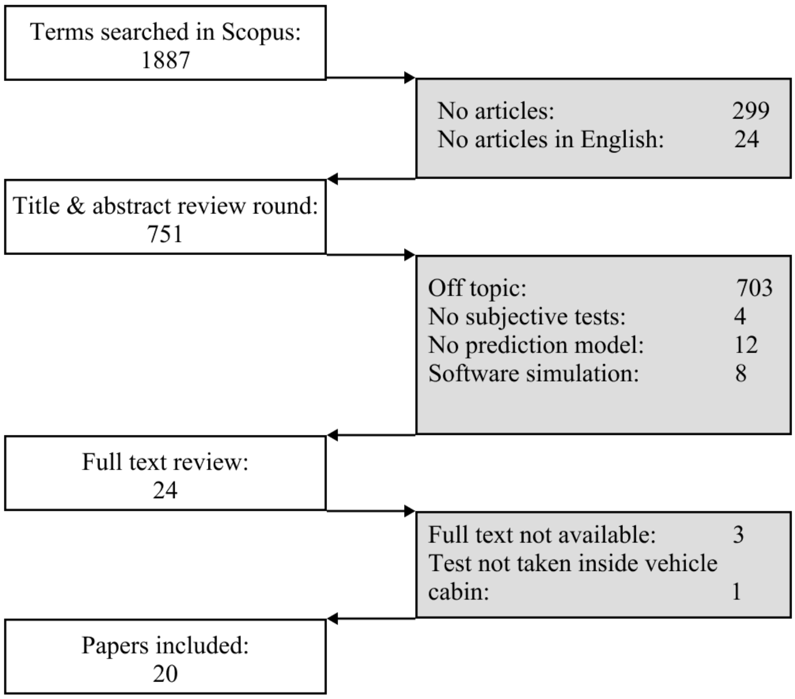

Fig. 2: numbers in the PICO strategy research

2.1.2. Tabulation

Each chosen paper underwent tabulation, capturing the most pertinent information relevant to this research. Specifically, **Table 2** presents the following details:

2.1.2.1. Provided information

- **Vehicle**: specifies the types of vehicles, i.e., tractor (T), fuel vehicle (FV), electric vehicle (EV), micro commercial vehicles (MCV), heavy commercial vehicles (HCV)
- **HVAC**: indicates whether the study specifically analyzed the vehicle or tractor's Heating, Ventilation, and Air Conditioning (HVAC) system concerning sound annoyance
- **Recording**: describes the scenarios considered in the field sound recordings (e.g., Idle (I), Maximum Torque (MT), Rated Power (RP), Different Speeds (DS), Fan Speeds (FS), Acceleration (A), Different Roads (DR), Breaking (B), Semi Anechoic Chamber (SAC), Idle Engine Run-up (IER)
- **Equipment**: indicates the quantity and type of microphones used, i.e., Artificial Head (AH), Vibration Sensors (VS), Microphones (M), Binaural Microphones (M(B)), Microphone Headset (MH)
- **Mic Position**: specifies microphone's position during recording time, i.e., Operator Ear Level (OEL), Different Positions (DP) for multiple positions inside the vehicle.
- **Samples**: the number of recorded samples
- **Acoustical Parameters**: details the specific acoustic features or characteristics analysed in the sound prediction models i.e., Loudness (L), Sharpness (S), Roughness (R), Fluctuation Strength (FS), Articulation Index (AI), Tonality (T), Linear Sound Pressure Level (L-SPL), A-Weighted Sound Pressure Level (A-SPL), Prominence (P), Tone to Noise Ratio (TNR), Hand Vibration (HV), Seat Vibration (SV), Waveform (WF), Mel frequency cepstral coefficients (MFCCs)
- **Subjective Evaluation Method** (SEM): Explain the approach employed for subjective evaluations of sound quality i.e., Pairwise Comparison Method (PCM), Rating Scale Method (RSM), Semantic Differential Method (SDM), Absolute Magnitude Estimation (AME), Semantic Differential with Anchor Stimulus (ASD)
- **Jury**: specifies the number of people used for the subjective tests

- **Prediction Method**: includes methodologies or techniques employed in developing the sound prediction models i.e., Multiple Linear Regression (MLR), Deep Belief Networks (DBN), Back Propagation Neural Network with Simulated Annealing (BPNN(SA)), Bidirectional Long Short-Term Memory with genetic algorithm (GA-BiLSTM), XGBoost algorithm, Back Propagation Neural Network with Genetic Algorithm (GA-BPNN), Convolutional Neural Network (CNN), Adaptable Learning Rate Trees CNNs (ALRT-CNN), Time Frequency Images CNNs (TF-CNN), Simulated Annealing and Genetic Algorithm BPNN (SAGA-BPNN), Laplacian Score Deep Belief Network (LS-DBN), Multiple Linear Regression (MLR), Synthetical Annoyance Evaluation for ANNs (ANN-SAE), Particle Sworm Optimization BPNN (PSO-BPNN), Deep Belief Network (DBN), Mahalanobis Distance (MD), Grey System Theory (GSM), Vehicle Noise Annoyance Neural Network Model (VNA-NNM)
- **Error/Correlation**: shows the error or correlation between subjective values and predicted values
- **Best parameters**: shows the most correlated parameters with the prediction model, the parameters are listed in order of importance

2.1.2.2. Specifications of the provided information on subjective evaluation methods

To evaluate sound quality, researchers use various methods to quantify subjective preferences. These methods include Pairwise Comparison Method (PCM), Rating Scale Method (RSM), Semantic Differential Method (SDM), Absolute Magnitude Estimation (AME) and Semantic Differential with Anchor Stimulus (ASD). PCM compares items in pairs to determine preference or importance. RSM assigns numerical scores based on predetermined criteria. SDM uses a scale of opposing adjectives to measure attitudes, and AME independently estimates stimulus size. ASD, similar to SDM, uses a reference point for more precise scoring.

2.1.2.3. Specifications of the provided information on prediction models

In the development of sound forecasting models, various methods and techniques are used, such as Multiple Linear Regression (MLR), Deep Belief Networks (DBN), Back Propagation Neural Network with Simulated Annealing (BPNN(SA)), Bidirectional Long Short-Term Memory with Genetic Algorithm (GA-BiLSTM), XGBoost algorithm, Back Propagation Neural Network with Genetic Algorithm (GA-BPNN), Convolutional Neural Network (CNN), Adaptable Learning Rate Trees CNN (ALRT-CNN), Time Frequency Images CNN (TF-CNN), Simulated Annealing and Genetic Algorithm BPNN (SAGA-BPNN), Laplacian Score Deep Belief Network (LS-DBN), Synthetic Annoyance Evaluation for ANNs (ANN-SAE), Particle Swarm Optimisation BPNN (PSO-BPNN), Deep Belief Network (DBN), Mahalanobis Distance (MD), Grey System Theory (GSM), and Vehicle Noise Annoyance Neural Network Model (VNA-NNM).

These methods cover a wide range of approaches, including traditional statistical methods such as MLR, machine learning techniques such as neural networks (NNs) and their variants, and hybrid models that integrate optimisation algorithms and neural networks.

Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It assumes a linear relationship between the input features and the output and seeks to find the best-fitting linear equation to make predictions.

On the other hand, **neural networks**, including various architectures such as BPNN, DBN, CNN and LSTM, are computational models inspired by the structure and function of the human brain. They consist of interconnected nodes, or neurons, organised in layers, with each neuron processing input data and transmitting signals to other neurons. Neural networks are capable of learning complex patterns and relationships within data, providing non-linear modelling capabilities, and are particularly effective at handling high-dimensional data and capturing intricate dependencies between input and output variables.

Table 2
Summary of the contents collected from the papers (REF) included in the review.

REF	Vehicle	HVAC	Recording	Equipment	Mic Position	Samples	Acoustical Parameters	SEM	Jury	PM	Error/Correlation	Best Parameters
[1]	T	No	I, MT, RP	AH, VS	OEL	30	A-SPL, L, S, R, AI, AI, HV, SV	PCM	40	BPNN(SA)	PE=4.4%	L, R, S, A-SPL, AI, HV, SV
[2]	FV	No	DS	4M (B)	OEL	37	WF, MFCCs	RSM	25	GA-BiLSTM	R=0.9829 MAPE=4.2793%	-
[3]	EV, FV	No	DS	6M (B)	OEL	12(35)	A-SPL, L, R, FS, S, T, AI	RSM	15	XGBoost	MAE=0.1 R^2=0.9805	L, S, A-SPL, R, AI, T, FS
[4]	T	No	DS	1M	DP	42	A-SPL, L, S, R, FS	RSM	20	GA-BPNN	PE=6.75% R^2=0.95	-
[5]	EV	No	A, B, DS	2M (B)	OEL	240	L, S, R, FS, AI, T	RSM	20	ALRT-CNNs	RMSE=0.766	S, R, L, AI, FS, T
[6]	FV	No	DS, DR	1M	OEL	480	No	RSM	15	TF-CNN	Accuracy=97.03%	-
[7]	EV	No	DS	AH	OEL	16	A-SPL, L, S, R, FS, AI, T, I	RSM	36	SAGA-BPNN	Max APE=5% R=0.99	S, L, A-SPL, R, AI, FS, I, T
[8]	EV	No	DS, DR	AH	OEL	1200	A-SPL, C-SPL, L, R, FS, S, T, AI	RSMc	24	LS-DBN	MAPE=2.67% R=0.99	S, L, AI, A-SPL, C-SPL, R
[9]	EV	No	DS	2M (B)	OEL	24	A-SPL, L, R, FS, S, T, AI	SDM	20	BPNN	MAPE=9%	A-SPL, S, R, L, FS, AI, [T
[10]	MCV	No	DS, DR	AH	OEL	74	A-SPL, L, R, S, AI	AME	30	MLR	R^2=0.98	L, S
[11]	FV	No	DS	2M (B)	OEL, DP	180	L-SPL, L, S, R	ASD	25	ANN-SAE	PE=10%	L, R, S, L-SPL
[12]	FV	No	SAC, DS	AH	OEL	30	A-SPL, L, R, FS, S, T, AI	RSM	32	PSO-BPNN	RE=4.17%	-
[13]	FV	No	DS	1M	OEL, DP	216	L-SPL, A-SPL, L, S, R, FS, AI, T, WTEt, WTEn	ASD	24	DBN	RMSE=0.064 R=0.94	-
[14]	HCV	No	DS	AH	OEL	76	A-SPL, L, S, R, FS, T, AI	RSM	40	BP-ANN	PE=6%	L, R, S, AI, A-SPL, T, FS
[15]	FV	No	DS	AH	OEL	12	A-SPL, L, S, R, FS	""	""	MD	PE=4.17%	-
[16]	FV	No	DS	AH	OEL	16	L, S, R, FS	PCM	30	GSM	PE=10%	-
[17]	FV	Yes	IER	NB	OEL	72	L, S, R, FS	SDM	31	BPNN	R=0.98	L, R, S
[18]	FV	No	DS	2M (B)	OEL, DP	42	L, S, R, A	ARSM	25	VNA-NNM	PE=7%	-
[19]	FV	Yes	FS	MH (B)	OEL	240	L, S, TNR, P	SDM	27	MLR	R=0.95	L, S, TNR, P
[20]	FV	Yes	FS	AH	OEL	7	A-SPL, L, R, S	RSM	30	MLR	R^2=0.92	L, S

2.1.2.4. Specifications of the provided information on error evaluation

RMSE (Root Mean Square Error) represents the average of squared differences between predicted and actual values. MAE (Mean Absolute Error) calculates the average of absolute differences between predicted and actual values. MAPE (Mean Absolute Percentage Error) measures the average percentage difference between predicted and actual values, relative to actual values. (PE) Percentage Error denotes the error as a percentage of the actual value. The Pearson Correlation Coefficient quantifies the strength and direction of the linear relationship between two variables. The Coefficient of Determination (R-squared) signifies the proportion of predictable variance in a dependent variable based on independent variable(s). Relative Error quantifies error in relation to the magnitude of the actual values.

2.1.3. Data analysis

There hasn't been a specific article that thoroughly investigates the sound of an HVAC system inside a tractor cabin. Existing research comprises only three articles that examine HVAC sound in fuel vehicles and two papers analysing sound quality within a tractor's cabin (HVAC not considered). Most articles, however, have focused on investigating sound quality inside on fuel and electric vehicles.

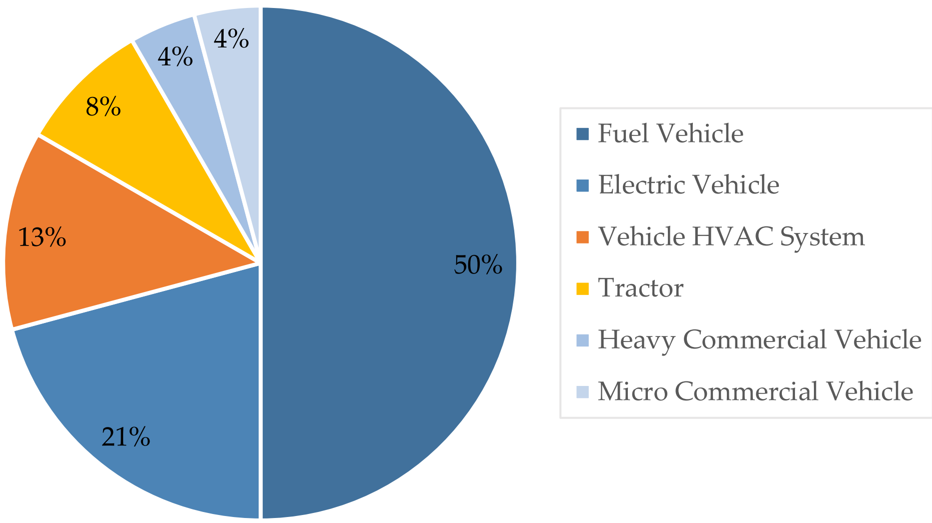

Fig. 3: percentage occurrence of vehicle types in the literature search.

Measurements for electric vehicles and fuel vehicles involved driving the vehicles at various speeds and on different roads. HVAC systems were tested inside the vehicle cabin at different fan speeds. As for tractors, an idle engine run-up was conducted for testing purposes, but the HVAC noise wasn't tested.
Most sound recordings were captured using binaural microphones or artificial heads to simulate the binaural masking effect at operator's ear level inside the vehicle. The average **number of noise samples** measured is 152. The most used psychoacoustic parameters for objective measurements are Loudness and Sharpness, each present in 85% of the occurrences. Following these, Roughness was used in 80% of the papers, while A-weighted Sound Pressure Level was utilized in 60%. Other parameters such as Fluctuation Strength, Articulation Index, Tonality, among others, were also employed, as detailed in Fig. 4 for further information.

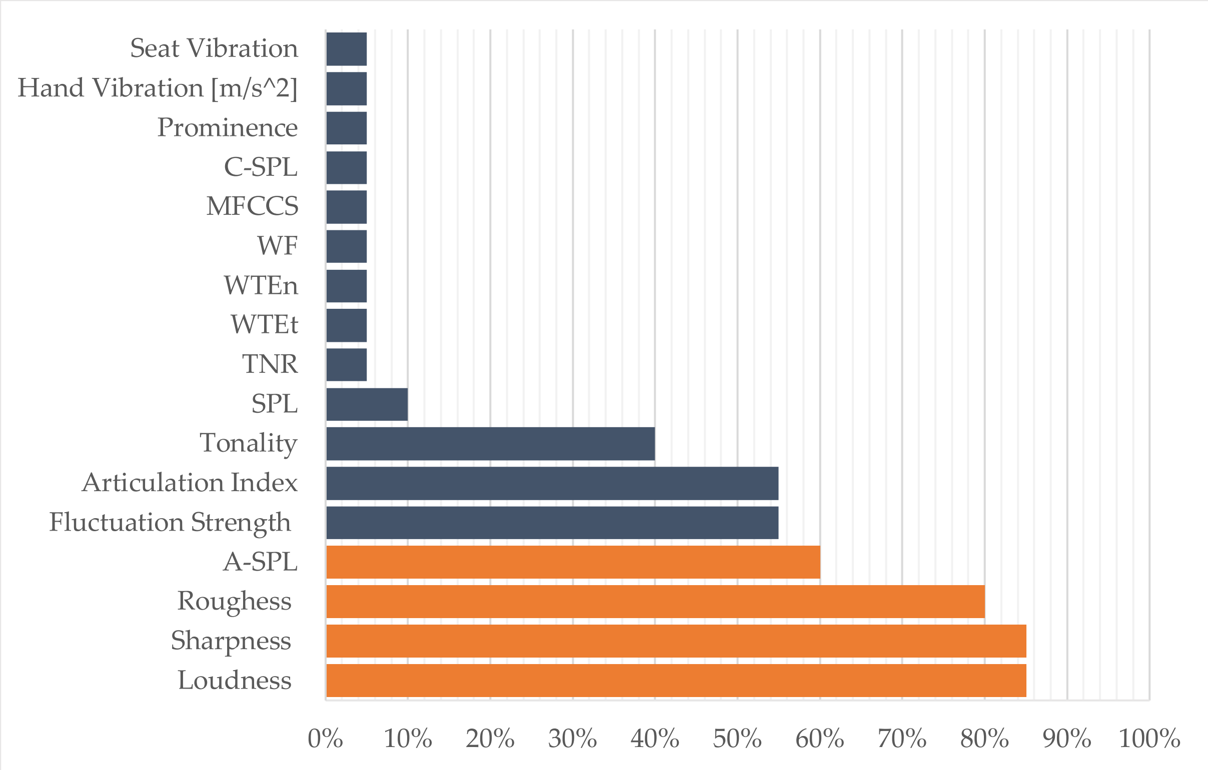

Fig. 4: Percentage occurrence of the psychoacoustic parameters used in the research: i.e, Loudness (L), Sharpness (S), Roughness (R), Fluctuation Strength (FS), Articulation Index (AI), Tonality (T), Sound Pressure Level (L-SPL), A-weighted Sound Pressure Level (A-SPL), Prominence (P), Tone to Noise Ratio (TNR), Hand Vibration (HV), Seat Vibration (SV), Waveform (WF), Mel Frequency Cepstral Coefficients (MFCCs), Wavelet Transform Energy (WTEn), Wavelet Transform Entropy (WTEt);

In relation to subjective tests, an average of 25 **jury members** were selected. There are several subjective evaluation methods available. Among the chosen papers, the Rating Scale Method (RSM) was the most used, representing 50% of occurrences, followed by the Pairwise Comparison Method at 15%, alongside the Semantic Differential with Anchor Stimulus (ASD), Semantic Differential Method (SDM) occurred in 10% of papers, while Absolute Magnitude Estimation (AME) was used in only 5% of the papers, and the rest 5% of papers the evaluation method is not specified (NS) as **Fig. 5** shows:

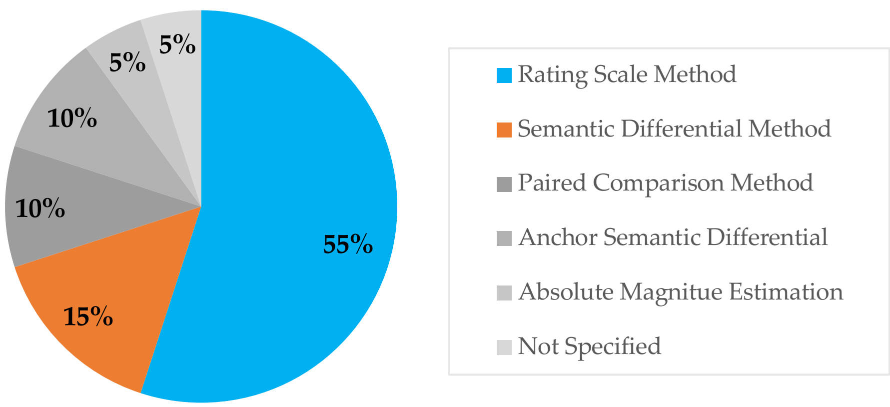

Fig. 5: occurrence, in percentage, of subjective evaluation methods used in the research, i.e., Paired Comparison Method (PCM), Rating Scale Method (RSM), Semantic Differential Method (SDM), Absolute Magnitude Estimation (AME), Semantic Differential with Anchor Stimulus (ASD).

The choice of type of prediction model has a significant impact on the accuracy of predictions. The data analysis shows that 70% of the papers use Neural Networks as their prediction model, followed by Multiple Linear Regression at 15%. In addition, other methods such as Mahalanobis Distance, Grey System Theory and the XGBoost algorithm are used in the remaining percentage of studies.

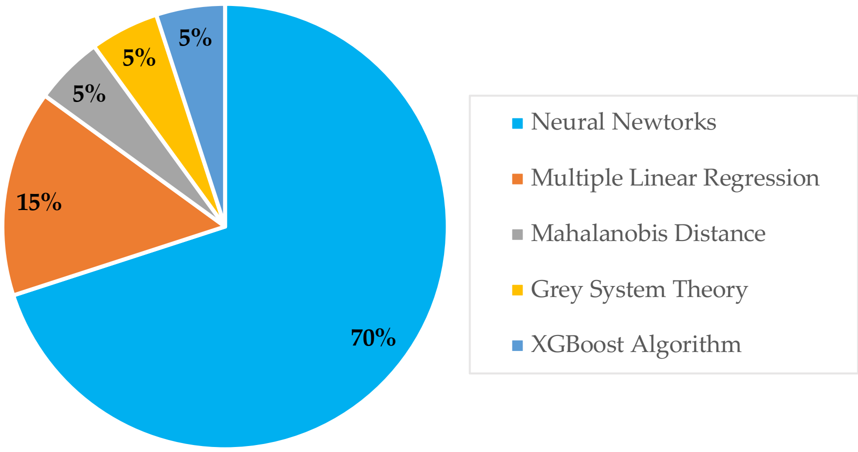

Fig. 6: occurrence, in percentage, of the prediction models used in various research.

Several papers have highlighted the most **influential parameters** affecting prediction outcomes. The findings consistently indicate that Loudness, Sharpness, and Roughness are highly correlated parameters overall. Specifically focusing on tractor sound, prioritizing the context of high noise and vibrations within tractor cabins, the most effective parameters for noise prediction are identified as Loudness, Roughness, and Sharpness. Notably, these three parameters also demonstrate efficacy in studies concerning HVAC systems for fuel vehicles, as demonstrated in **Table 4**.

Table 3: top 3 psychoacoustic parameters in order of prediction influence for every type of vehicle studied.

Type of Vehicle	Top 3				
Tractor	L, R, S				
Electric Vehicle	L, S, A-SPL	S, R, L	S, L, A-SPL	S, L, AI	A-SPL, S, R
Fuel Vehicle	L, R, S				
Heavy Commercial Vehicle	L, R, S				
Fuel Vehicle with HVAC	L, R, S	L, S, TNR	L, S		
Micro Commercial Vehicle	L, S				

2.1.4. Research on Noise Annoyance Assessment

In addition to the objective measurement of psychoacoustic parameters, a socio-acoustic survey is planned for exposed persons working in the tractor cab. The aim is to identify a standardised approach for tractors, alternatively, if a standardised method is not readily available, the search aims to identify a widely used method that provides reliable scaling results and is easily comparable.

As a first step, the subjective evaluation methods used in **Table 2** are analysed, in particular for studies focusing on tractors, heavy commercial vehicles (HCVs) and vehicle HVAC studies.

- Published in 2023 [1] investigates the sound quality within a tractor's cabin using the Paired Comparison Method (**PCM**). In the subjective test, jury members are tasked with comparing randomly paired tractor sounds. This comparison is conducted on a comfort level scale ranging from 1 to 10, where each numerical value corresponds to an anchor category spanning from "very bad" to "excellent" in terms of comfort levels.
- Published in 2022, [4] uses the **RSM**, to investigate the sound quality within a tractor's cabin. Reviewers assess the samples on a unipolar scale of 1 to 10, accompanied by a set of 5 verbal anchors ranging from "little uncomfortable" to "extremely uncomfortable" to gauge the level of sound discomfort.
- Published in 2016, [14] focuses on heavy commercial vehicles and uses “according to the international standards” a unipolar scale from 1 to 10 RSM with 10 verbal anchors for each number that go from “Very Bad” to “Excellent” in terms of sound quality.
- Published in 2012, [17] studies the sound quality of an HVAC system inside a car, the subjective responses were measured by the Semantic Differential Method (**SDM**). Five pairs of adjectives that can represent HVAC noise were specified on a seven-point scale: quiet-loud, soft-sharp, smooth-rough, pleasant-unpleasant, expensive-cheap.
- Published in 2009, the aim of the paper [19] is to investigate the subjective impression of HVAC noise, the semantic differential technique (**SDM**) was chosen. There were no existing studies on auditory descriptors or semantic differentials for sound evaluation in Brazilian Portuguese. Initially, an open questionnaire was introduced to gather insights on the characteristics and importance of automobile ventilation and air conditioning system noise. The subjective impressions were categorized into five aspects: quality (not annoying-annoying), roughness (smooth-rough), loudness (quiet-loud), spectral composition (dull-sharp), and tonality (not-whistling-whistling). The rating scale consists of 7 degrees.
- Published in 2021, paper [20] investigates on a vehicle’s HVAC system, for the jury test, the score range followed a eleven-point scale **RSM** with the following numerical and text anchors: 0–2 (imperceptible pleasantness or coolness), 2–4 (perceptible pleasantness or cool- ness), 4–6 (weak pleasantness or coolness), 6–8 (pleasantness or coolness similar to a normal car), 8–10 (more pleasant or cooler than a normal car), and 10 (very pleasant or cool). The rating was given as a continuous value by moving a slide bar on a listening test program.

After this analysis, research on international standards about noise annoyance was performed:

- **ISO 15666:2021** [21] uses the 0-10 Rating Scale Method as standard, outlines specifications for socio-acoustic surveys and social surveys that encompass inquiries about the impacts of noise. It encompasses questions to be posed, response scales, essential aspects of survey implementation, and guidelines for reporting results. However, it is important to note that the document's focus is limited to surveys concerning noise annoyance specifically "at home."
- The Society of Automotive Engineers **(SAE)** recommends a 1-10 category rating scale for noise and discomfort. [Subjective Rating Scale for Vehicle Ride and Handling, 2016, SAE] [22]
- Guidelines for jury evaluations 1999 [23]

These three documents could be a reference for survey’s structure, but additional research is required to find out if RSM is the optimal method for evaluating vehicle interior noise annoyance. The literature search performed on SCOPUS focuses on articles in English, and uses the PICO strategy for the right selection of the parameters:

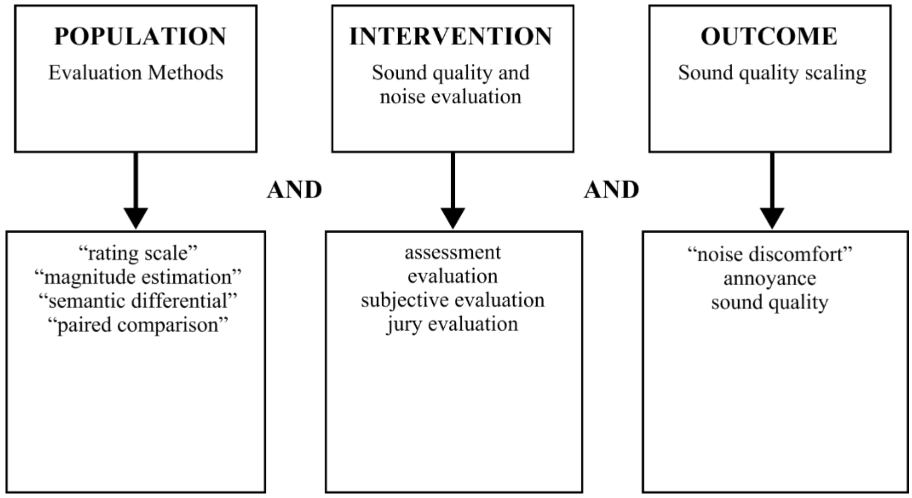

Fig 7: parameters used in the PICO research

Papers underwent revision and were included or excluded following these criteria:

Table 4: inclusion and exclusion criteria for the selection of papers

Inclusion	Exclusion
• Papers related to subjective evaluation methods' assessment • Guidelines for vehicle interior noise assessment	• Papers unrelated to subjective evaluation • Proceedings of conferences • Papers not published in English • Papers already included in the references

After conducting a review of titles and abstracts, 3 documents have been chosen for further analysis out of the initial pool of 155 documents:

- [29] (2020) investigates the subjective discomfort caused by noise in the methods of relative magnitude estimation (RME), absolute magnitude estimation (AME), and Rating Scale Method (RMS), using real sound recorded with an artificial head inside a car driving at constant speed, and an artificial sound with random noise, both at the same SPL of 55dBA. Results indicated that the **RSM** yields **linear interval scales,** and the **ME** method yields **logarithmic interval scales** for the discomfort of noise, so the sensory continua of noise discomfort could be described in linear relation or power function, depending on the evaluating methods. The 'discomfort magnitude of noise' is a quantitative continuum in the ratio scale and the 'discomfort extent of noise' is a qualitative continuum in the category scale.
- [30] (2023) investigates the effect of rating scales, individual characteristics, and number of subjects on annoyance, and may facilitate application in laboratory sound quality assessment experiments. The test results revealed that the **age and the familiarity of the target noise** can lead to **differences** in the perceived annoyance of the subjects. The findings showed that the correlation coefficient between the Mean Noise Annoyance (MNA) of 30 subjects and the MNA of 88 (or 104) individuals was greater than 0.98, and the maximum gap between MNAs is no more than 15 points on the 100-point linear scale.
- [26] (1999) proposes guidelines, positive and negative aspects for every type of jury test focusing on vehicle noise assessment, RSM, SDM, AME, PCM are mentioned.

2.2. Signal Acquisition

On 20th July 2023, noise measurements were conducted on a high horsepower Fendt tractor inside DENSO's shed. The primary objective of the test was to evaluate the noise at the operator's ear level inside the tractor cab. The evaluation took into account various operating scenarios involving both the HVAC system and the engine. The main sources of HVAC noise are the six air vents located on both the left and right sides beneath the steering wheel, along with the air filter situated on the left side below the operator's seat.



Fig 8-9: position of the air vents inside the tractor’s cab
Fig 10: air filter is positioned at the left side of operator’s seat.

Feedback received within the tractor cabin indicates a pronounced sound reflection originating from the rear glass of the cabin. Further investigation and mitigation measures may be necessary to address and alleviate this specific source of noise disturbance.

Fig 11: sound reflection coming from the back glass panel.

During all test scenarios, the tractor remained stationary. The evaluation covered three operational modes of the HVAC system: low speed (speed 4), medium speed (speed 7), and high speed (speed 10). Each speed setting was tested with both the engine on and off. Furthermore, the assessment included a test of speed 7 with the engine on, where the compressor was turned off.

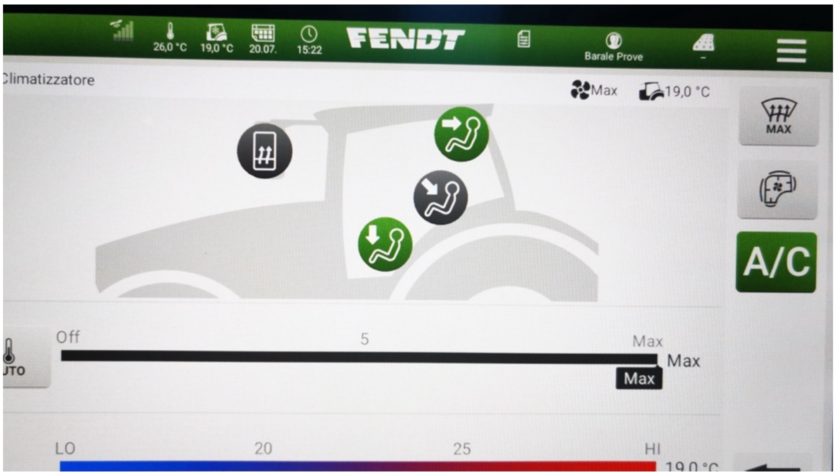

Fig 12: interface of the operational modes within the cabin, green buttons on

It is important to note that the air flow was evaluated with both green buttons activated for all measurements. In future analyses, it will be necessary to comprehensively evaluate various types of air flow, including different air vent inclinations.

Table 5: list of all the measurement type for each operating mode and their abbreviations

Measurement	**Abbreviation**
Background Noise	BG
Engine Off, HVAC Speed 4	S4
Engine Off, HVAC Speed 7	S7
Engine Off, HVAC Speed 10	S10
Engine On, HVAC Off	E S0
Engine On, HVAC Speed 4	E S4
Engine On, HVAC Speed 7	E S7
Engine On, HVAC Speed 7, Compressor Off	E S7 Coff
Engine On, HVAC Speed 10	E S10

2.2.1 Microphones

The microphones used in the assessment were all placed at a consistent height of 1.13 meters ± 0.05 meters. Sample recordings were taken for 60 seconds, and a 10-second portion free of unwanted noises was selected for psychoacoustic calculations. The microphones used included the Zylia ZM-1, NTi M2230, B&K 4101 headset, and Siemens ABH04 Headset. Standardized microphone placement is crucial to ensure accurate and comparable results across evaluations. In acoustic assessments, achieving precise measurements is essential and requires paying meticulous attention to equipment specifications. Each microphone recording underwent calibration utilizing a B&K type 4231 microphone calibrator, which produces a 1 kHz sine wave at 94 dB SPL. However, an exception was made for the Zylia ZM-1 microphone array, whose volume was manually adjusted within the listening environment of the Audio Space Lab, housed within the 16-speaker sphere, to align with the volume of NTi M2230 omnidirectional recordings.

2.2.1.1. Zylia ZM-1

- Positioning: operator ear level
- Frequency range: 20Hz - 20kHz
- Type of recording: ambisonics audio

The ZYLIA ZM-1 is a specialised microphone array designed for high-fidelity multi-track audio recording. It features 19 omnidirectional capsules that use state-of-the-art MEMS (Micro-Electro-

Mechanical Systems) technology. These capsules are strategically positioned to capture sound sources with precision in specific directions and distances.
Samples were recorded at the operator's ear level and used for subjective assessments in the Audio Space Lab at Politecnico di Torino. The recorded tracks were played in 3D order ambisonics using a sphere of 16 audio speakers built around the subject. These results will then be compared to the objective results obtained through psychoacoustic analysis of the NTi microphone recorded samples.

Fig 13: Zylia ZM-1 in recording position.

Fig 14: control software for ZM-1.

ZM-1 uses MEMS-based omnidirectional condenser capsules, taking advantage of silicon technology advancements for a compact design, stable parameters, and low power consumption. The tight tolerances guarantee consistent sound reproduction across microphones, maximizing the effectiveness of ZYLIA's DSP algorithms.

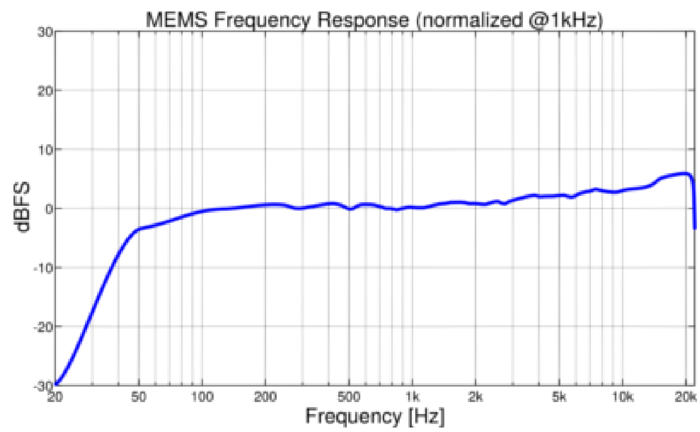

Fig 15: ZM-1 frequency response of a single capsule.

2.2.1.2. NTi M2230

The NTiM2230 microphone was used for omnidirectional noise recording at operators' ear level inside the cabin. It is powered by 48V phantom power and contains a preamplifier in its body. It offers a high dynamic range and wide frequency range while maintaining low noise levels. This measurement microphone can be connected to the XL2 Audio and Acoustic Analyzer via the ASD Cable, XL2 automatically the microphone model and calibration data.

- Sensitivity typical @1kHz: 42mV/Pa
- Measured Quantity: pressure
- Frequency range: 5Hz - 20kHz
- Positioning: operator ear level

Fig 16: NTi microphone recording position.

Fig 17: NTi XL2 acoustic analyzer.

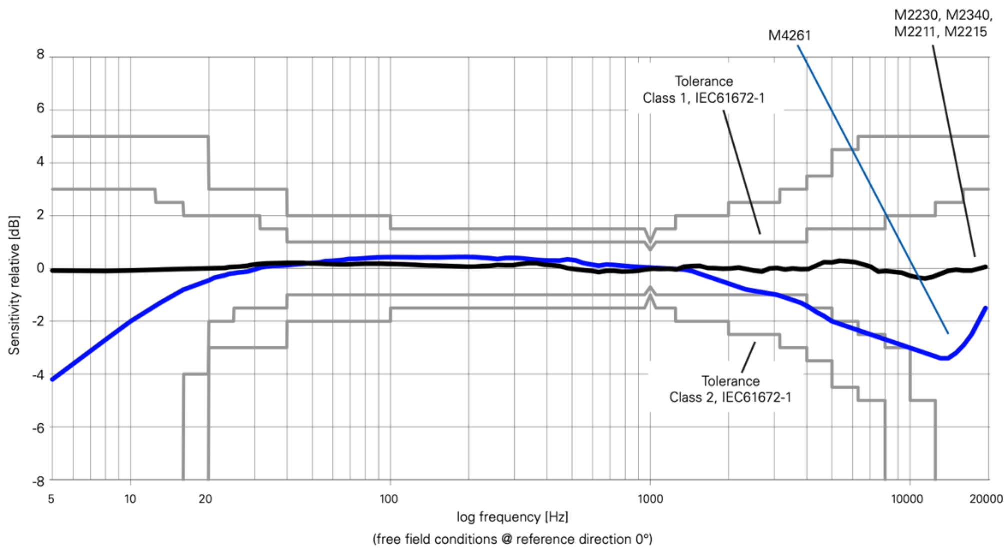

Fig 18: NTi M2230 frequency response (black line) in free field conditions.

2.2.1.3. B&K 4101

Binaural recordings were made using a B&K **Head and Torso Simulator** type 4128, which was equipped with realistic ear and mouth simulators. This apparatus accurately replicates the acoustic properties of an average adult human head and torso, ensuring a lifelike reproduction during the evaluation of two types of headphones. Both headsets were connected to the **Simcenter SCADAS XS**, a handheld data acquisition system capable of simultaneously acquiring dynamic data at 50,000 samples per second on up to 12 dynamic channels. The system has a built-in battery for autonomous operation, or data can be acquired with a host PC.

- Sensitivity typical @1kHz: 20mV/Pa
- Measured Quantity: pressure
- Frequency range: 20Hz - 20kHz
- Positioning: operator ear canal

The B&K 4101 microphones are specifically designed for binaural recordings taken near the entrance of the human ear canal. Binaural recording is particularly effective in capturing sound as it is perceived by the human test subject, resulting in a 3D stereo sensation. This is the microphone frequency response when equipped with a head and torso simulator 4128 with incidence directly from the front:

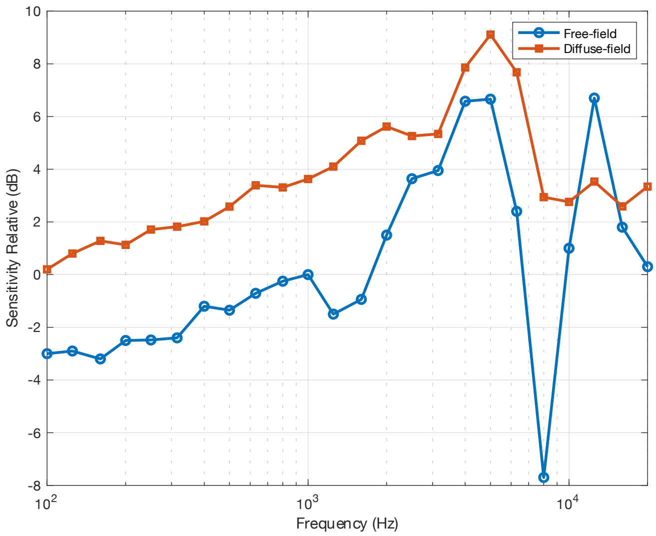

Fig 19 (left): frequency response of B&K 4101 microphone on B&K 4128 head and torso simulator.
Fig 20 (right): B&K 4101 headset mounted on B&K 4128 head and torso simulator.

2.2.1.4. Siemens ABH04

The Siemens ABH04 headset features microphones positioned at both headphone pavilion for binaural recordings. The acquired data can also be listened to through the headset providing a calibrated sound chain for accurate sound playback. The calibrated playback is only available with the headset connected to the scadas XS and a tablet.

- Sensitivity typical @1kHz: 31.7 mV/Pa
- Measured Quantity: pressure
- Frequency range: 20Hz - 20kHz
- Positioning: operator ear external

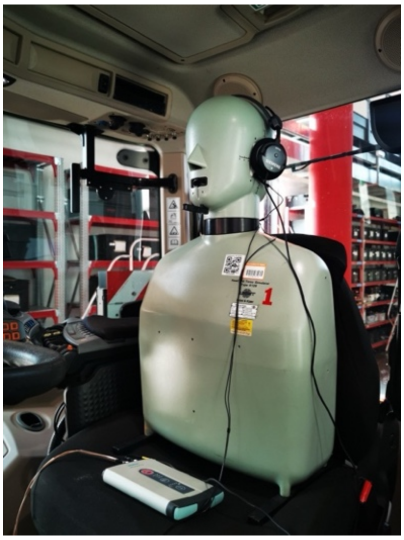

Fig 21-22: Siemens ABH04 headset and Scadas XS.

2.2.2. Psychoacoustic Parameters

Psychoacoustic indicators are metrics used to understand how people perceive sound and how the human auditory system processes it. They go beyond the physical characteristics of sound and encompass the subjective human experience of a particular auditory stimulus. These indicators are derived from a combination of different acoustic factors, offering insights into our perception of sound. This section provides an overview of psychoacoustic parameters, including their definitions, just noticeable differences, and applications in the context of tractors. It also discusses the differences between mono and binaural listening.

2.2.2.1. Linear Equivalent Sound Pressure Level

The LZeq (Equivalent Continuous Sound Pressure Level) characterizes the consistent sound pressure level that, within a specified timeframe, encompasses the equivalent total energy of the fluctuating noise. Therefore, LZeq essentially represents the Root Mean Square (RMS) sound level, wherein the duration of measurement functions as the averaging period. This definition highlights LZeq's role as a standardized metric for gauging the overall intensity of sound experienced over a specific duration, aiding in the assessment of noise exposure levels. The 'Z' letter has replaced Linear or Flat and is defined as being a flat frequency response of 8Hz to 20kHz ±1.5dB.

2.2.2.2. A-weighted Equivalent Sound Pressure Level

The LAeq (A-weighted Equivalent Continuous Sound Pressure Level) is the equivalent continuous sound pressure level with a standard weighting of audible frequencies ('A' weighting) that reflects human ear's response to noise. Measurements made with this frequency weighting will be displayed as dB(A) or dBA.

2.2.2.3. Loudness

Fastl & Zwicker [24] define loudness as the perceived intensity of a sound, influenced by its acoustic qualities and the listening environment, as judged by normally hearing individuals. Loudness is determined by the sound's pressure level but is also affected by frequency, waveform, bandwidth, and duration of the sound. A sone is the loudness of a sound with a loudness level of 40 phon. Doubling the number of sones indicates a sound that's perceived as twice as loud as another.
For this study Loudness was calculated using the Zwicker's method, ISO 532:1975, method B, using a diffuse field equalization, this choice was made because, according to ISO 532-1:2017, Annex D, page 52, a free-field equalization should only be applied if there is just one acoustic source in the free field in front of the head and torso simulator (elevation angle and azimuth angle 0°, distance greater than 1,5 m); the sound inside the tractor's cabin comes from various sources, such as the engine, the various fans of the HVAC system, and the reflections from the side of the cabin.

Loudness according to ISO 532B, can be calculated as follows:

$$N' = 0.08 \left(\frac{E_{TQ}}{E_0}\right)^{0.23} \left[\left(0.5 + 0.5\frac{E}{E_{TQ}}\right)^{0.23} - 1\right] \quad (1)$$

where E is the sound excitation; E_0 is the excitation under benchmark sound intensity; E_{TQ} is the excitation under the absolute listening valve. The specific loudness N' (1) exhibits the distribution of loudness across the critical bands, its unit is "sone/Bark", The total loudness N (2) is the result of the specific loudness values N' through integration of the critical band rate:

$$N = \int_0^{24Bark} N'(z)dz \text{ (sone)} \quad (2)$$

Pedrielli & Cadretti [25] note that the difference in loudness that is just noticeable becomes more pronounced as the overall sound pressure level of the signal increases. In the case of earth-moving machines, where the maximum SPL hovers around 80 dB, the discernible difference in loudness is evaluated to be 0.8 sone. Within the confines of a tractor cabin where the maximum measured SPL is also approximately 80 dB, it is reasonable to consider 0.8 sone as the threshold for the just noticeable difference. For measurements using a head and torso simulator the individual loudness of both the right and the left channel should be reported. The maximum or the average of both channels is regarded as a representative single-digit value.

2.2.2.4. Sharpness

Sharpness is a psychoacoustics metric that provides a numerical measure of sensation based on the number of high-frequency components in a sound. Its unit is the "acum", 1 acum corresponding to the sharpness of a broad-band noise centered on 1 kHz, with a width of 1 critical band and a level of 60 dB. It is a linear scale: double the frequency content, double the sharpness. According to Fastl & Zwicker [24], sharpness increases for a level increment from 30 to 90 dB by a factor of two. This means that the dependence on level can be ignored as a first approximation, especially if the level differences are not very large. The most important parameters influencing sharpness are the overall spectral content and the center frequency of narrow band sounds and is not dependent on loudness level or the detailed spectral content of the sound. It is computed as a weighted sum of the specific loudness levels reported to the global loudness. Several weighting functions exist (by Aures, Fastl, Von Bismarck) but they all increase toward the highest frequency bands to model the fact that the hearing system is more sensitive to high-frequency components level.

Sharpness calculation as described in DIN45692 is based on a prior loudness calculation according to Zwicker method:

$$g(z) = \begin{cases} 1, & z \leq 15.8\,Bark \\ 0.85 + 0.15e^{0.42(z-15.8)}, & z > 15.8\,Bark \end{cases} \quad (3)$$

$$S = 0.11\,\frac{\int_0^{24\,Bark} N\prime g(z) z\,dz}{\int_0^{24\,Bark} N\prime\,dz}\ (acum) \quad (4)$$

where S (2) is the sharpness to be calculated and the denominator gives the total loudness N (2); the upper integral is like the first moment of specific loudness over critical-band rate, but uses an additional factor, $g(z)$ (3), that is critical-band-rate dependent and increases above 16 Bark. The just noticeable difference, defined as the minimum variation in sharpness detected at least by 75% of the jury subjects is assessed as 0.04 acum [25].

2.2.2.5. Fluctuation Strength

The hearing sensation of fluctuation strength is produced at low modulation frequencies up to a modulation frequency about 20 Hz, at higher modulation frequencies the hearing sensation of roughness is produced. A fixed point is therefore defined for a 60-dB, 1-kHz tone 100% amplitude-modulated at 4Hz, as producing 1 vacil.

Fluctuation strength can be described with the following expression:

$$F = \frac{0.008 \int_0^{24\,\mathrm{Bark}} \Delta L_E(\mathrm{z}) dz}{(f_{\mathrm{mod}}/4\mathrm{Hz})+(4\mathrm{Hz}/f_{\mathrm{mod}})}\ (vacil) \quad (5)$$

$$\Delta L_E(\mathrm{z}) = 20 log_{10}\left[\frac{N'_{max}(z)}{N'_{min}(z)}\right] \quad (6)$$

where f_{mod} is the modulation frequency, and $\Delta L_E(\mathrm{z})$ is the variation of the sound signal.

2.2.2.6. Roughness

Roughness is a sensation created by the relatively quick changes produced by modulation frequencies in the region between about 15 to 300 Hz, there is no need for exact periodical modulation, but the spectrum of the modulating function must be between 15 and 300 Hz in order to produce roughness. To define the roughness of 1 asper, it's chosen a 60-dB, 1-kHz tone that is 100% modulated in amplitude at a modulation frequency of 70 Hz.

For amplitude modulation, the important parameters are the degree of modulation and modulation frequency, while for frequency modulation it is the frequency modulation index and modulation frequency.

$$R = 0.3 f_{mod} \int_0^{24\,Bark} 20log \left[\frac{N'_{max}(z)}{N'_{min}(z)}\right] dz \; (asper) \tag{7}$$

The studies of Fastl and Zwicker have revealed that in the case of amplitude-modulated pure tones an increment of roughness becomes perceptible if the degree of modulation is increased by 10%, the corresponding increment in roughness is 17%.

2.2.2.7. Articulation Index

The Articulation Index (AI) serves as a measure indicating how background noise levels can impact human speech comprehension, ranging from 0% (no speech understood) to 100% (complete speech clarity). Initially developed for assessing speech privacy and communication system effectiveness, the AI metric has expanded its applications. Today, it's utilized to evaluate factors like vehicle interior noise, the sound levels of household appliances, and other areas where speech intelligibility is crucial.

$$AI = \frac{M_1}{M_{total}} \times 100\ \% \tag{8}$$

Where M_1 is the number of listened-to sound units and M_{total} is the number of total sound units.

2.2.2.8. Speech Interference Level

Leo Beranek (1947) introduced the Speech Interference Level (SIL) to assess the impact of noise on speech communication within passenger aircraft. SIL is calculated as the average of sound pressure levels measured across octave bands (500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz) and is expressed in dB. This metric provides a single-number rating and serves as a practical way to evaluate how noise interferes with speech communication, both indoors and outdoors.

2.3 Subjective Assessment

The inclusion of subjective evaluation experiments in this study is important as they can directly capture consumers' perceptions, this feedback is necessary for establishing the prediction model. The subjective assessment process involves the evaluator listening to a sound through a replay device, such as headphones or speakers, and providing an evaluation based on various approaches. The assessment results can be slightly affected by the choice of playback system and evaluation method. To ensure accuracy, we will compare two playback systems and two subjective evaluation methods. Each playback system will be tested on 10 individuals, for a total of 20 subjects between 24 and 60 years old, they had no hearing impairment. The recorded noise samples from **Table 5** were cut to a length of 10 seconds, and the same portion was used to calculate psychoacoustic parameters.

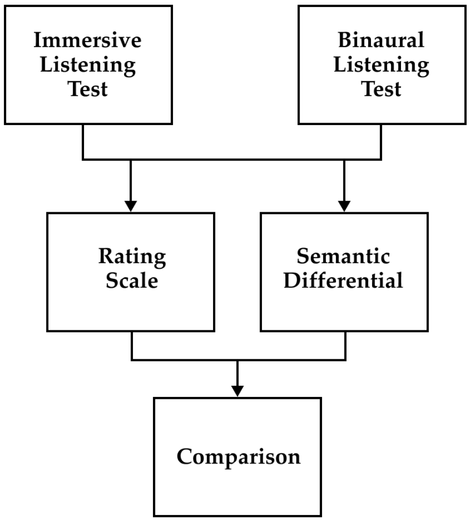

Fig 23: subjective assessment pipeline.

Ten participants underwent an **immersive listening test**, which provided an exclusive simulation of the tractor cab through the use of a VR headset and a sphere with 16 speakers reproducing 3rd order ambisonics audio. Ten participants underwent a **binaural listening test** using a pair of Sennheiser HD650 headphones, which played the binaural recordings made with both B&K 4101 and Siemens ABH04 binaural microphone headsets. Playback and evaluation of the recordings is possible using Head Acoustics' ArtemiS Suite software. The literature review shows that there are several methods available for evaluation. In order to be consistent with the research findings, two widely used and recommended methods were selected: a 1-10 category rating scale focusing on sound annoyance, as used in Chen's research [4], and a semantic differential method, as used by Leite [19]. This methodological choice improves the robustness of the study by aligning it with established practices in the field.

2.3.1. Rating Methods

To ensure comparability, the following subjective rating methods are applied consistently to all types of playback systems. Prior to the formal test on each playback system, the tester is presented with three different sounds selected from the HVAC operating modes, one at low intensity, one at medium intensity and one at high intensity. The test sounds are rated using the 1-10 Annoyance Rating Scale and the four scales of the Semantic Differential Method. For the semantic scales Dull/Sharp and Smooth/Rough a reference with an example of this type of sound is proposed. The purpose of the training session is to familiarise the tester with the scoring procedure as recommended in paper [23] and to improve the accuracy of the scores during the official test.

2.3.1.1. Rating Scale Method

The 1-10 category rating scale involves assessing the subjective sensations experienced by the human body on a scale of 1 to 10, ranging from comfort to discomfort, as shown in **Table 6** The scale consists of 5 categories, ranging from 'Little Annoyance' to 'Extreme Annoyance,' and includes ten distinct levels. It is important to note that the scale will be presented in Italian, in accordance with the nationality of the participants. **Table 7** provides the corresponding Italian translation. The comfort level for the sample is determined by the average of all evaluators' scores, with lower ratings indicating lower comfort levels.

Table 6: 1-10 Annoyance Rating Scale.

Verbal Descriptors	**Little Annoyance**	**Moderate Annoyance**	**High Annoyance**	**Very High Annoyance**	**Extreme Annoyance**
Category	1-2	3-4	5-6	7-8	9-10

Table 7: 1-10 Annoyance Rating Scale translated into Italian.

Verbal Descriptors	**Poco Fastidioso**	**Leggermente Fastidioso**	**Moderatamente Fastidioso**	**Molto Fastidioso**	**Estremamente Fastidioso**
Category	1-2	3-4	5-6	7-8	9-10

Fig 24: Annoyance rating scale displayed in VR headset and translated in Italian.

Fig 25: 1-1 Annoyance Rating Scale displayed on SQala software.

2.3.1.2. Semantic Differential Method

The Semantic Differential Method (SDM) allows for the evaluation of different sound attributes and will provide a more comprehensive evaluation of psychoacoustic parameters such as sharpness, roughness, and loudness. Participants use descriptive response scales with bipolar adjective pairs to rate sounds. These pairs consist of an adjective and its opposite representing the following psychoacoustic parameters and sensations: loudness (quiet/loud), noise annoyance (not annoying/annoying), sharpness (dull-sharp), roughness (smooth-rough). The sounds must be rated for each pair of adjectives on a 7-point scale. For the evaluation test, the semantic pairs were translated into Italian using the translation words provided by D. Dal Palù and E. Buatti work on semantic differential scales [27]. In the study, bilingual subjects, proficient in both English and Italian, translated the adjectives using different words. They then collaboratively selected the adjective that most accurately captured the intended translation.

Table 8: SDM scales and their corresponding translations to Italian.

SDM Scale	Not-Annoying/Annoying	Quiet/Loud	Dull/Sharp	Smooth/Rough
Italian Translation	Non-Fastidioso/Fastidioso	Debole/Forte	Sordo/Penetrante	Regolare/Irregolare

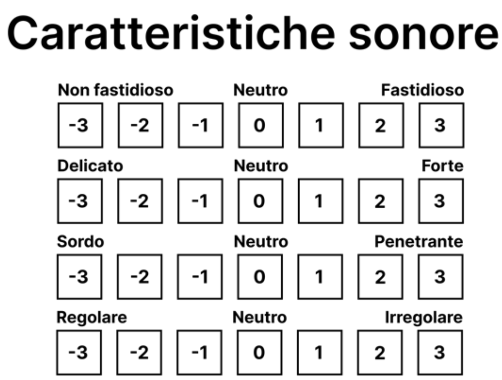

Fig 26: Semantic Differential Method (SDM) scales displayed on VR headset, Italian translations.

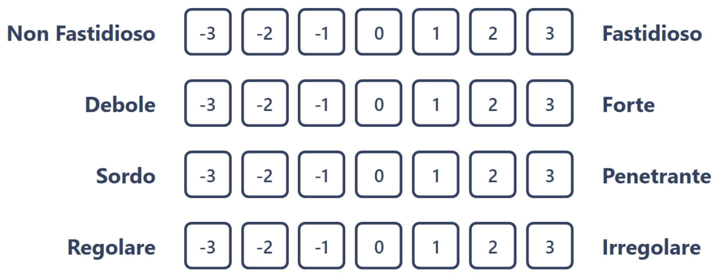

Fig 27: Semantic Differential Method (SDM) scales displayed on SQala software, Italian translations.

2.3.2 Sound Playback Systems

2.3.2.1. Immersive Test

Ten participants were selected to assess different HVAC working modes based on **Table 5**. They immersed themselves in spatialized sound recordings and experienced a lifelike representation of a tractor cab through a Meta Quest VR Headset. This approach, which leverages the omnidirectionality of the playback system, promises a more realistic experience. It allows for an accurate representation of the reflections within the tractor cabin and a more authentic portrayal of the low-frequency sounds produced by the tractor engine.

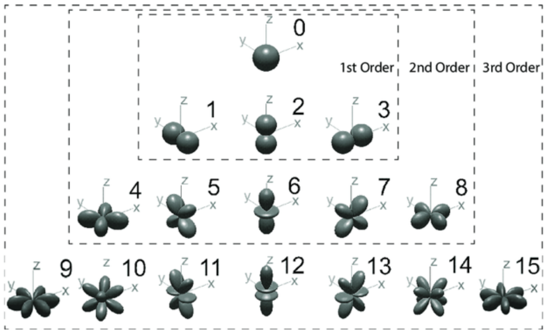

Fig 28-29: an immersive test in Audio Space Lab (left); orders of ambisonics (right).

Ambisonics is a technique for capturing, reproducing, and manipulating sound in a three-dimensional space. It is commonly used in virtual reality, augmented reality, and immersive audio applications.
The order of Ambisonics refers to the complexity and precision of the spatial information captured or reproduced, higher-order Ambisonics is crucial in scenarios where precise sound localization and a realistic audio environment are critical. The three common orders of Ambisonics are 1st, 2nd, and 3rd.
1st Order Ambisonics:

- Captures sound information in a spherical manner using four channels: W (omnidirectional), X (front-back), Y (left-right), and Z (up-down).
- It is suitable for basic spherical sound reproduction but lacks detailed spatial resolution.
- It is often used in basic applications where a simple surround sound field is sufficient.

2nd Order Ambisonics

- Adds additional channels to enhance spatial resolution. The system employs nine channels, namely W, X, Y, Z, U, V, T, S, and R.
- It provides a more precise spatial representation than 1st order, capturing sound from various directions with greater accuracy.
- This makes it suitable for more sophisticated applications that require a higher level of spatial accuracy.

3rd Order Ambisonics

- Further increases spatial resolution and accuracy by introducing 16 channels. W, X, Y, Z, U, V, T, S, R, Q, P, O, N, M, L, and K.
- This technology offers a more detailed and nuanced representation of sound sources in a three-dimensional space.
- It is ideal for advanced applications, particularly in virtual reality and augmented reality environments where a highly realistic and immersive audio experience is desired.

The virtual reality (VR) environment was created by recording a 360-degree video of the tractor cab using an Insta 360 One X2 360 camera in a monoscopic format. Monoscopic videos project a sequence of flat images onto a sphere surrounding the viewer at a rate of 25 frames per second. The video was experienced using a Meta Quest 1 VR headset.

Fig 30: monoscopic view captured with Insta360 One X2 camera.

Fig 31-32: Insta 360 One X2 camera (left); Meta Quest 1 VR headset (right).

The volume of each recording from the Zylia ZM-1 microphone array was manually fine-tuned to align with the LZeq of the NTi omnidirectional microphone recordings taken inside the tractor cabin at the operator's ear level. This calibration process involved placing the NTi microphone at the listener's ear level within the audio space lab, playing back each sample corresponding to the operating mode listed in Table 5, and adjusting the playback volume according to the respective LZeq value.
The script of the immersive test allows for a coordinated test environment involving three programs: Bidule for audio control, Unreal Engine for playing audio stimuli, and MATLAB for facilitating communication between the two through Open Sound Control (OSC). It establishes OSC connections, sets file paths and selects specific sound files. The script manages subject information, manages training stimuli, and prompts users to rate audio stimuli on a scale of 1 to 10. The script collects Rating Scale (RS) and Semantic Differential Method (SDM) ratings for all modes of operation and organises all collected data into Excel files. It also provides an automated structure for running experiments with audio stimuli, managing subject data and exporting results.

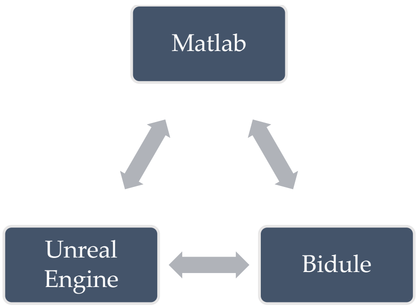

Fig 33: communication between programs for the immersive test.

2.3.2.2. Binaural Listening Test

The binaural listening test consisted in listening 10 seconds audio samples recorded by B&K 4141 e Siemens ABH04 binaural microphone headsets. The test was conducted with the SQala module of ArtemiS Suite software developed by Head Acoustics. The playback system consisted in a pair of Senheiser HD650 headphones, a laptop and a Roland Octa-Capture audio interface. To ensure a correct playback simulating the recorded conditions with both microphones, some steps are needed.

If a head and torso simulator is used for the measurements, an adapted equalization to the measurement environment should be performed in order to reproduce the sounds and directional effects as close to the original as possible. For measurements in vehicles, the ID (Independent Direction) equalization has proven to be advantageous [28], as it only compensates for the direction-independent parts (resonances) of the outer ear's transmission function. This compensation is necessary because the sound that has passed through the ear canal before it reaches the microphone returns to the ear canal during playback, so we have a double resonance; with equalization, the signal that reaches the eardrum is the same as if the listener were present in the sound field.

Fig 34-35: B&K 4101 microphone headset position inside the ear (left); Siemens ABH04 microphone headset position outside the ear (right).

When using a B&K microphone positioned in the cavum conchae before the ear canal, it is important to consider the resonances of the cavum conchae. Although the ID equalization may seem important, a closer inspection reveals that its effect at medium-low frequencies is nearly linear, with maximum attenuation occurring around 5kHz at 10 dB. Notably, the recorded sound samples exhibit high sound pressure levels at lower frequencies, primarily due to HVAC and tractor noise caused by engine. However, binaural listening tests demonstrated no perceptual differences between equalized and non-equalized playback, ID equalization has minimal impact on accurate sound playback for this specific scenario and may not be used.

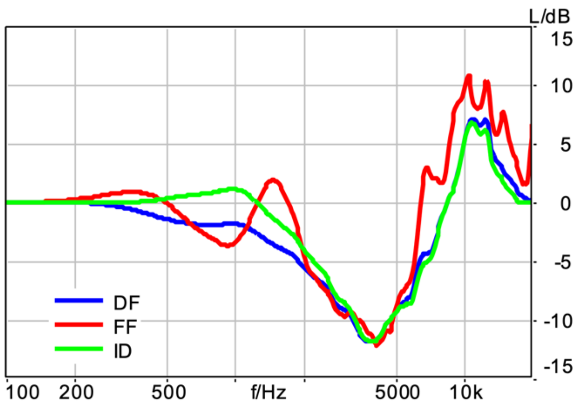

Fig 32: independent-direction, diffuse field and free field equalization curves.

Regarding the Siemens binaural headset, the microphone is positioned outside of the ear. Therefore, no ID equalization that takes into account the outer ear resonances should be used, whether diffuse or free field, due to the specific environment of the tractor cabin.
To ensure fidelity in playback volume, adjustments were made by conducting tests in an anechoic chamber. The same artificial head equipped with Sennheiser HD650 headphones and the B&K binaural microphone were used to fine-tune the playback volume. The sound pressure level and loudness level of the tractor recordings were matched to the B&K recordings with headphones on in the anechoic chamber through iterative gain adjustments. This process ensures a precise and realistic reproduction of sound in the tractor cabin, accounting for the complexities of the recording environment.
ArtemiS Suite automatically applies headphones equalization to eliminate any influence of the playback system on the listening experience.

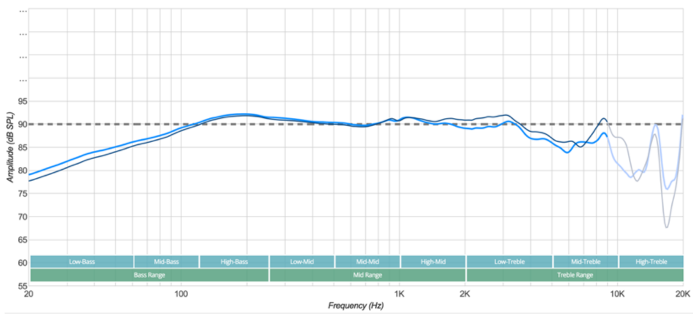

Fig 36-37: playback volume adjustment in anechoic chamber (left); Sennheiser HD650 frequency response (right).

3. Results

This section presents the results of psychoacoustic analyses and predictive model outcomes. In the psychoacoustic analysis, microphone recordings for each HVAC system operating mode are trimmed to ten seconds to eliminate accidental noises. Psychoacoustic parameters described in the previous section are computed and compared with the subjective assessment results. The tests employed two methods: the 1-10 Annoyance Rating Scale and the Semantic Differential Method (SDM), which uses four adjective pairs to represent Annoyance, Loudness, Sharpness, and Roughness. Mean values were calculated for each assessment method across all evaluators for each HVAC operating mode. The prediction model will be developed using **multiple linear regression**. Mean ratings will serve as dependent variables, while previously calculated psychoacoustic parameters will serve as independent variables. The results exclude psychoacoustic parameters such as Fluctuation Strength and Tonality due to their low and stationary values across the various HVAC modes. The primary parameters included are the Equivalent Sound Pressure Level (**LZeq**), the A-weighted Equivalent Sound Pressure Level (**LAeq**), **Loudness**, **Sharpness**, and **Roughness**. Supplementary material also includes Speech Interface Level and Articulation Index.

3.1. Psychoacoustics Analyses

The recorded samples of every microphone were cut to a length of 10 second, this portion will be analysed by computing psychoacoustic parameters and then used as hearing test for the subjective assessment. The calibration for Siemens ABH04 B&K 4101 microphone headset was automatically loaded by Testlab, this could be checked by viewing the SPL of the calibration file signals. The calibration was loaded manually for the NTi omnidirectional microphone noise recordings. The psychoacoustic parameters were calculated using Simcenter Testlab with the following settings:

Acquisition Parameters: Tracking and Triggering: Measurement mode > Stationary, Tracking method> Time, Duration> 300 s, Acquisition rate> 2 avg/s, Number of averages> 601; FS Acquisition: Resolution> 50.0 - 1.0 Hz] [Channel Processing: Format> RMS] [Sections: Frame Statistics > RMS, Mean, 90th Percentile; Overall Level > ✓ Overall Level, Acoustic Channels > ✓ Additional overall level > weighting> A; Level calculation > LAeq – A – weighted equivalent continuous sound level> Add; Psychoacoustic Metrics> Loudness ISO 532A, Sharpness - diffuse field; Modulation Metrics > Roughness, Fluctuation Strenght; the frequency spectrum and FFT vs time were calculated using ArtemiS Suite software with an FFT size of 4096 and a Hanning window of 50%.

3.1.1. Binaural Microphones

Eleven sound samples were chosen from the binaural headset noise recordings, each representing a noise condition corresponding to different HVAC operating modes at various speeds and with the engine both on and off (as detailed in **Table 5**). Specifically, seven operating modes were selected from the **B&K 4101** microphone recordings, while four operating modes were chosen from the **Siemens ABH04** microphone recordings. The 10-second noise samples captured by both headsets will be played for subjective assessment. The subjective ratings obtained will be utilized as the dependent variable, while the psychoacoustic results will serve as the independent variables for the prediction model.

3.1.1.1. Equivalent Sound Pressure Level

The sound pressure levels exhibit a linear rise corresponding to the increase in HVAC speed and the activation of the engine. The sound level at 'ES7' with the compressor off is equivalent to that at 'ES4' with the compressor on.

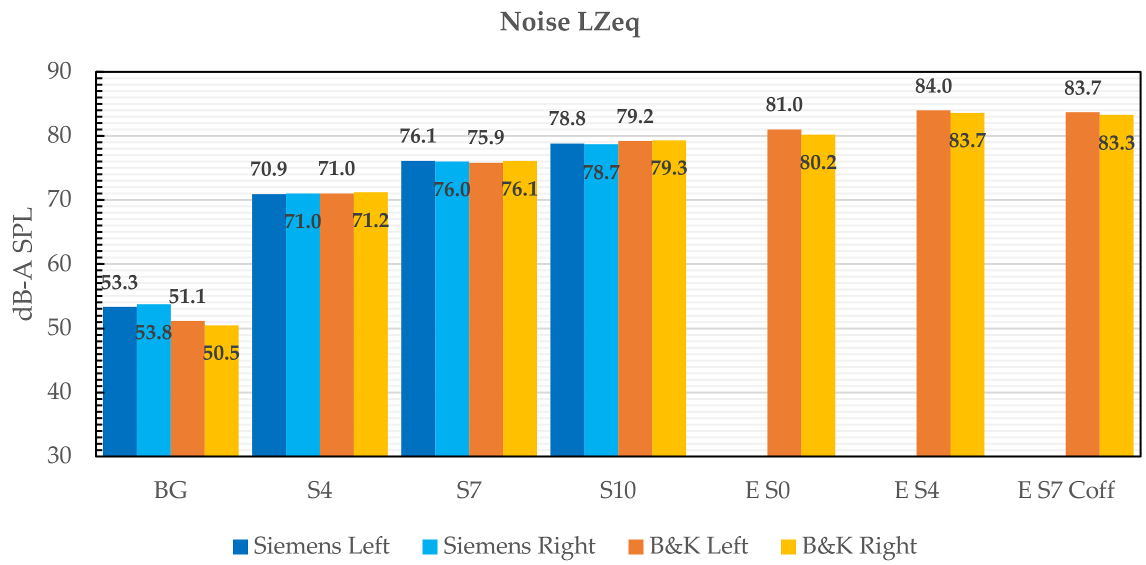

Fig 38:(y-axis) noise equivalent sound pressure level expressed in dB; (x-axis) type of measurement from **Table 5**; left and right ear are considered for Siemens (blue) and B&K headsets (orange); Just Noticeable Difference (JND) is 1dB.

3.1.1.2. A-weighted Equivalent Sound Pressure Level

The differences in A-weighted sound pressure levels between the operating modes are more noticeable. Specifically, the A-weighted value of 'ES7' with the compressor off is higher than that of 'ES4' with the compressor on.

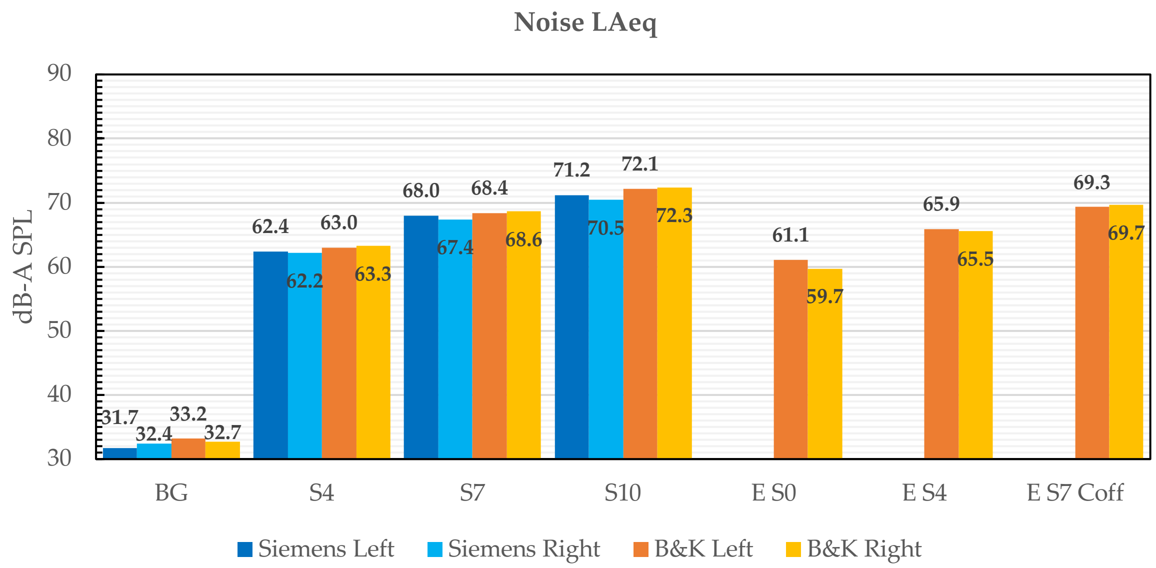

Fig 39:(y-axis) noise A-weighted sound pressure level expressed in dBA; (x-axis) type of measurement from **Table 5**; left and right ear are considered for Siemens and B&K headsets; Just Noticeable Difference (JND) is 1dB.

3.1.1.3. Loudness

The differences in loudness level, measured in sones, are significantly more pronounced compared to the differences in LZeq and LAeq SPLs. Although 'ES4' has the highest linear SPL in dB, in terms of loudness, it is equivalent to 'S4' with the engine off. The highest condition in terms of loudness is 'S10' with the HVAC set to speed 10 and the engine off. A distinct difference in loudness can be observed for each operating mode between the two microphone headsets, with the largest disparity occurring at 'S10', amounting to 4 sones.

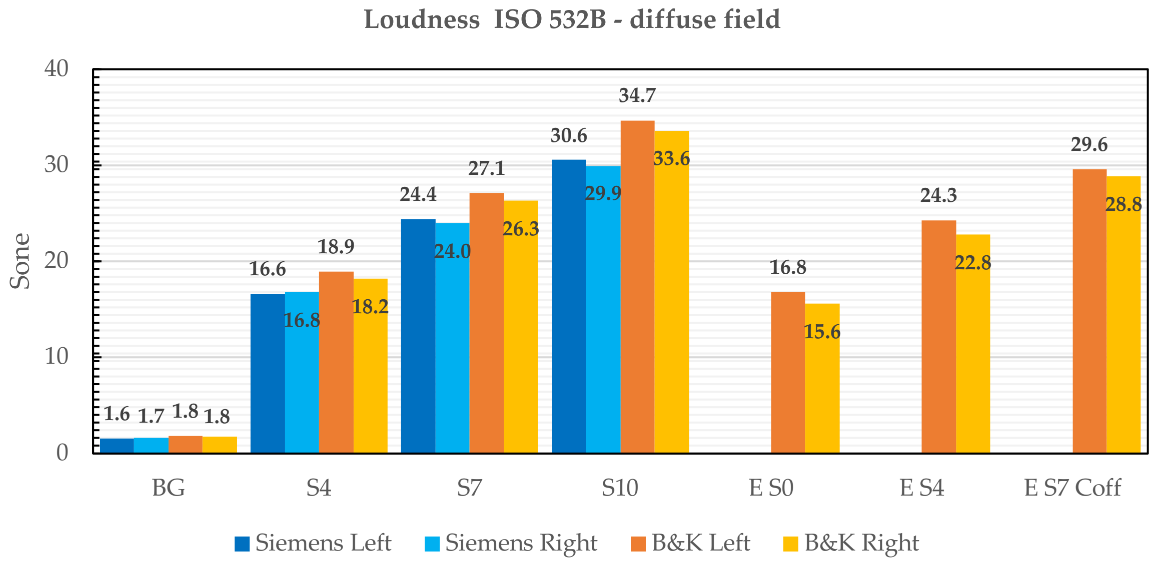

Fig 40:(y-axis) mean Loudness level expressed in sone; (x-axis) type of measurement from **Table 5**; left and right ear are considered for Siemens (blue) and B&K headsets (orange); Just Noticeable Difference (JND) is 0.8 sone.

3.1.1.4. Sharpness

Sharpness values increase slightly for every operating mode, with a maximum value of 1.38 acum in 'S10'. The Siemens ABH04 microphone headset noise samples result in less sharpness than the B&K 4101 headset.

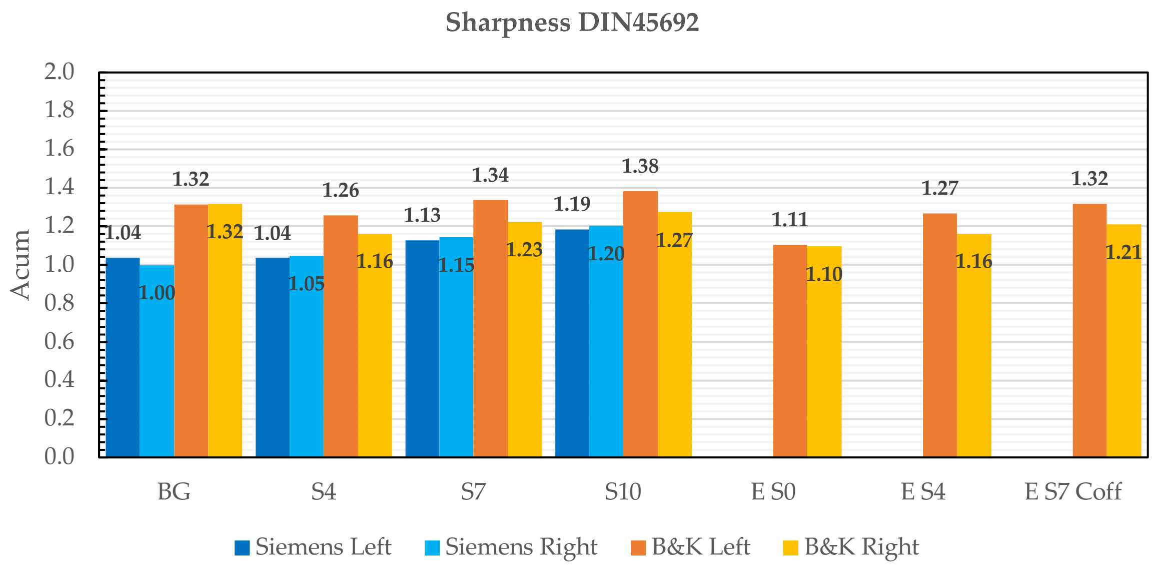

Fig 41:(y-axis) mean Sharpness level expressed in acum; (x-axis) type of measurement from **Table 5**; left and right ear are considered for Siemens and B&K headsets; Just Noticeable Difference (JND) is 0.04 acum.

3.1.1.5. Roughness

The Roughness values are generally low, but they escalate with the increase in HVAC speed, peaking at 0.27 asper at 'S10'. The Roughness of the motor at HVAC speed 4 ('ES4') is equivalent to that of 'S4' with the engine off.

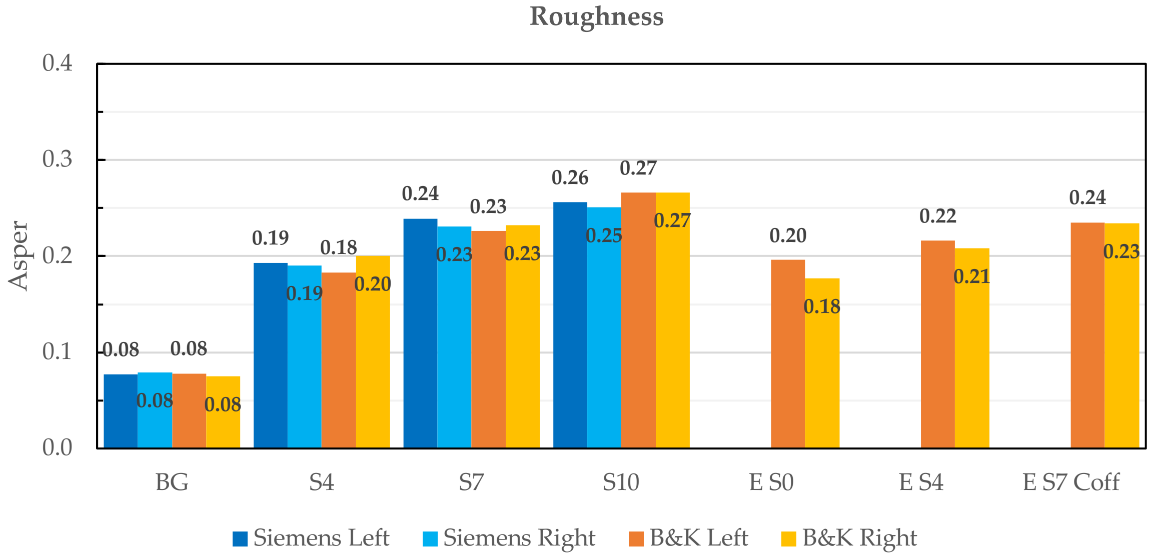

Fig 42:(y-axis) mean Roughness level expressed in sone; (x-axis) type of measurement from **Table 5**; left and right ear are considered for Siemens and B&K headsets; Just Noticeable Difference (JND) is 0.8 sone.

3.1.1.6. Frequency Spectrum

A frequency-smoothed spectrum is generated to enhance the comprehension of certain frequency components' influence. The most significant impact on SPL occurs in the low frequencies. For high frequencies, differences in SPL become smaller. The spectral characteristics of the B&K 4101 spectrum in Figure X-X closely resemble those of the Siemens ABH04 spectrum in Figure X-X.

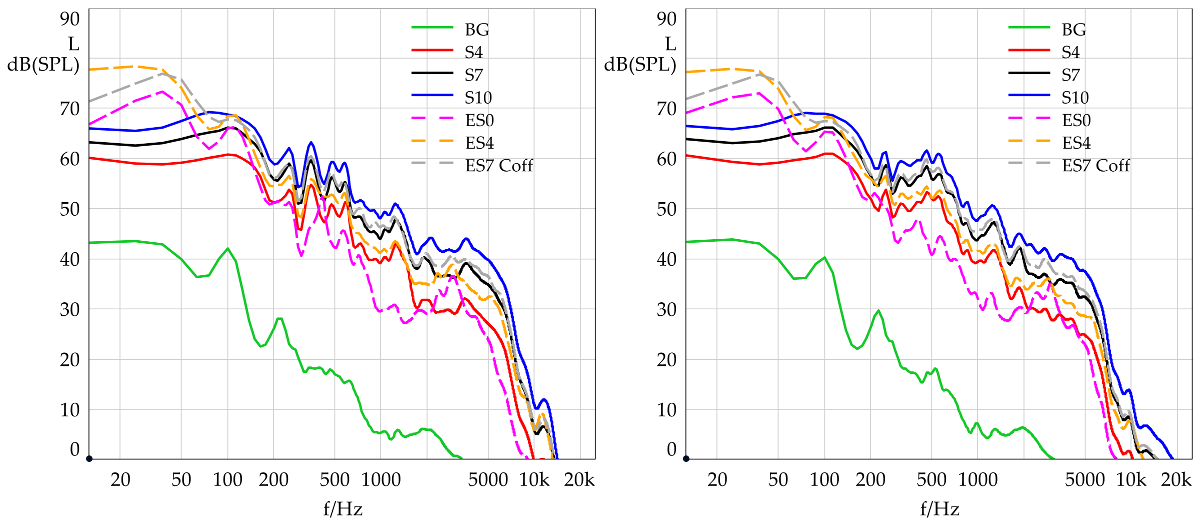

Figure 43-44: Spectrum Analysis of the B&K 4101 binaural microphone headset, depicting the left (43) and right (44) channels. Dashed lines indicate HVAC speeds with the tractor engine running.

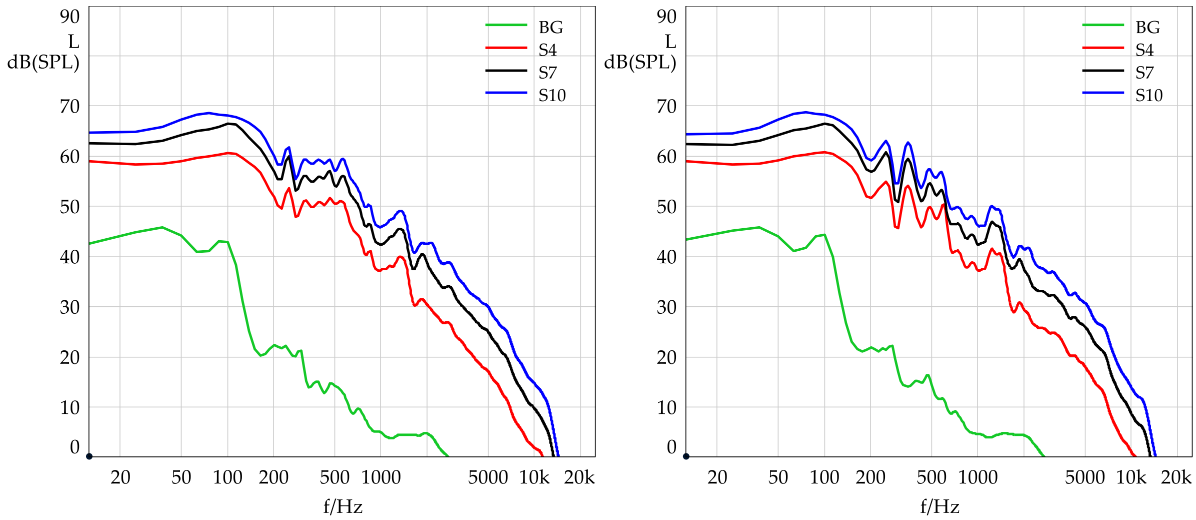

Figure 45-46: Spectrum Analysis of the Siemens ABH04 binaural microphone headset, depicting the left (45) and right (46) channels. Dashed lines indicate HVAC speeds with the tractor engine running.

3.1.1.7. FFT vs Time

By conducting FFT vs Time analysis, it becomes possible to compare the frequency influence over time. From the graphs, it is evident that we are dealing with stationary signals, as the SPL signal remains constant both over time and across frequencies.

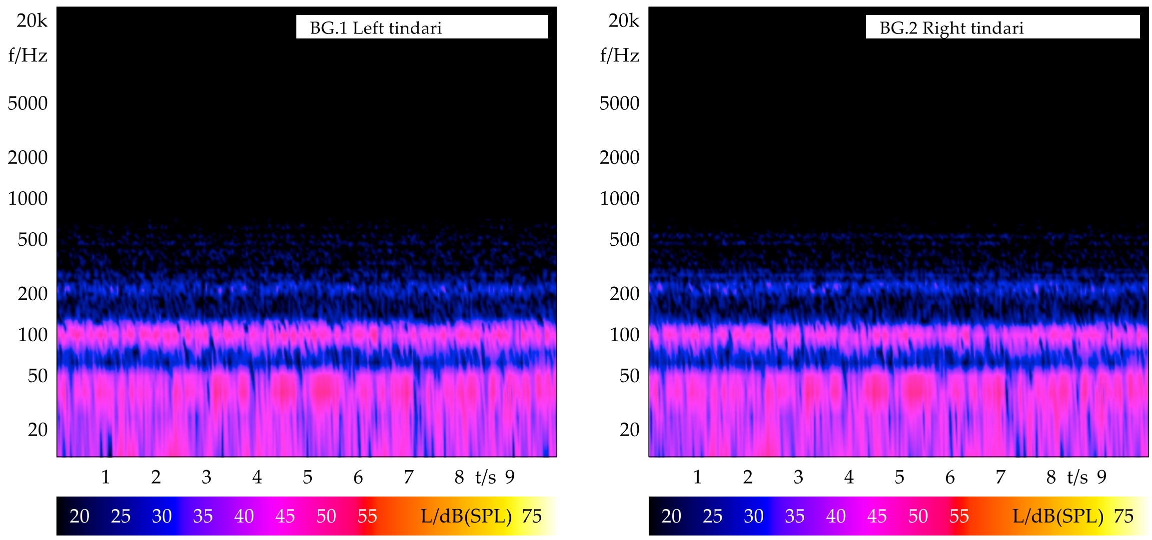

Figure 47-48: FFT vs Time of the B&K 4101 binaural microphone headset, depicting the left (47) and right (48) channels. BG is the current operating mode.

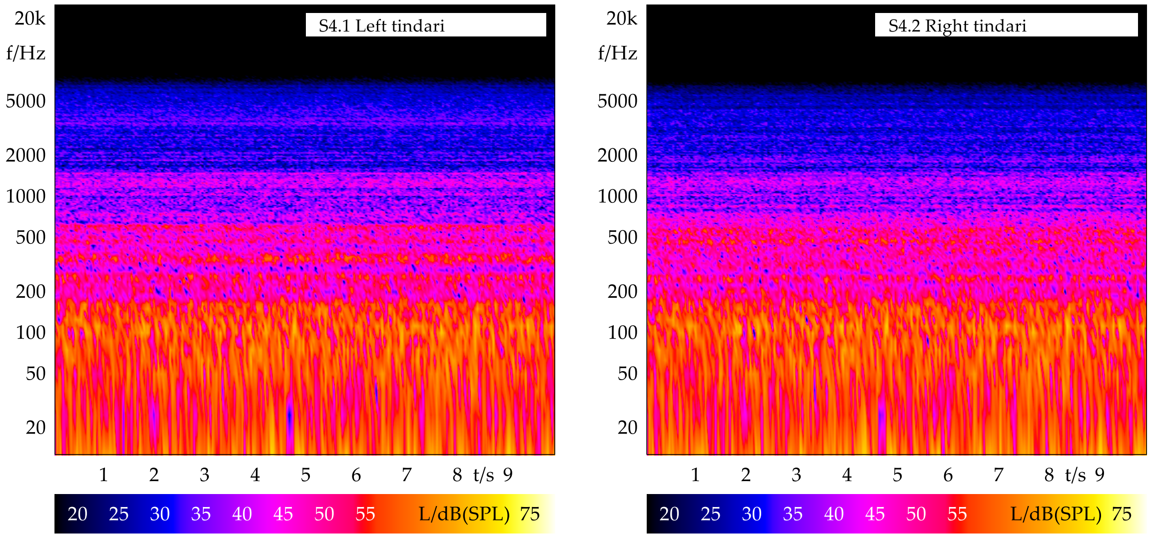

Figure 49-50: FFT vs Time of the B&K 4101 binaural microphone headset, depicting the left and right channels. S4 is the current operating mode.

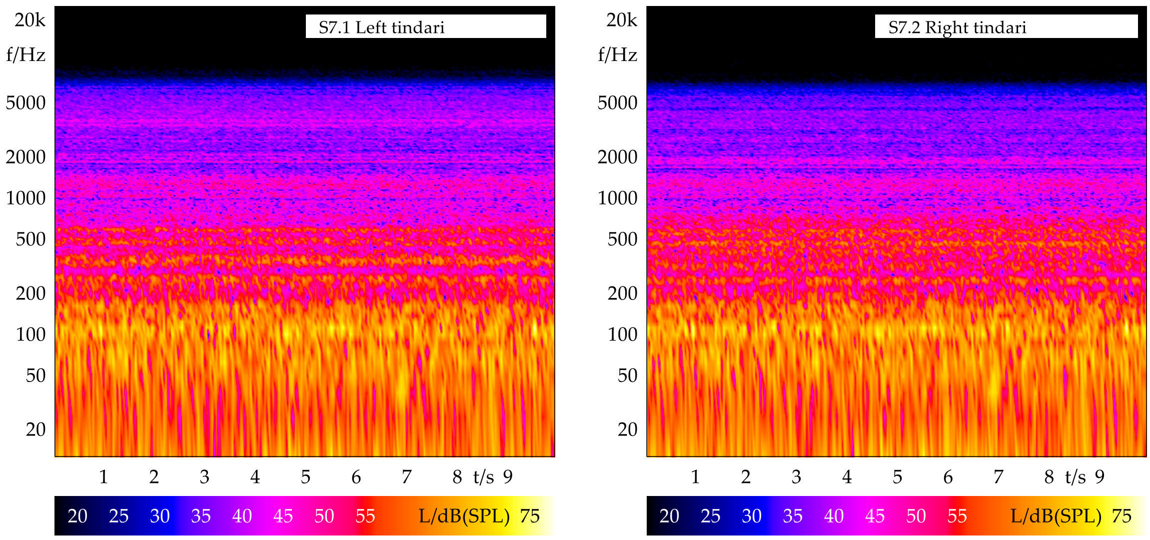

Figure 51-52: FFT vs Time of the B&K 4101 binaural microphone headset, depicting the left and right channels. S7 is the current operating mode.

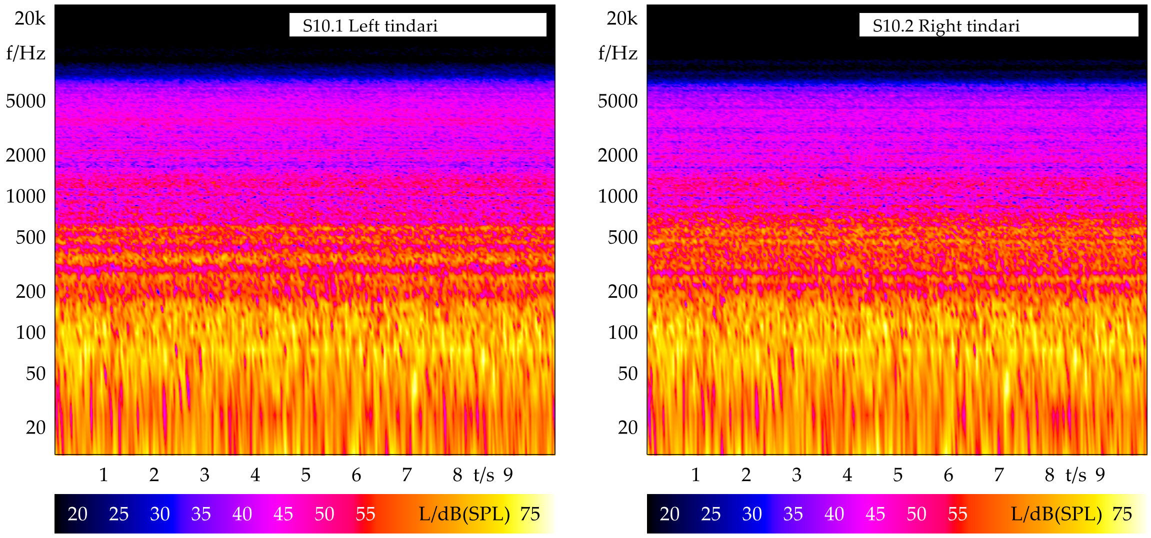

Figure 53-54: FFT vs Time of the B&K 4101 binaural microphone headset, depicting the left and right channels. S10 is the current operating mode.

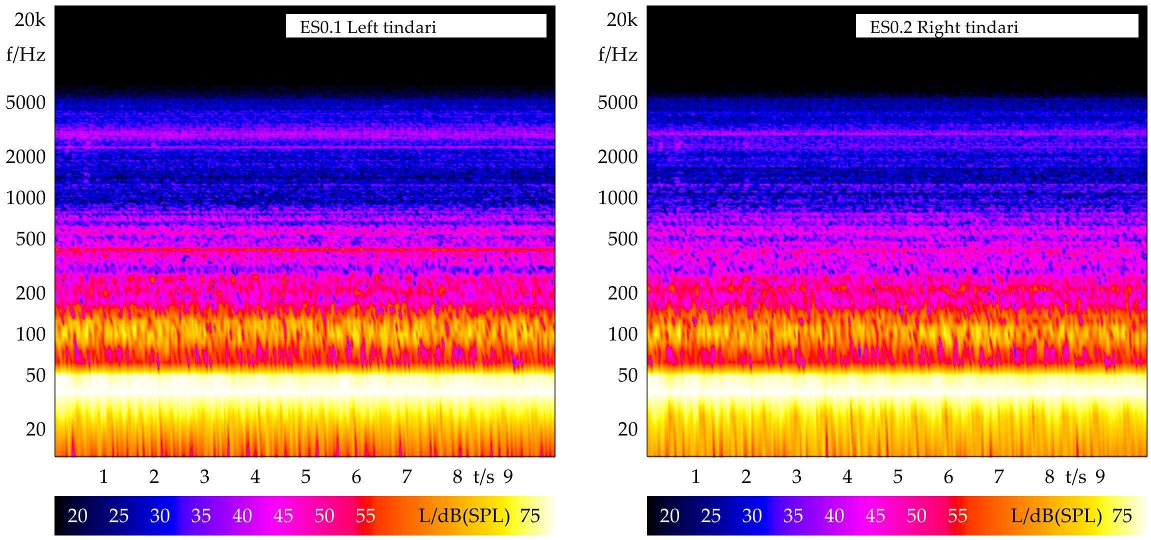

Figure 55-56: FFT vs Time of the B&K 4101 binaural microphone headset, depicting the left and right channels. ES0 is the current operating mode.

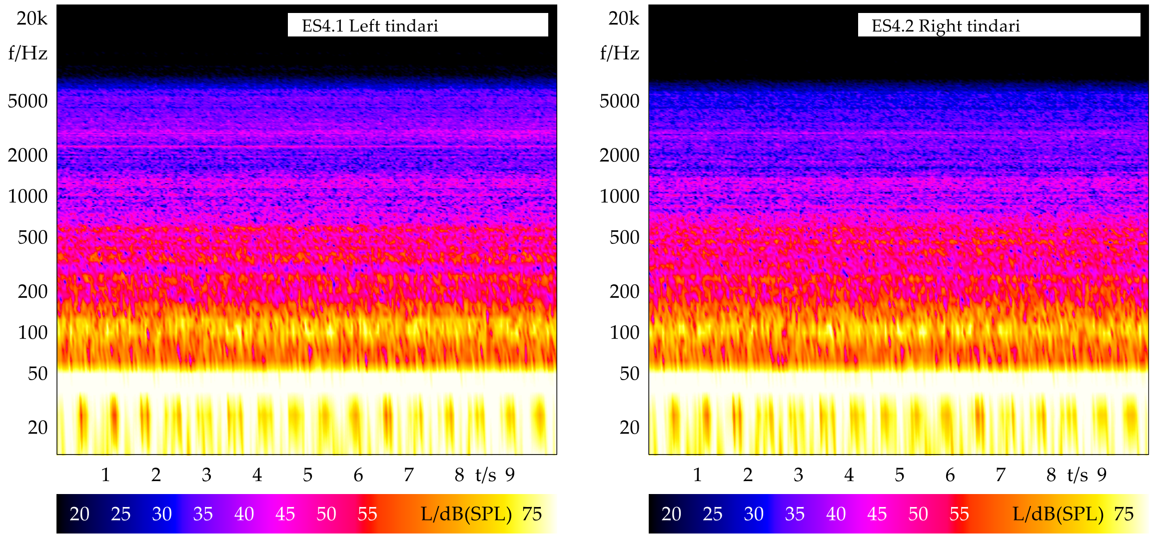

Figure 57-58: FFT vs Time of the B&K 4101 binaural microphone headset, depicting the left and right channels. ES4 is the current operating mode.

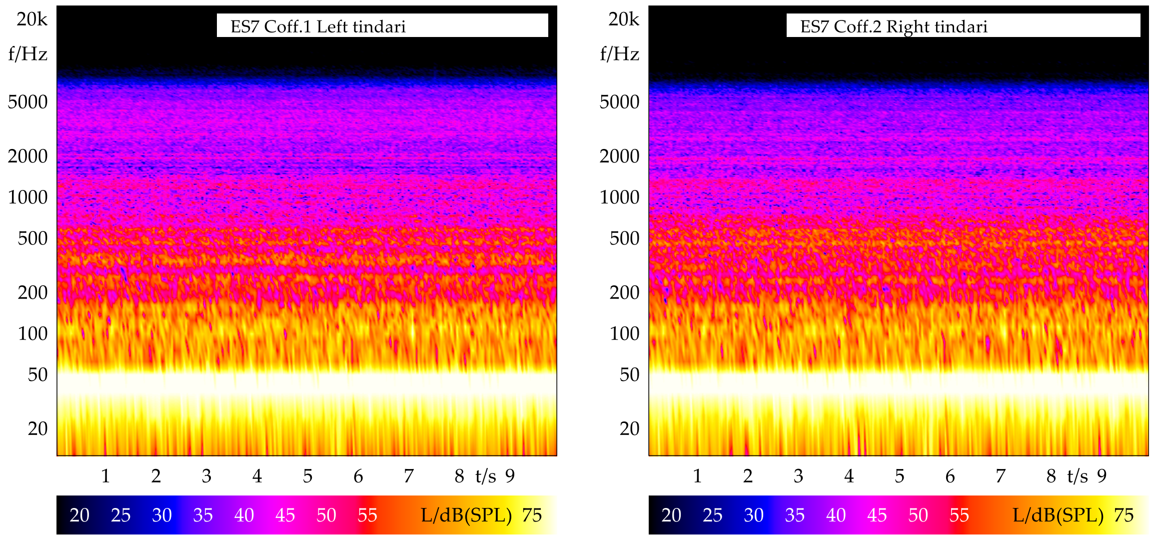

Figure 59-60: FFT vs Time of the B&K 4101 binaural microphone headset, depicting the left and right channels. ES7 is the current operating mode.

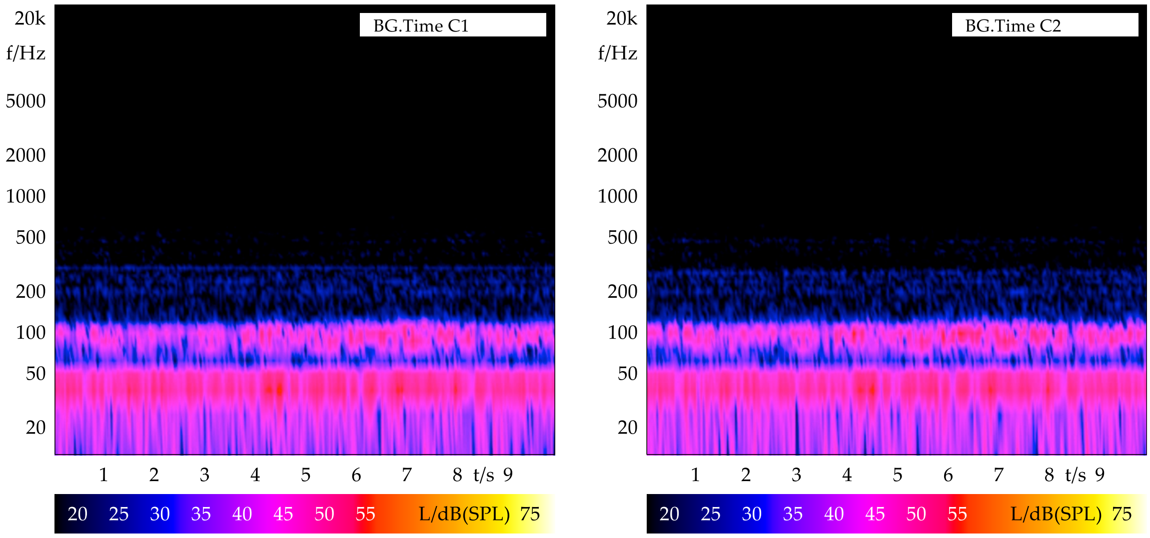

Figure 61-62: FFT vs Time of the Siemens ABH04 binaural microphone headset, depicting the left (C1) and right (C2) channels. BG is the current operating mode.

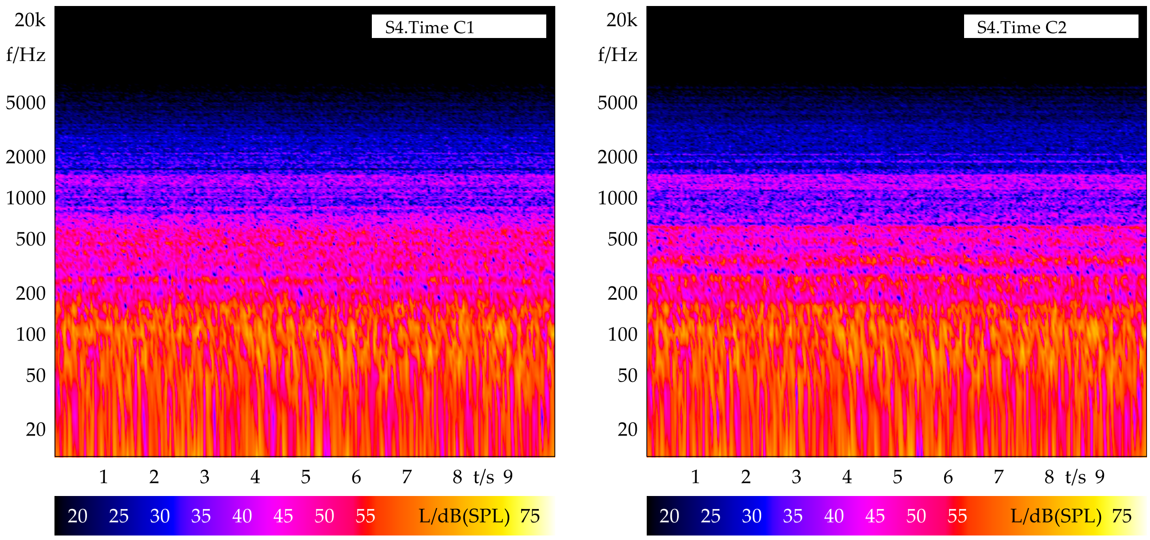

Figure 63-64: FFT vs Time of the Siemens ABH04 binaural microphone headset, depicting the left (C1) and right (C2) channels. S4 is the current operating mode.

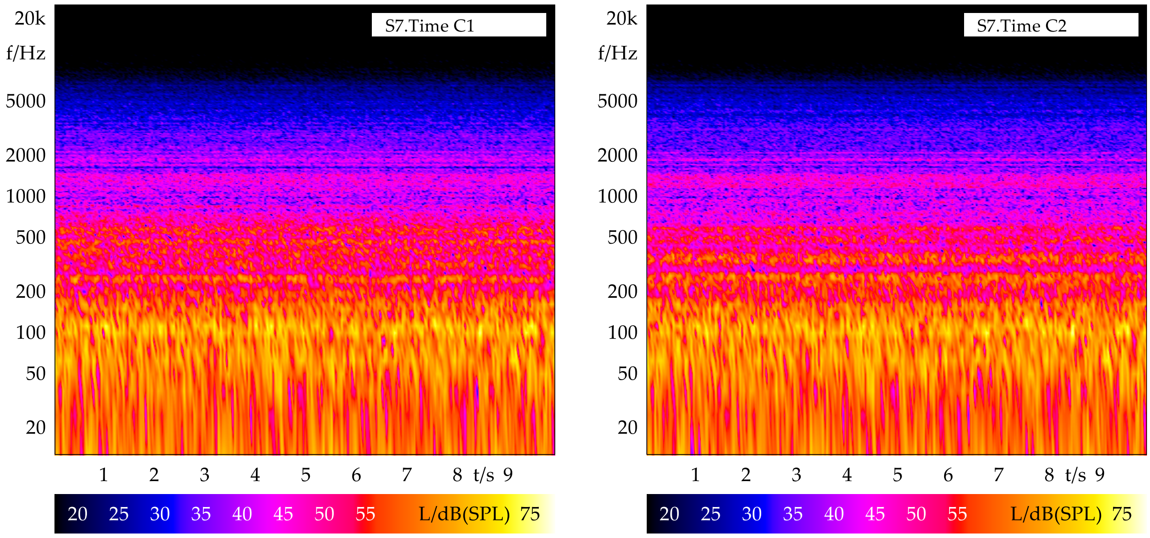

Figure 65-66: FFT vs Time of the Siemens ABH04 binaural microphone headset, depicting the left (C1) and right (C2) channels. S7 is the current operating mode.

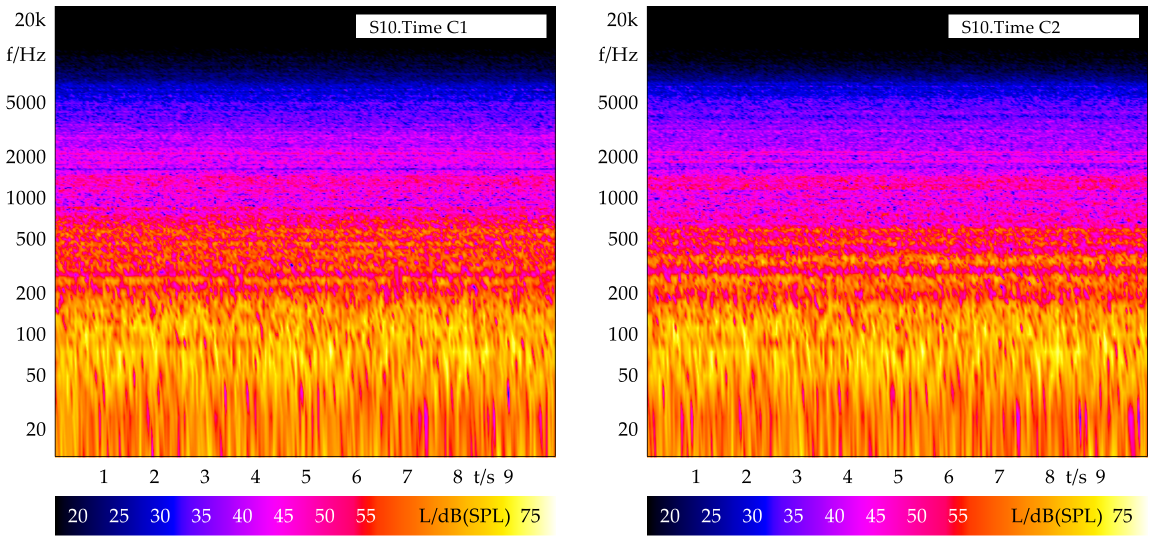

Figure 67-68: FFT vs Time of the Siemens ABH04 binaural microphone headset, depicting the left (C1) and right (C2) channels. S10 is the current operating mode.

3.1.1.8. Supplementary material

The articulation index (AI) quantifies speech intelligibility in noisy environments by analysing the relative levels of speech and background noise. The AI value closer to 1 or 100% indicates a higher ability to hear speech accurately.

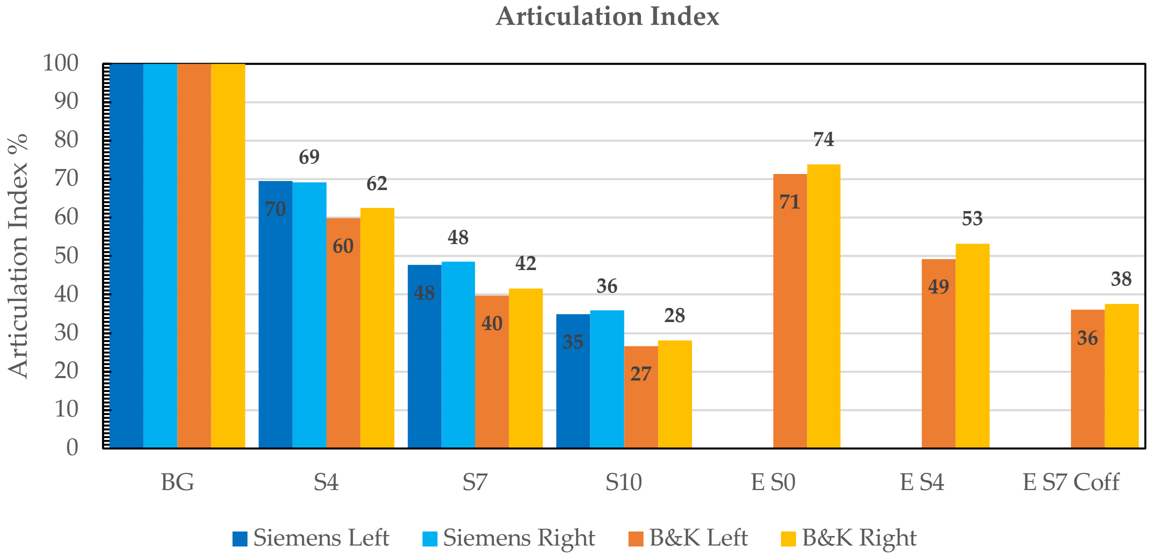

Fig 69:(y-axis) Articulation Index level expressed in percentage; (x-axis) type of measurement from **Table 5**; left and right ear are considered for Siemens and B&K headsets.

The Speech Interference Level (SIL) measures the amount of interference caused by background noise on speech communication. It provides insight into the clarity and intelligibility of speech in noisy environments.

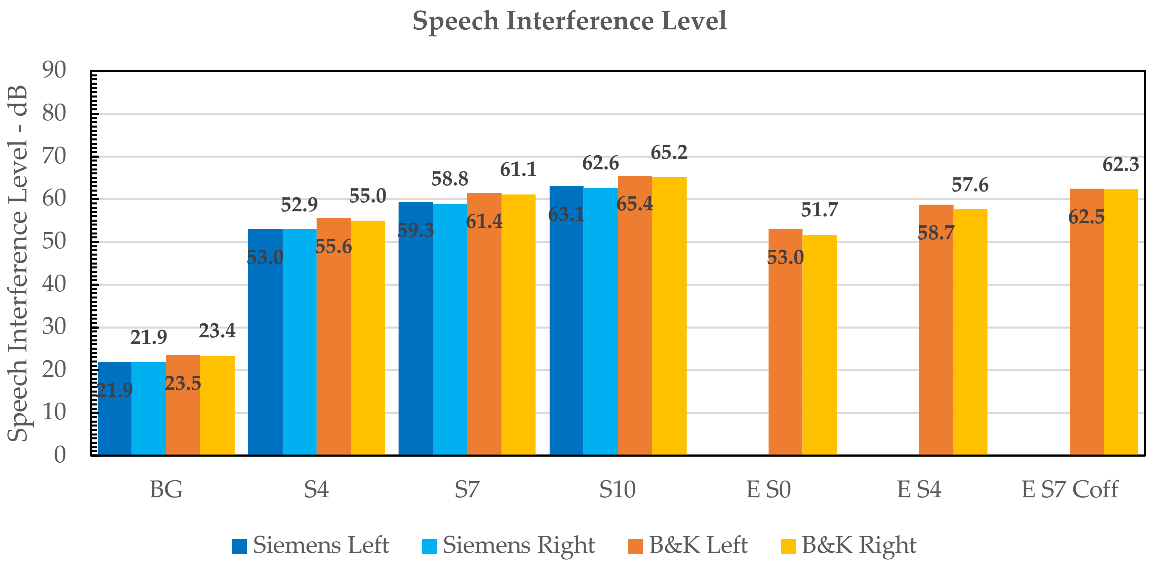

Fig 70:(y-axis) Speech Interface Level expressed in percentage; (x-axis) type of measurement from **Table 5**; left and right ear are considered for Siemens and B&K headsets.

3.1.2. Omnidirectional Microphones

Nine sound samples were chosen from the NTi microphone omnidirectional noise recordings, each representing a noise condition corresponding to different HVAC operating modes at various speeds and with the engine both on and off (as detailed in **Table 5**). Each parameter is computed using a 10-second excerpt derived from the original 60 seconds of the recorded sample. The selection of this 10-second segment involves assessing portions of the signal free from accidental sounds. The 10-second version will also be employed for subjective evaluation.

3.1.2.1. Linear Equivalent Sound Pressure Level

The sound pressure levels exhibit a linear rise corresponding to the increase in HVAC speed and the activation of the engine. The maximum SPL is reached with engine on and HVAC speed 10 'ES10', at 85.6 dB.

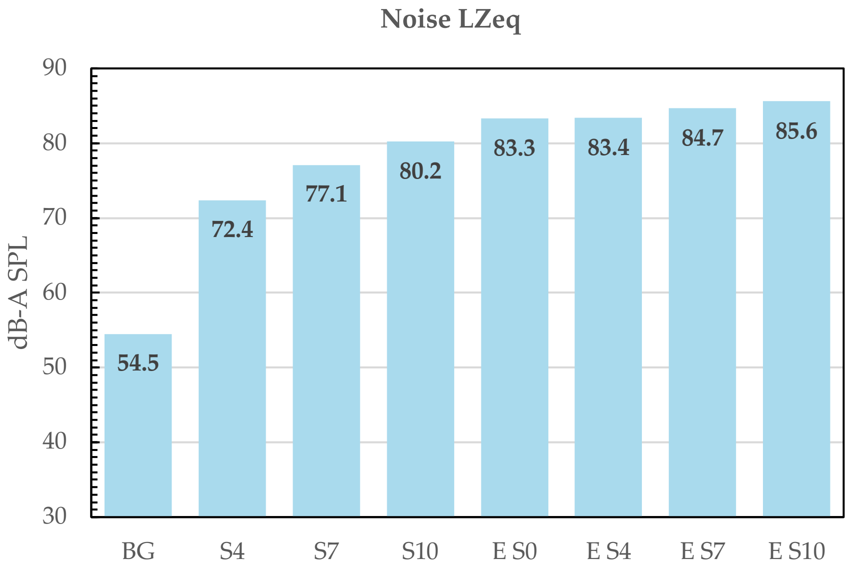

Fig 71:(y-axis) noise sound pressure level expressed in dB; (x-axis) type of measurement from **Table 5**; omnidirectional NTi microphone; Just Noticeable Difference (JND) is 1dB.

3.1.2.2. A-weighted Equivalent Sound Pressure Level

The discrepancies in A-weighted sound pressure levels among the operating modes are more conspicuous. Specifically, there is only a 2.6 dBA difference between 'S10' with the engine off and 'ES10' with the engine on.

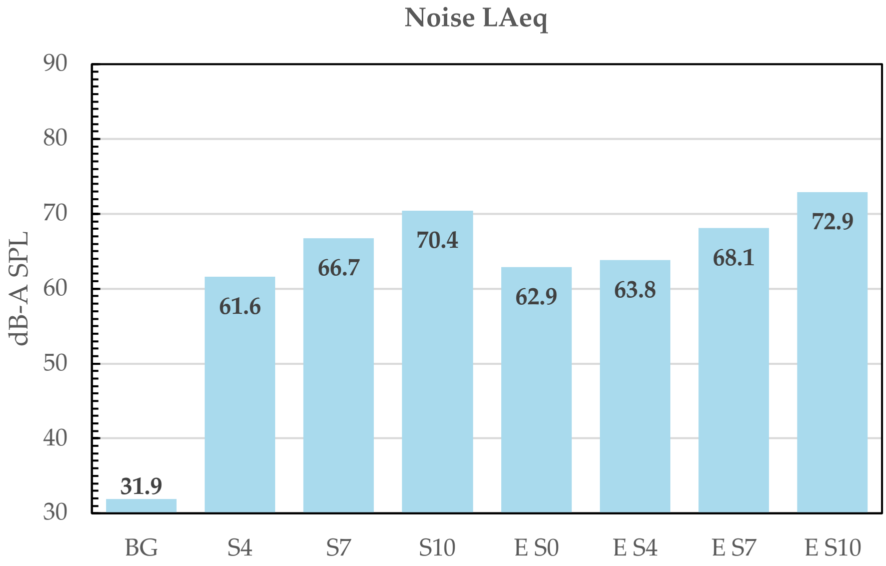

Fig 72:(y-axis) noise A-weighted sound pressure level expressed in dB; (x-axis) type of measurement from **Table 5**; omnidirectional NTi microphone; Just Noticeable Difference (JND) is 1dB.

3.1.2.3. Loudness

The disparities in loudness level, quantified in sones, are notably more pronounced than the differences in LZeq and LAeq SPLs. Particularly, the contrast in loudness between 'ES10' and 'S10' is more distinct. 'S10' remains louder than 'ES7' even with the engine turned on.

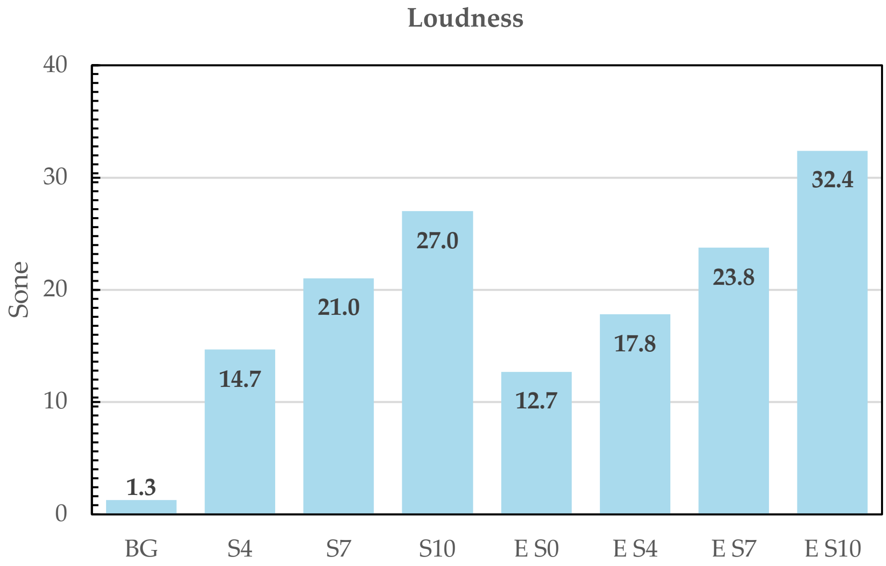

Fig 73:(y-axis) mean Loudness values expressed in sone; (x-axis) type of measurement from **Table 5**; omnidirectional NTi microphone; Just Noticeable Difference (JND) is 0.8 sone.

3.1.2.4. Roughness

The Roughness values are generally low, but they increase with the rise in HVAC speed, reaching a peak of 0.27 asper at 'S10'. The Roughness of the motor at HVAC speed 4 ('ES4') is merely 1 asper higher than that of 'S4' with the engine off.

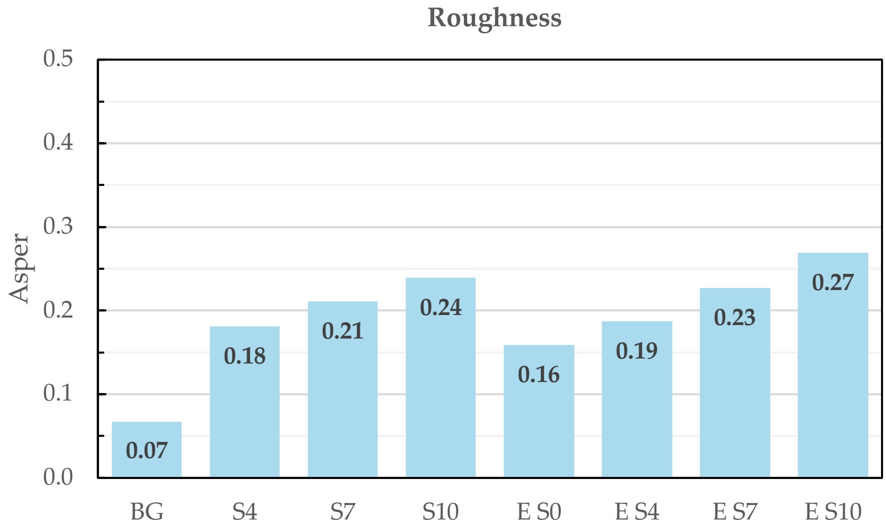

Fig 74:(y-axis) mean Roughness value expressed in asper; (x-axis) type of measurement from **Table 5**; omnidirectional NTi microphone; Just Noticeable Difference (JND) is 0.05 asper.

3.1.2.5. Sharpness

Sharpness values exhibit a slight increase for each operating mode, reaching a maximum value of 1.18 acum in 'ES10'.

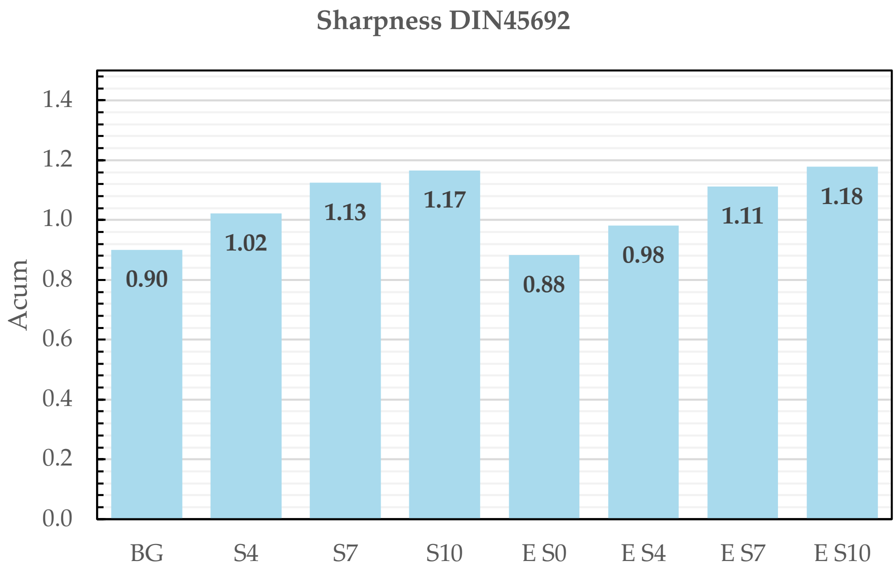

Fig 75:(y-axis) mean Sharpness value expressed in acum; (x-axis) type of measurement from **Table 5**; omnidirectional NTi microphone; Just Noticeable Difference (JND) is 0.04 acum.

3.1.2.6. Frequency Spectrum

A frequency-smoothed spectrum is generated to enhance the comprehension of certain frequency components' influence. The most significant difference in SPL occurs in the low frequencies. For high frequencies, differences in SPL become smaller.

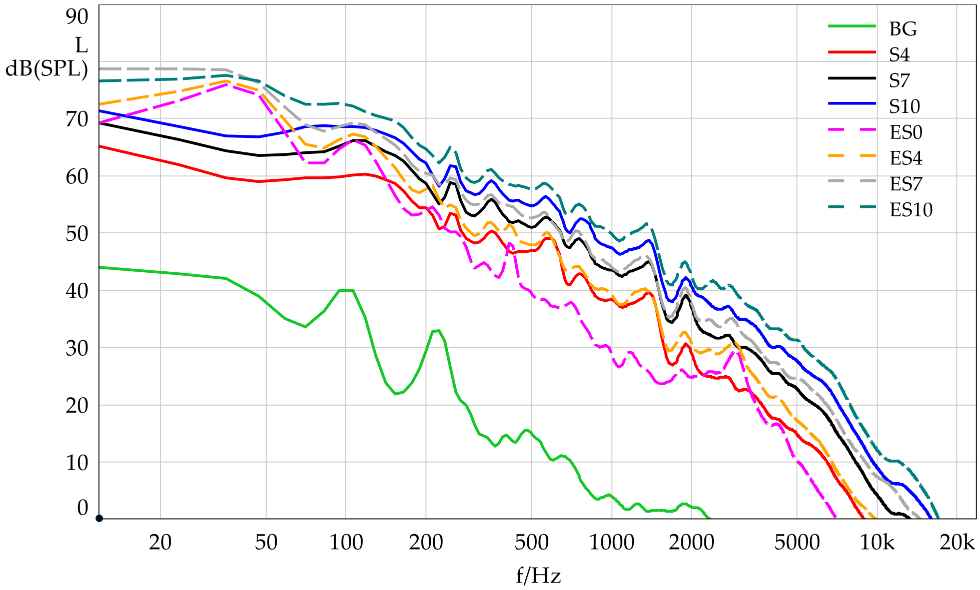

Figure 76: Spectrum Analysis of the NTi omnidirectional microphone; dashed lines indicate HVAC speeds with the tractor engine running.

3.1.2.7. FFT vs Time

By performing FFT vs Time analysis, we can compare the frequency influence over time. The graphs indicate that we are dealing with stationary signals, as the SPL signal remains consistent over time and across frequencies. Additionally, the omnidirectional spectrum closely resembles the binaural spectrums.

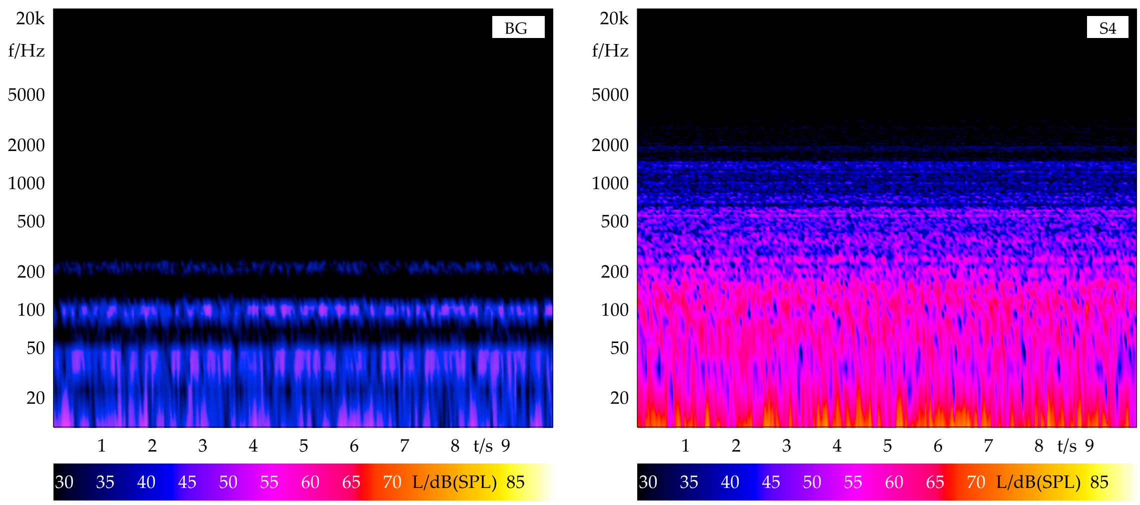

Figure 77-78: FFT vs Time of the NTi omindirectional microphone headset, BG mode on the left, S4 on the right.

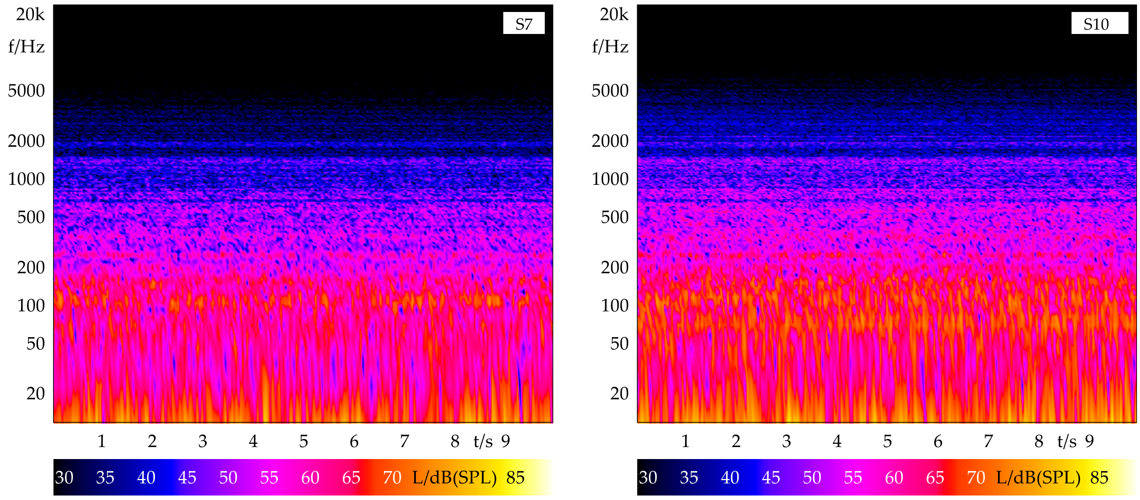

Figure 79-80: FFT vs Time of the NTi omindirectional microphone headset, S7 mode on the left, S10 on the right.

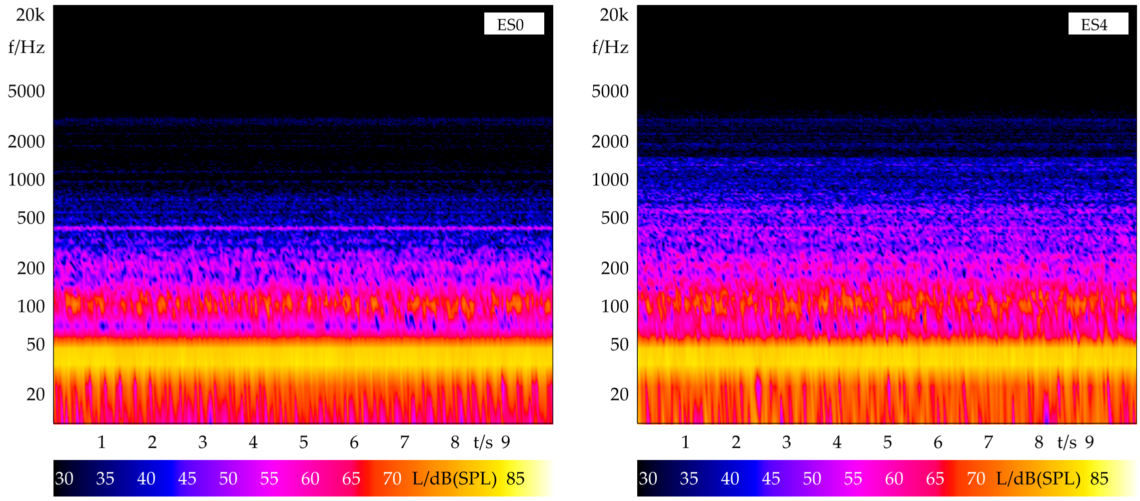

Figure 81-82: FFT vs Time of the NTi omindirectional microphone headset, ES0 mode on the left, ES4 on the right.

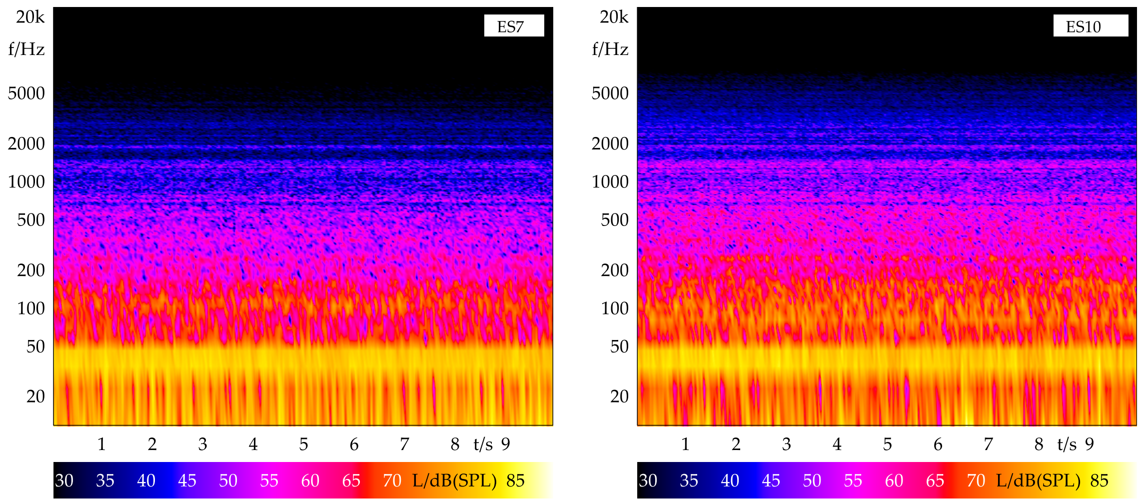

Figure 83-84: FFT vs Time of the NTi omindirectional microphone headset, ES7 mode on the left, ES10 on the right.

3.1.2.8. Supplementary Material

The articulation index (AI) quantifies speech intelligibility in noisy environments by analyzing the relative levels of speech and background noise. The AI value closer to 1 or 100% indicates a higher ability to hear speech accurately.

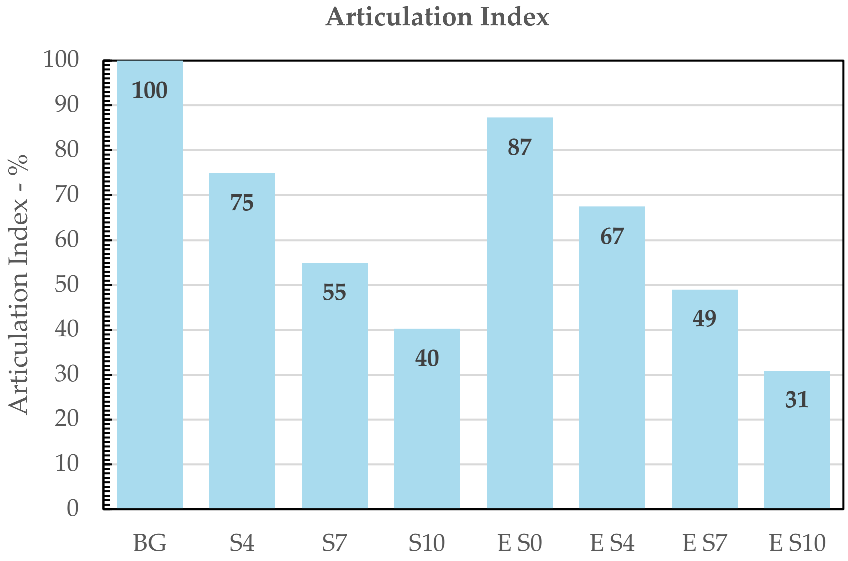

Fig 85:(y-axis) Articulation Index level expressed in percentage; (x-axis) type of measurement from **Table 5**; noise samples from NTi omnidirectional microphone.

The Speech Interference Level (SIL) measures the amount of interference caused by background noise on speech communication. It provides insight into the clarity and intelligibility of speech in noisy environments.

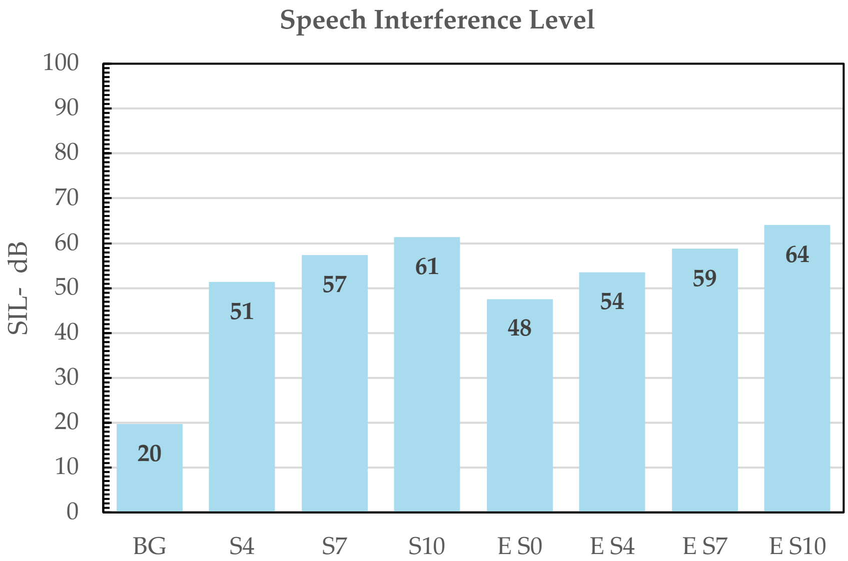

Fig 86:(y-axis) Speech Interface Level expressed in percentage; (x-axis) type of measurement from **Table 5**; noise samples from NTi omnidirectional microphone.

3.2. Subjective Assessment

A comprehensive analysis and comparison of subjective evaluations have been conducted for both the immersive and binaural listening test. The primary objective is to determine the average subjective scores across various HVAC operating modes and to perform an in-depth correlation analysis between subjective ratings and psychoacoustic parameters including Linear SPL (LZeq), A-SPL (LAeq), Loudness, Sharpness, and Roughness. This analysis holds significance in selecting independent variables for the prediction models, thereby enhancing its accuracy and reliability.

3.2.1. Immersive Listening Test

The immersive test involved listening spatialised audio recordings in ambisonics audio of 3rd order and viewing the interior of the tractor cabin through the use of a VR headset. The operating modes of the HVAC system are recorded with the Zylia ZM-1 microphone and played back to the subjects through a sphere of 16 speakers. The results of the subjective ratings are then compared with the psychoacoustic parameters obtained from the NTi omnidirectional microphone. **Table 9** displays the mean values of the 1-10 Rating Scale for each HVAC mode, and the Annoyance, Loudness, Roughness, and Sharpness scales of semantic differential method.

Table 9: For each operating mode, the mean values of the subjective evaluation are provided for the Annoyance Rating Scale and the Semantic Differential Method scales, encompassing Annoying/Not-Annoying, Loud/Quiet, Sharp/Dull, and Rough/Smooth attributes.

	Annoyance Rating Scale	**SDM: Annoying**	**SDM: Loud**	**SDM: Sharp**	**SDM: Rough**
BG	1.2	-2.5	-2.6	-0.8	-1.9
S4	3.7	-1.2	-0.7	-0.1	-1
S7	5.5	0.7	0.7	-0.1	-0.3
S10	6.1	1.5	1.5	0.9	0.4
ES0	5.2	0.6	0.3	0.3	0
ES4	6	1.2	0.7	-0.8	0.1
ES7	6.5	1.7	1.2	-0.6	0.2
ES7 Coff	6.8	2	1.5	-0.8	0.3
ES10	7.3	2.2	1.8	0.6	0.5

3.2.1.1. Rating Scale Method

After hearing the sound sample, evaluations are conducted using the 1-10 rating scale reported in **Table 10,** which ranges from a rating of 1 (Little annoyance) to 10 (Extreme annoyance).

Table 10: 1-10 rating scale in terms of annoyance.

Verbal Descriptors	**Little Annoyance**	**Moderate Annoyance**	**High Annoyance**	**Very High Annoyance**	**Extreme Annoyance**
Category	1-2	3-4	5-6	7-8	9-10

In **Fig 87** a box plot provides a summary of the distribution of the ratings for each operating mode of the HVAC system, including interquartile values, median, mean, maximum and minimum evaluations. This visual aid is valuable for understanding the central tendency, spread, and overall shape of the dataset. It is important to note that the first outlier for the "BG" operating mode, with a rating of 2, does not qualify as a true outlier as it still corresponds to "Little Annoyance" category. Similarly, the rating of 8 in the "ES7" mode should not be considered an outlier, as it is only one category higher than the mean category.

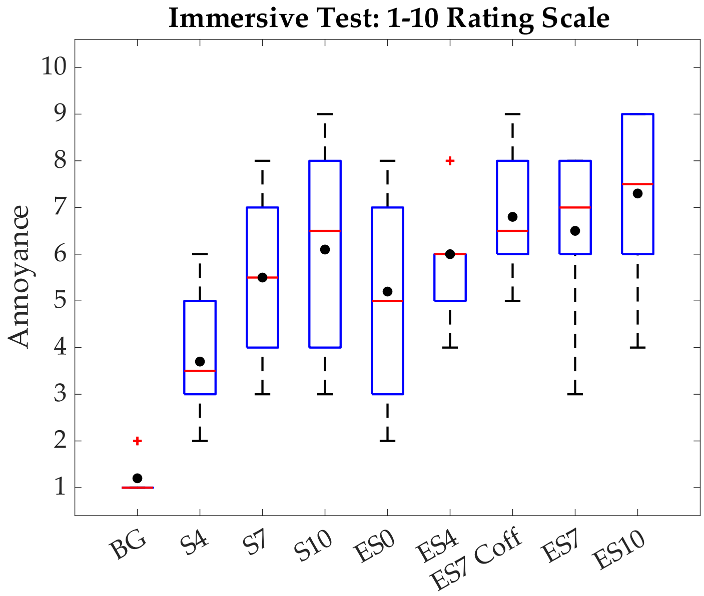

Fig 87: immersive test box plot for 1-10 rating scale; the lower and upper quartile ranges are visually depicted as the bottom and top edges of the blue box, respectively; red line: median; black dot: mean; black T-lines: lowest and highest ratings; '+' symbols: outliers in the data; The HVAC operating modes are shown on the abscissa (see Table 5).

3.2.1.2. Semantic Differential Method

This section analyses the results of the subjective evaluations using the Semantic Differential Method. This assessment method evaluates four scales of semantic adjective pairs reported in **Fig 88**, ranging from -3 to 3 (inclusive of 0). Annoyance is represented by the Not-Annoying/Annoying scale, Loudness by the Quiet/Loud scale, Sharpness by the Dull/Sharp scale, and Roughness by the Smooth/Rough scale.

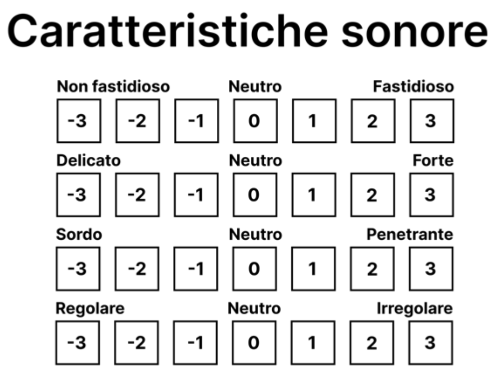

Fig 88: semantic adjective pairs and their relative assessment values; Not-Annoying(-3)/Annoying(3), Quiet(-3)/Loud(3), Dull(-3)/Sharp(3), Smooth(-3)/Rough(3).

Analysing all the ratings from the different semantic adjective pairs in **Fig 89-92**, we can see higher inter-quartile ranges compared to the 1-10 Annoyance Rating Scale, especially for the Annoying/Not-Annoying scale, which also presents some outliers. All the outliers in the Annoying and Loud scale are preserved as sound ratings could be influenced by the random playback order of sounds. For instance, if a very loud sound is played before a medium intensity sound, the evaluator may rate the medium sound lower.

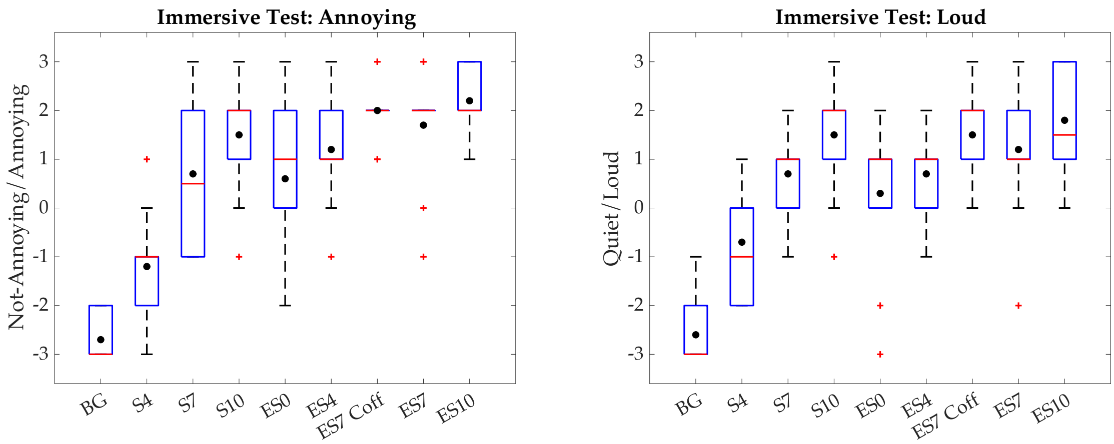

Fig 89-90: shows the box plot for the SDM scale related to annoyance (left) and loudness (right). The blue box represents the lower and upper quartile ranges, while the red line represents the median and the black dot represents the mean. The lowest and highest ratings are represented by the black T-lines, and any outliers in the data are represented by the '+' symbols. The HVAC operating modes are shown on the abscissa (see Table 5).

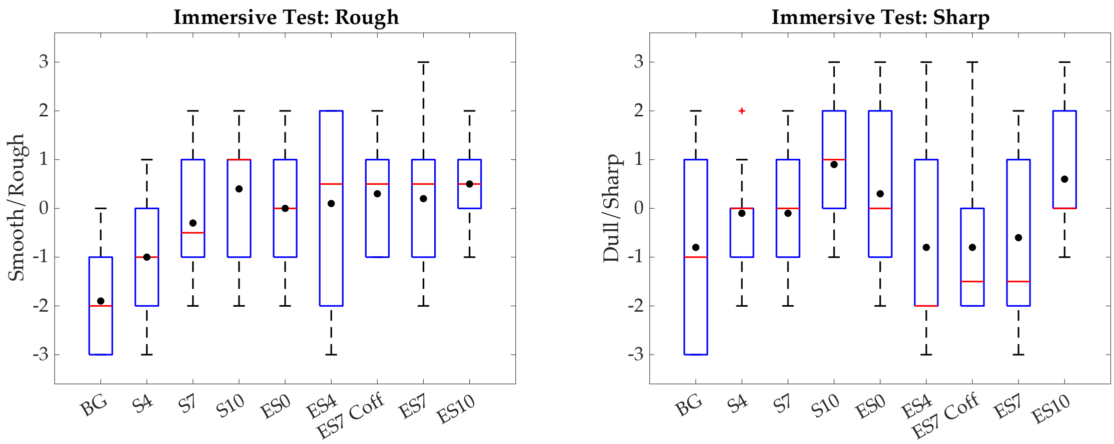

Fig 91-92 shows the box plot for the SDM scale related to roughness (left) and sharpness (right). The blue box represents the lower and upper quartile ranges, while the red line represents the median and the black dot represents the mean. The lowest and highest ratings are represented by the black T-lines, and any outliers in the data are represented by the '+' symbols. The HVAC operating modes are shown on the abscissa (see Table 5).

3.2.2. Binaural Listening Test

During the test, evaluators listened to binaural recordings made with the B&K 4101 and the Siemens ABH04 microphone headsets. The playback was listened with a pair of Sennheiser HD650 headphones, subjects rated the sounds using the 1-10 Annoyance Rating Scale and the Semantic Differential Method. In handling binaural recordings, psychoacoustic parameters are computed individually for each channel. Subsequently, the mean values of these parameters for the left and right channels are determined for each recording across various operational modes. This approach aligns with the guidelines outlined in ISO 15666 for comparative analysis. **Table 11** and **12** show the results obtained for each microphone type.

Table 11: mean subjective ratings listening Siemens ABH04 microphone recordings.

	Annoyance Rating Scale	SDM: Annoying	SDM: Loud	SDM: Sharp	SDM: Rough
BG	1.2	-2.6	-2.8	-2.1	-2
S4	4.5	0.7	-0.4	0.1	-0.3
S7	7.1	1.3	1.3	0.6	-0.3
S10	8.5	1.9	1.5	0.4	0

Table 12: mean subjective ratings listening B&K 4101 microphone recordings.

	Annoyance Rating Scale	SDM: Annoying	SDM: Loud	SDM: Sharp	SDM: Rough
BG	1.8	-2.2	-2.6	-1.1	-2.2
S4	6.0	0.1	-0.3	0.2	-0.3
S7	7.9	1.4	1.6	-0.4	-0.5
S10	9.4	2.5	2.6	0.4	0.2
ES0	6.7	1.2	0.8	-0.5	1.7
ES4	8.1	1.7	1.4	-0.2	1.0
ES7 Coff	9.2	2.5	2.2	0.2	0.3

3.2.2.1. Annoyance Rating Scale Results

The results presented in **Figure 93** indicate lower inter-quartile ranges compared with the immersive test, with some outliers observed in the 'ES0', 'S7', and 'S10' operating modes. These outliers will be preserved as sound ratings could be influenced by the random playback order of sounds.

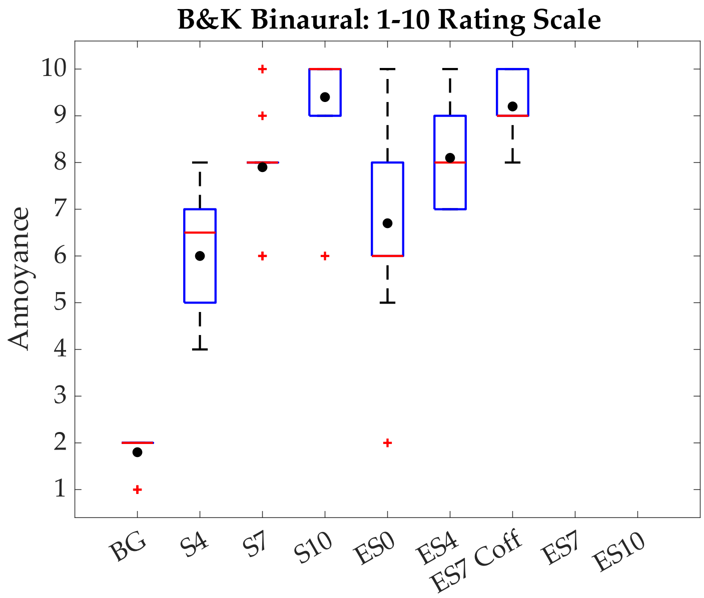

Fig 93: Annoyance Rating Scale Method with B&K 4101 recordings; the lower and upper quartile ranges are visually depicted as the bottom and top edges of the blue box, respectively; red line: median; black dot: mean; black T-lines: lowest and highest ratings; '+' symbols: outliers in the data; The HVAC operating modes are shown on the abscissa (see Table 5).

Concerning the Siemens ABH04 recordings, subjects provided more precise ratings of the sounds. Only two outliers were observed at 'S4' and 'BG', with the latter differing by only one category. Therefore, it's advisable not to consider them as significant outliers.

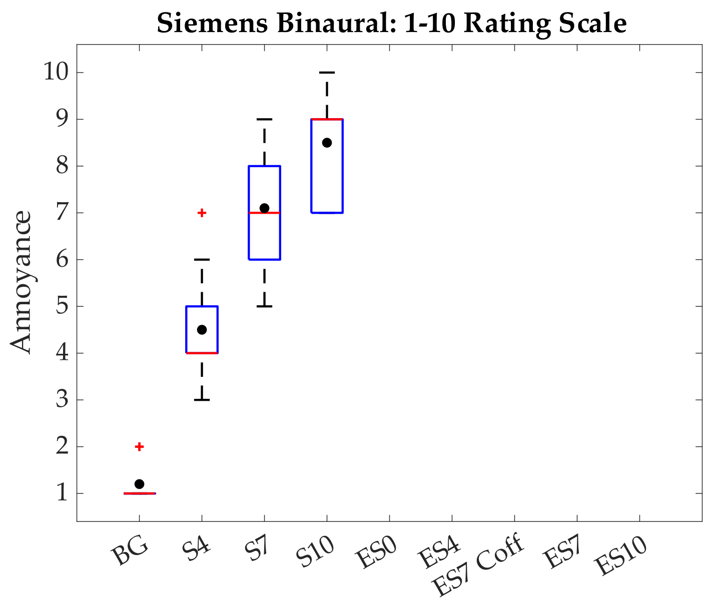

Fig 94: Annoyance Rating Scale Method with Siemens ABH04 recordings; the lower and upper quartile ranges are visually depicted as the bottom and top edges of the blue box, respectively; red line: median; black dot: mean; black T-lines: lowest and highest ratings; '+' symbols: outliers in the data; The HVAC operating modes are shown on the abscissa (see Table 5).

3.2.2.2. Semantic Differential Method Results

The results of the semantic differential method for the B&K 4101 microphone headset are presented in the following four **figures 95-98**. All outliers have been retained, as in previous sections, due to the potential influence of random playback order on judgement. The Quiet/Loud scale exhibits smaller inter-quartile ranges compared to the Not-Annoying/Annoying scale.

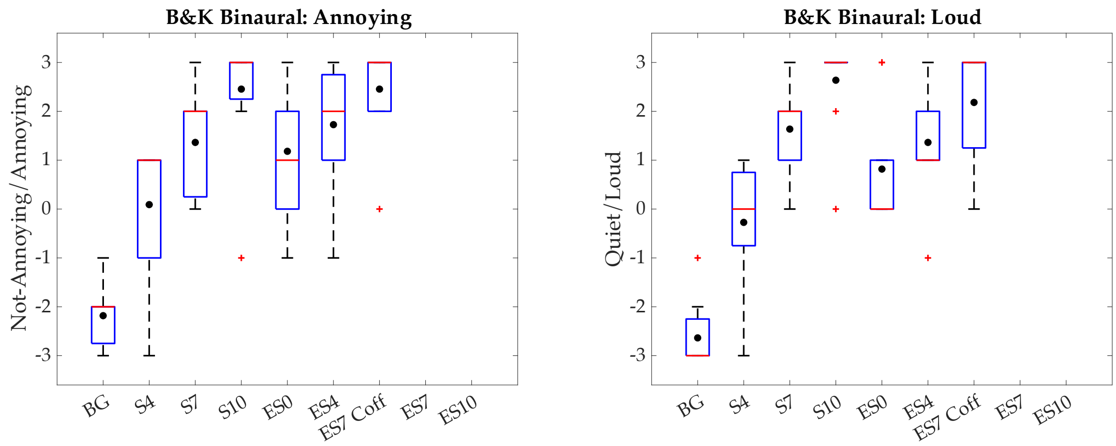

Fig 95-96: B&K 4101 microphone recordings, the box plot shows the SDM scale related to Annoyance (left) and Loudness (right). The blue box represents the lower and upper quartile ranges, while the red line represents the median and the black dot represents the mean. The lowest and highest ratings are represented by the black T-lines, and any outliers in the data are represented by the '+' symbols. The HVAC operating modes are shown on the abscissa (see Table 5).

The Rough/Smooth and Sharp/Dull scales in **Fig 97-98** exhibit larger interquartile ranges and wider ranges of maximum and minimum rating scores. This highlights the challenge of providing precise ratings for these two psychoacoustic parameters, reflecting a similar difficulty observed in the immersive test.

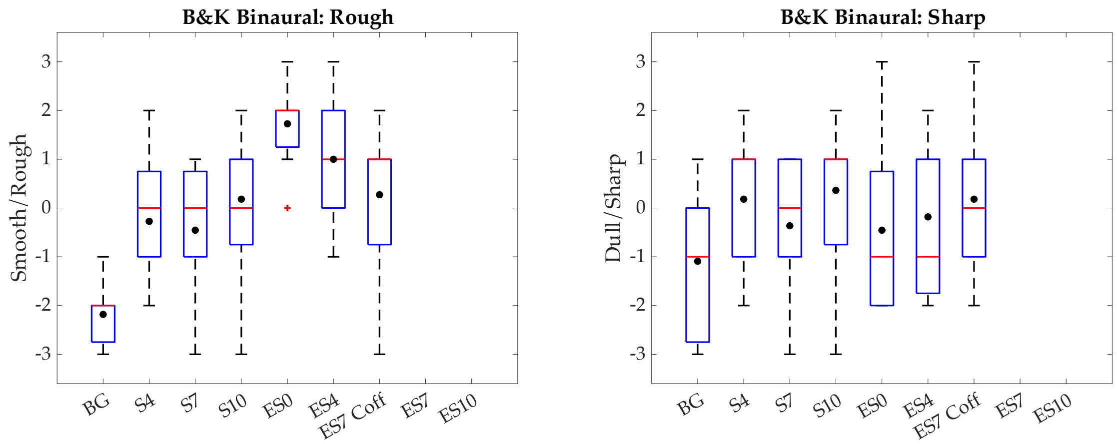

Fig 97-98: B&K 4101 microphone recordings, the box plot shows the SDM scale related to Roughness (left) and Sharpness (right). The blue box represents the lower and upper quartile ranges, while the red line represents the median and the black dot represents the mean. The lowest and highest ratings are represented by the black T-lines, and any outliers in the data are represented by the '+' symbols. The HVAC operating modes are shown on the abscissa (see Table 5).

Figures 99-100 show the results of the semantic differential rating for the Siemens ABH04 microphone headset. The inter-quartile values of the Not-Annoying/Annoying scale are thinner compared to the B&K 4101 recordings. The Smooth/Rough and Dull/Sharp scales present larger inter-quartile ranges.

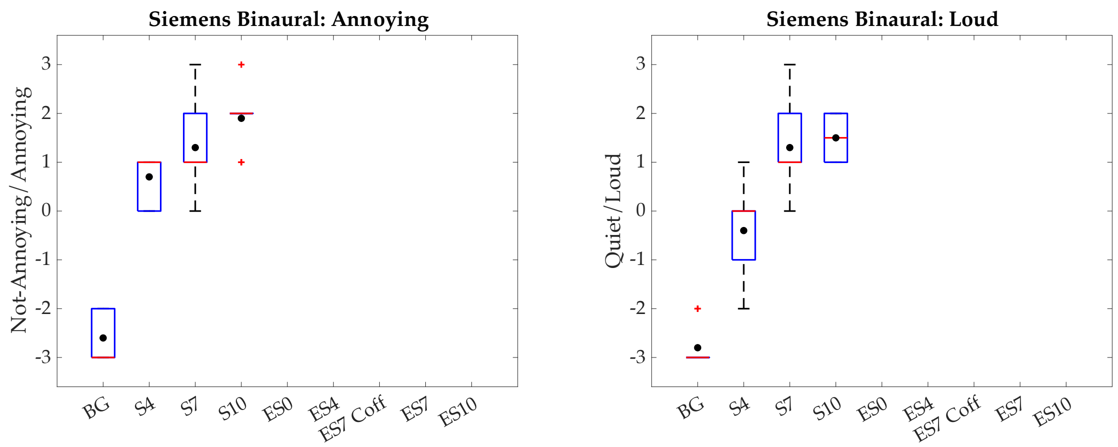

Fig 99-100: Siemens ABH04 microphone recordings, the box plot shows the SDM scale related to annoyance (left) and loudness (right). The blue box represents the lower and upper quartile ranges, while the red line represents the median and the black dot represents the mean. The lowest and highest ratings are represented by the black T-lines, and any outliers in the data are represented by the '+' symbols. The HVAC operating modes are shown on the abscissa (see Table 5).

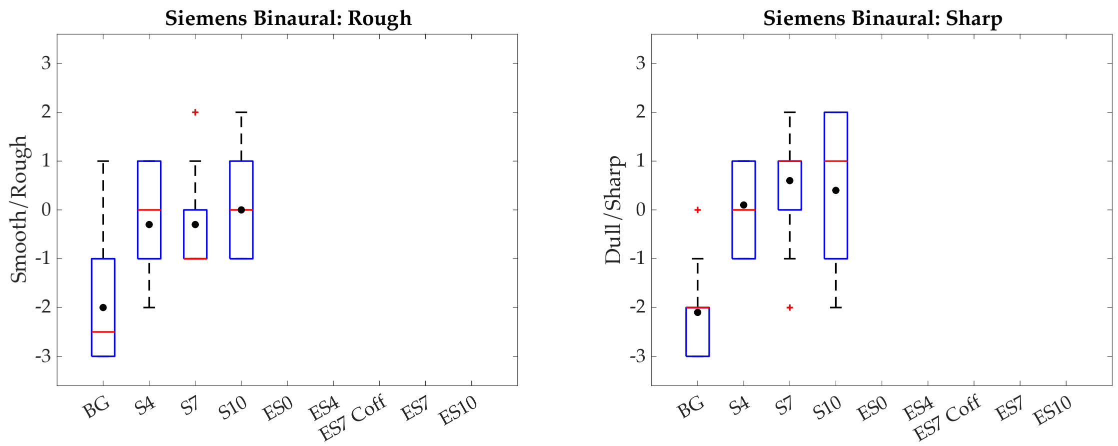

Fig 101-102: Siemens ABH04 microphone recordings, the box plot shows the SDM scale related to Roughness (left) and Sharpness (right). The blue box represents the lower and upper quartile ranges, while the red line represents the median and the black dot represents the mean. The lowest and highest ratings are represented by the black T-lines, and any outliers in the data are represented by the '+' symbols. The HVAC operating modes are shown on the abscissa (see Table 5).

3.2.3. Comparison

The subjective assessment gave higher evaluation of annoyance for the binaural listening test, **Fig 104**. Inter-quartile ranges are wider for the immersive test in **Fig 103** except for 'ES4'.

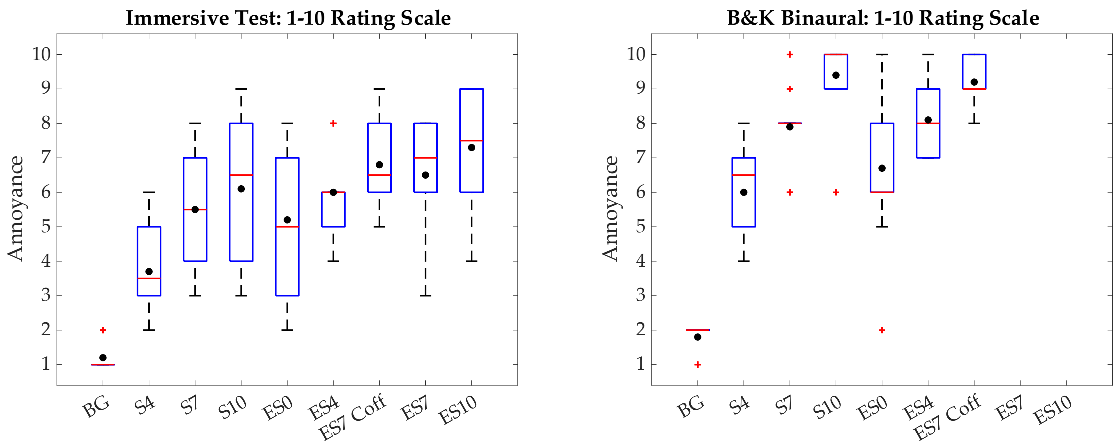

Fig 103-104: box plot comparison of the 1-10 annoyance rating scale for the immersive listening test (left) and the binaural listening test (right).

Inter-quartile ranges are similar for both the immersive test and the binaural listening test. The binaural test shows higher annoyance and loudness ratings.

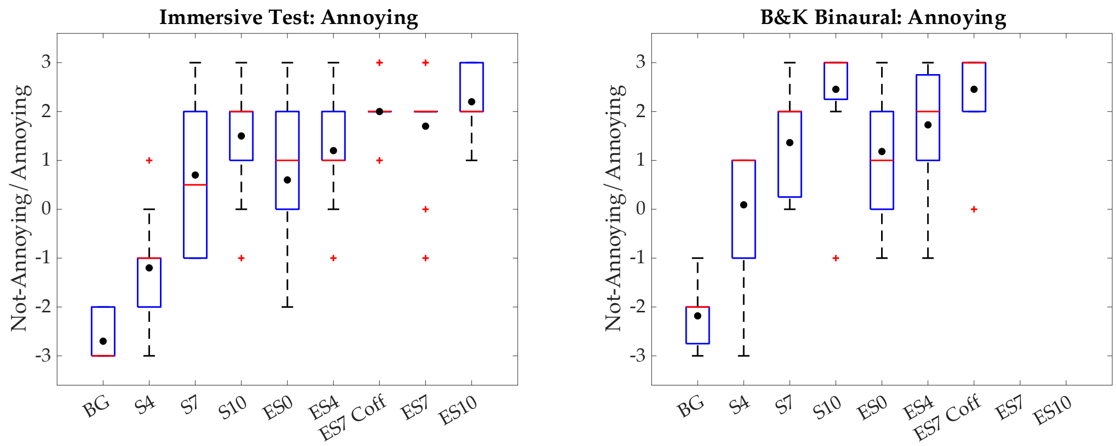

Fig 105-106: box plot comparison of the SDM Annoying scale for the immersive listening test (left) and the binaural listening test (right).

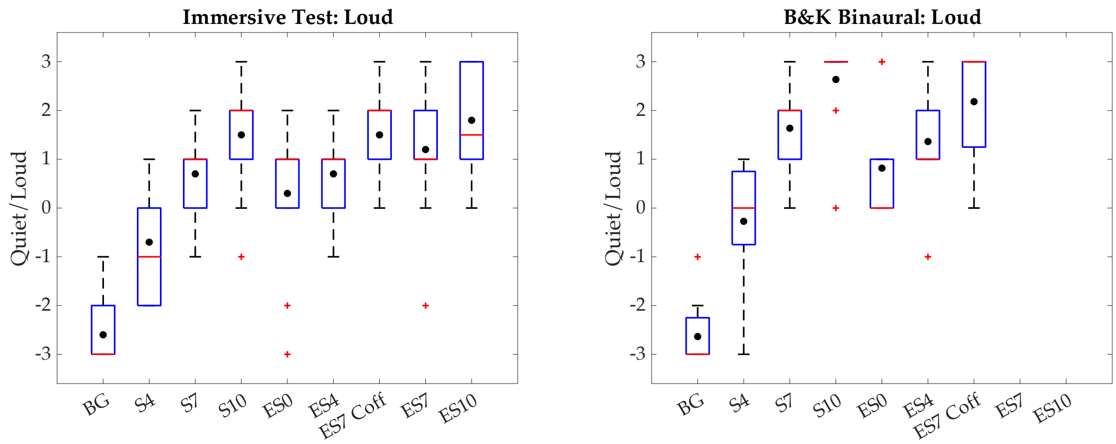

Fig 107-108: box plot comparison of the SDM Loud scale for the immersive listening test (left) and the binaural listening test (right).

For the Rough semantic scale, the results of the immersive test (**Fig x**) are the closest representation of the calculated roughness, while the binaural test in **Fig x** showed a non-linear tendency, from 'ES0' to 'ES7 Coff' the perceived roughness decreases slightly, whereas in the objective data it should increase.

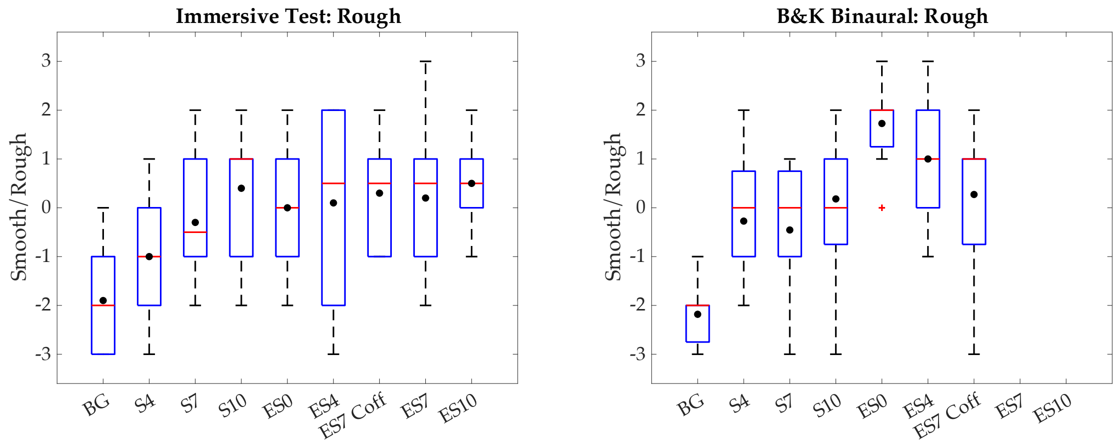

Fig 109-110: box plot comparison of the SDM Rough scale for the immersive listening test (left) and the binaural listening test (right).

Sharpness values show a non-linear tendency for both the immersive and binaural listening tests, with the binaural test appearing to give slightly more linear results than the immersive test.

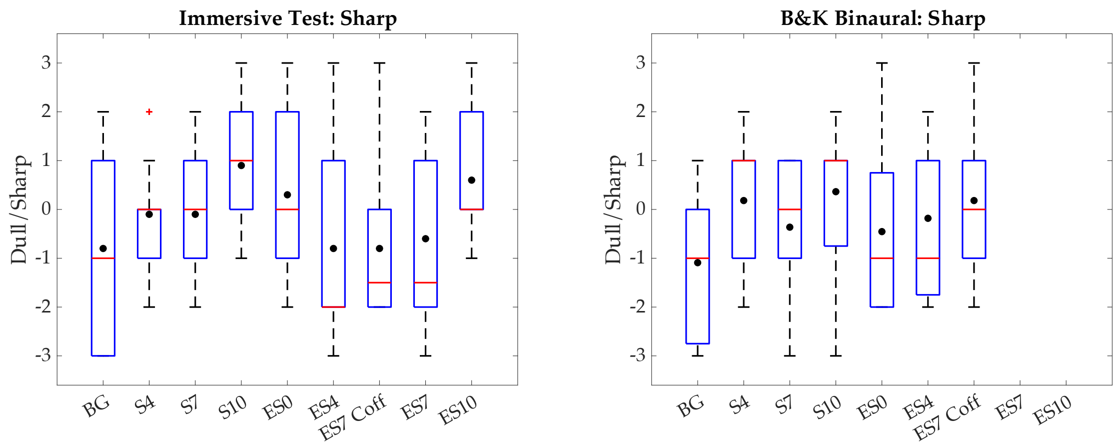

Fig 111-112: box plot comparison of the SDM Sharp scale for the immersive listening test (left) and the binaural listening test (right).

3.3. Prediction Model

Multiple linear regression analyses were conducted to explore the relationship between subjective assessments and psychoacoustic parameters for both the immersive test and the binaural listening test. Initially, a correlation analysis between subjective and objective data was conducted to identify the psychoacoustic parameters most relevant for predicting subjective ratings. Selection of independent variables for the prediction model was based on their correlation with subjective ratings and consideration of multicollinearity among the psychoacoustic parameters. High correlation between independent variables could compromise the reliability of the model. Therefore, variables with low multicollinearity were preferred. Additionally, significance tests including the F-test and t-test were performed to ensure the reliability of the prediction model.

3.3.1. Immersive Listening Test

Table 13 displays all the data utilised for the prediction model. The independent variables employed for multiple linear regression are the psychoacoustic parameters computed from the NTi omnidirectional microphone recorded samples. The subjective ratings will be regarded as dependent variables. If an off-diagonal element of P is smaller than the significance level (default is 0.05), then the corresponding correlation in R is considered significant.

3.3.1.1. Rating Scale Method

Based on the results of the subjective evaluation using the 1-10 Annoyance Rating Scale, the independent variables of the model were selected by examining correlation coefficients between the psychoacoustic parameters and the rating scale, as presented in **Table 14**. The analysis revealed a hierarchy of correlation strength, with Linear SPL (LZeq) demonstrating the highest correlation at 0.95, followed by LAeq at 0.94, and Loudness and Roughness at 0.93. In contrast, Sharpness exhibited the lowest correlation at 0.68. Sharpness was omitted from the model due to its low correlation with subjective annoyance. LAeq is not used due to its collinearity with LZeq. After some tests, the best prediction model for noise annoyance relied on LZeq and Loudness, resulting in an **adjusted R-square** value of 0.97 and a Root Mean Square Error (**RMSE**) of 0.35. Importantly, all p-values were below 0.05, indicating the statistical significance of the findings. The multiple linear regression equation (9) has been determined. **Figure 113** illustrates the subjective response of the test versus the predicted values of the equation. The results closely align with the red line, indicating that the subjective response is in line with the model response.

$$\text{Annoyance} = -4.4705 + 0.1043 \times LZeq + 0.0844 \times Loudness \quad (9)$$

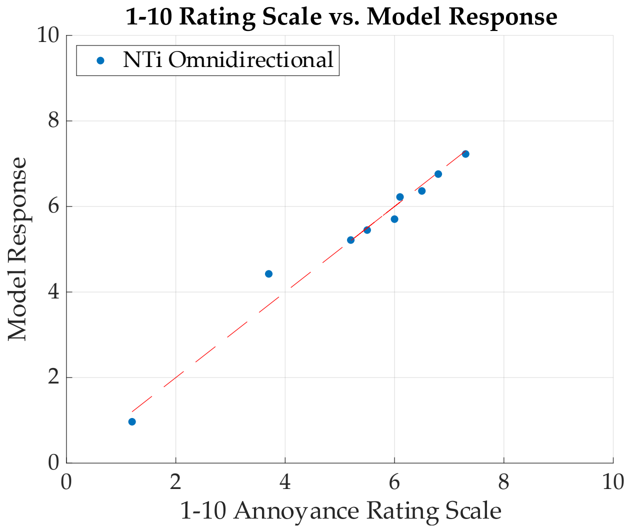

Fig 113: Comparison of subjective responses for the 1-10 Annoyance Rating Scale and the prediction model responses by Eq. (9). The straight red line indicates the cases where model responses are equal to subjective responses.

Table 13: Psychoacoustic parameters: LZeq, LAeq, Loudness, Sharpness; Subjective ratings: RS (1-10 Rating Scale Annoyance), Semantic Differential Method (SDM): Annoying, Loud, Sharp, Rough.

	LZeq	LAeq	Loudness	Sharpness	Roughness	1-10 Rating Scale	SDM: Annoying	SDM: Loud	SDM: Sharp	SDM: Rough
BG	51.0	30.7	1.4	0.90	0.07	1.2	-2.5	-2.6	-0.8	-1.9
S4	72.9	60.3	15.3	1.02	0.18	3.7	-1.2	-0.7	-0.1	-1
S7	77.3	65.5	22.0	1.13	0.21	5.5	0.7	0.7	-0.1	-0.3
S10	79.7	69.1	28.2	1.17	0.24	6.1	1.5	1.5	0.9	0.4
ES0	82.1	57.5	13.3	0.88	0.16	5.2	0.6	0.3	0.3	0
ES4	82.5	62.4	18.6	0.98	0.19	6	1.2	0.7	-0.8	0.1
ES7	83.8	66.8	24.8	1.11	0.23	6.5	1.7	1.2	-0.6	0.2
ES7 Coff	83.7	69.7	29.6	1.32	0.24	6.8	2	1.5	-0.8	0.3
ES10	84.8	71.6	33.8	1.18	0.27	7.3	2.2	1.8	0.6	0.5

Table 14: Correlation coefficients for the immersive listening test. Both objective and subjective data are reported. The 1-10 Rating Scale corresponds to the Annoyance Rating Scale, while SDM: Annoying, Loud, Sharp, Rough correspond to the Semantic Differential Method (SDM) scales. The bold numbers don't make the assumption (p-value<0.05).

	LZeq	LAeq	Loudness	Sharpness	Roughness	1-10 Rating Scale	SDM Annoying	SDM Loud	SDM Sharp	SDM Rough
LZeq	1.00	-	-	-	-	-	-	-	-	-
LAeq	0.93	1.00	-	-	-	-	-	-	-	-
Loudness	0.81	0.93	1.00	-	-	-	-	-	-	-
Sharpness	**0.51**	0.72	0.86	1.00	-	-	-	-	-	-
Roughness	0.85	0.97	0.99	0.80	1.00	-	-	-	-	-
1-10 Rating Scale	0.95	0.94	0.93	0.68	0.93	1.00	-	-	-	-
SDM Annoying	0.92	0.89	0.91	0.68	0.89	0.99	1.00	-	-	-
SDM Loud	0.93	0.95	0.95	0.72	0.95	0.99	0.98	1.00	-	-
SDM Sharp	**0.30**	**0.40**	**0.40**	**0.11**	**0.44**	**0.30**	**0.28**	**0.39**	1.00	-
SDM Rough	0.95	0.90	0.88	**0.59**	0.88	0.98	0.98	0.98	**0.40**	1.00

The equation was validated using the technique adopted by [10]. This involved correlating the predicted rating of the model with the median ratings of 5 random subjects from the assessment. The process was repeated three times, resulting in correlation values of 0.96, 0.94, and 0.94 ($p<0.05$).

3.3.1.2. Semantic Differential Method

Table 14 reports correlation coefficients between psychoacoustic parameters and the four scales of Semantic Differential Method. Considering the Annoying/Not-Annoying scale, Linear SPL (LZeq) has the highest correlation at 0.92, followed by Loudness at 0.91, A-SPL (LAeq) and Roughness at 0.89 and Sharpness at 0.68. When considering the Loud/Quiet scale, the correlation yields notably higher results. Specifically, Loudness, A-SPL, and Roughness demonstrate correlations of 0.95, followed by Loudness at 0.92, and Sharpness at 0.72. However, the Sharp/Dull scale exhibits the lowest correlation values, rendering it unsuitable for predicting sharp sounds in the model. Notably, Sharpness presents the lowest value at 0.11, which was anticipated to be the highest. Conversely, the Rough/Smooth scale performs more effectively for SPL than for Roughness, demonstrating a linear correlation of 0.88 for Roughness values.

The **Not-Annoying/Annoying** scale, reported with the mean evaluations for each operating mode is used as dependent variable, while the psychoacoustic parameters of LZeq, Loudness and Roughness were used as independent variables. The model has an **adjusted R-square** value of 0.95, with a **RMSE** of 0.36. **Figure 114** shows the subjective response plotted against the predicted values from the equation (10). The results closely follow the red line, indicating that the model's predictions closely match the subjective responses.

$$\text{Annoying} = -5.9915 + 0.1243 \times LZeq + 0.3222 \times Loudness - 48.5507 \times Roughness \qquad (10)$$

The equation was validated using the technique adopted by [10]. This involved correlating the predicted rating of the model with the median ratings of 5 random subjects from the assessment. The process was repeated three times, resulting in correlation values of 0.95, 0.95, and 0.93 ($p<0.05$).

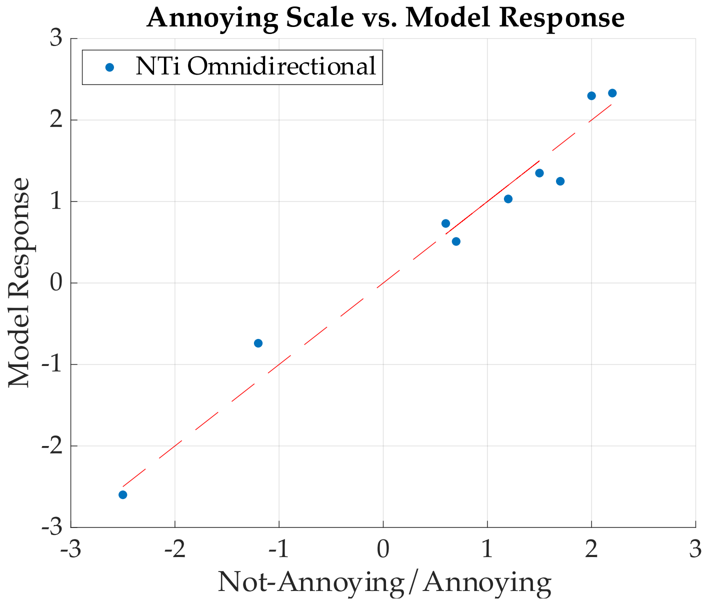

Fig 114: Comparison of subjective responses for the SDM Annoying/Not-Annoying scale and the prediction model responses by Eq. (10). The straight red line indicates the cases where model responses are equal to subjective responses.

The **Quiet/Loud** scale was employed as the dependent variable, and the psychoacoustic parameters of LZeq and Loudness served as independent variables for prediction. The model achieved an **adjusted R-square** of 0.97 and a root mean square error (**RMSE**) of 0.25. **Figure 115** shows the subjective response plotted against the predicted values from the equation (11). The results closely follow the red line, indicating that the subjective response closely matches the model's predictions.

$$\text{Loud} = -6.0031 + 0.063 \times LZeq + 0.0773 \times Loudness \qquad (11)$$

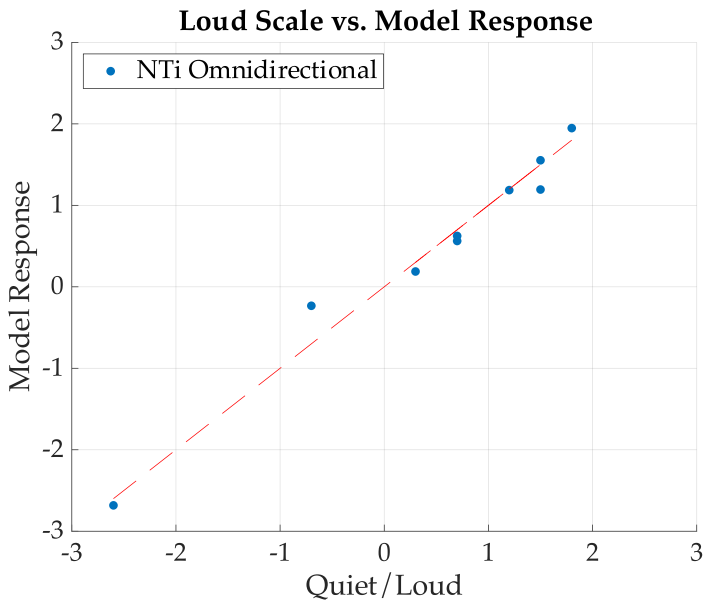

Fig 115: Comparison of subjective responses for the SDM Quiet/Loud scale and the prediction model responses by Eq. (11). The straight red line indicates the cases where model responses are equal to subjective responses.

The equation was validated using the technique adopted by [10]. This involved correlating the predicted rating of the model with the median ratings of 5 random subjects from the assessment. The process was repeated three times, resulting in correlation values of 0.92, 0.96, and 0.99 ($p<0.05$).
The **Rough/Smooth** scale was employed as the dependent variable, and the psychoacoustic parameters of Roughness and LZeq are used as independent variables for prediction. The model achieved an **adjusted R-square** of 0.88 and a root mean square error (**RMSE**) of 0.27. **Figure 116** shows the subjective response plotted against the predicted values from the equation (12).

$$\text{Rough} = -4.9950 + 3.5662 \times Roughness + 0.0529 \times LZeq \qquad (12)$$

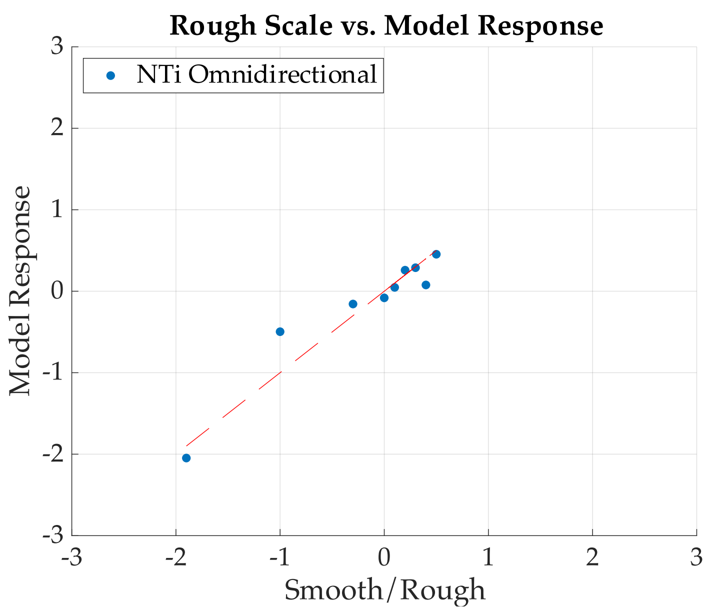

Fig 116: Comparison of subjective responses for the SDM Smooth/Rough scale and the prediction model responses by Eq. (12). The straight red line indicates the cases where model responses are equal to subjective responses.

The equation was validated using the technique adopted by [10]. This involved correlating the predicted rating of the model with the median ratings of 5 random subjects from the assessment. The process was repeated three times, resulting in correlation values of 0.67, 0.66, and 0.85 ($p<0.05$). The correlation coefficient of 0.66 indicates a p-value greater than 0.05, suggesting that the results may be unreliable.

3.3.2. Binaural Listening Test

Table 15 and **16** show the mean values of subjective assessment for each type of assessment method, along with the corresponding psychoacoustic parameters extracted from recordings captured by both B&K 4101 and Siemens ABH04 binaural microphone headsets. The psychoacoustic parameters are reported as the average between the left and right channels for each recording. The recordings obtained from both microphone setups will be used to develop the prediction model. This approach aims to increase the number of observations to enhance the reliability and robustness of the prediction.

Table 15: B&K 4101 microphone headset recordings; Psychoacoustic parameters: LZeq, LAeq, Loudness, Sharpness; Subjective ratings: RS (1-10 Rating Scale Annoyance), Semantic Differential Method (SDM): Annoying, Loud, Sharp, Rough.

	LZeq	**LAeq**	**Loudness**	**Roughness**	**Sharpness**	**RS**	**Annoying**	**Loud**	**Sharp**	**Rough**
BG	50.8	33.0	1.79	0.08	1.32	1.8	-2.2	-2.6	-1.1	-2.2
S4	71.1	63.1	18.57	0.19	1.21	6.0	0.1	-0.3	0.2	-0.3
S7	76.0	68.5	26.70	0.23	1.28	7.9	1.4	1.6	-0.4	-0.5
S10	79.3	72.2	34.14	0.27	1.33	9.4	2.5	2.6	0.4	0.2
ES0	80.6	60.4	16.20	0.19	1.10	6.7	1.2	0.8	-0.5	1.7
ES4	83.8	65.7	23.53	0.21	1.22	8.1	1.7	1.4	-0.2	1.0
ES7 Coff	83.5	69.5	29.23	0.23	1.26	9.2	2.5	2.2	0.2	0.3

Table 16: Siemens ABH04 microphone headset recordings; Psychoacoustic parameters: LZeq, LAeq, Loudness, Sharpness; Subjective ratings: RS (1-10 Rating Scale Annoyance), Semantic Differential Method (SDM): Annoying, Loud, Sharp, Rough.

	LZeq	**LAeq**	**Loudness**	**Roughness**	**Sharpness**	**RS**	**Annoying**	**Loud**	**Sharp**	**Rough**
BG	53.53	32.07	1.62	0.08	1.02	1.2	-2.6	-2.8	-2.1	-2
S4	70.97	62.29	16.71	0.19	1.04	4.5	0.7	-0.4	0.1	-0.3
S7	76.05	67.67	24.21	0.24	1.14	7.1	1.3	1.3	0.6	-0.3
S10	78.74	70.82	30.25	0.25	1.19	8.5	1.9	1.5	0.4	0

Table 17: Correlation coefficients for the immersive listening test. Both objective and subjective data are reported. The 1-10 Rating Scale corresponds to the Annoyance Rating Scale, while SDM: Annoying, Loud, Sharp, Rough correspond to the Semantic Differential Method (SDM) scales. The bold numbers don't make the assumption (p-value<0.05).

	LZeq	**LAeq**	**Loudness**	**Sharpness**	**Roughness**	**1-10 Rating Scale**	**SDM Annoying**	**SDM Loud**	**SDM Sharp**	**SDM Rough**
LZeq	1.00	-	-	-	-	-	-	-	-	-
LAeq	0.93	1.00	-	-	-	-	-	-	-	-
Loudness	0.88	0.96	1.00	-	-	-	-	-	-	-
Sharpness	**0.18**	**0.26**	**0.42**	1.00	-	-	-	-	-	-
Roughness	0.90	0.99	0.98	**0.31**	1.00	-	-	-	-	-
1-10 Rating Scale	0.94	0.94	0.97	**0.45**	0.95	1.00	-	-	-	-
SDM Annoying	0.96	0.96	0.96	**0.32**	0.96	0.97	1.00	-	-	-
SDM Loud	0.94	0.94	0.97	**0.40**	0.96	0.99	0.98	1.00	-	-
SDM Sharp	0.76	0.90	0.83	**0.30**	0.88	0.79	0.84	0.79	1.00	-
SDM Rough	0.90	0.74	0.62	**0.00**	0.68	0.75	0.79	0.74	**0.58**	1.00

3.3.2.1. Rating Scale Method

A correlation analysis was conducted in **Table 17** to select the psychoacoustic parameters most correlated with the 1-10 rating scale assessment results. The analysis also assessed for multicollinearity among the psychoacoustic parameters. Subjective ratings show the highest linearity match with Loudness at 0.97, followed by Roughness at 0.95, LZeq and LAeq at 0.94 and Sharpness at 0.45.

When compared to the immersive test, linear regression yields superior outcomes in binaural listening. The dependent variable is the mean subjective annoyance assessed using the 1-10 **Annoyance Rating Scale**, while the independent variables remain consistent with those of the immersive test: Linear SPL (LZeq) and Loudness show the best results. The model predicts noise annoyance with an **adjusted R-square** of 0.97 and a **RMSE** of 0.49. **Figure 117** shows the subjective response plotted against the predicted values from the equation (13). Subjective responses closely align with the prediction model responses, as indicated by the proximity to the red line in the plot. This suggests that the prediction model effectively captures the relationship between the independent variables and the subjective assessments.

$$Annoyance = -3.9563 + 0.0956 \times LZeq + 0.1660 \times Loudness \qquad (13)$$

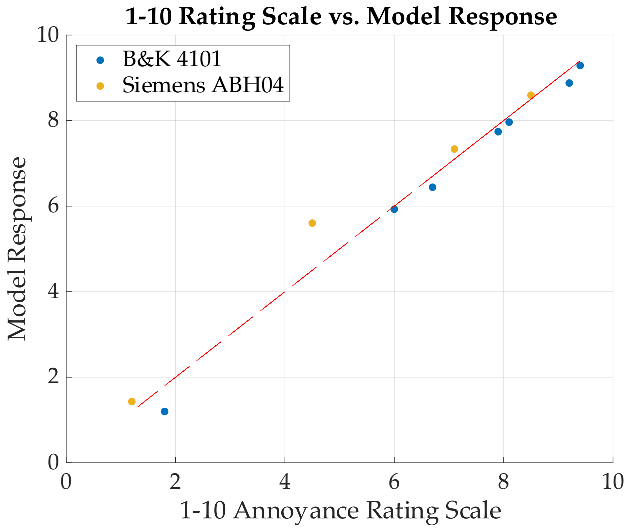

Fig 117: Comparison of subjective responses for the Annoyance Rating Scale and the prediction model responses by Eq. (13). The straight red line indicates the cases where model responses are equal to subjective responses.

The equation was validated using the technique adopted by [10]. This involved correlating the predicted rating of the model with the median ratings of 5 random subjects from the assessment. The process was repeated three times, resulting in correlation values of 0.92, 0.96, and 0.99 ($p<0.05$).

3.3.2.2. Semantic Differential Method

Table 17 presents the correlation results between the psychoacoustic parameters of the binaural recordings and the four scales of the Semantic Differential Method. The Annoying/Not Annoying scale exhibits high correlation with LZeq, LAeq, Loudness, and Roughness at 0.96, while demonstrating low correlation with Sharpness at 0.32. The Loud/Quiet scale demonstrates high correlation with Loudness at 0.97, followed by Roughness at 0.96, LZeq and LAeq at 0.94, and Sharpness at 0.40. However, the Sharp/Dull scale shows low correlation with Sharpness and appears to be better correlated with LAeq and Roughness at 0.90 and 0.88, respectively. Moreover, the Rough/Smooth scale exhibits stronger correlation with LZeq at 0.90 than with Roughness at 0.68.

The **Annoying/Not-Annoying** scale of Semantic Differential Method was predicted through the use of Loudness and LZeq as independent variables, the results show an **adjusted R-squared** of 0.97 with a **RMSE** of 0.27. **Figure 118** shows the subjective response plotted against the predicted values from the equation (14). Subjective responses closely align with the prediction model responses, as indicated by the proximity to the red line in the plot.

$$Annoying = -6.3205 + 0.07582 \times LZeq + 0.0768 \times Loudness \qquad (14)$$

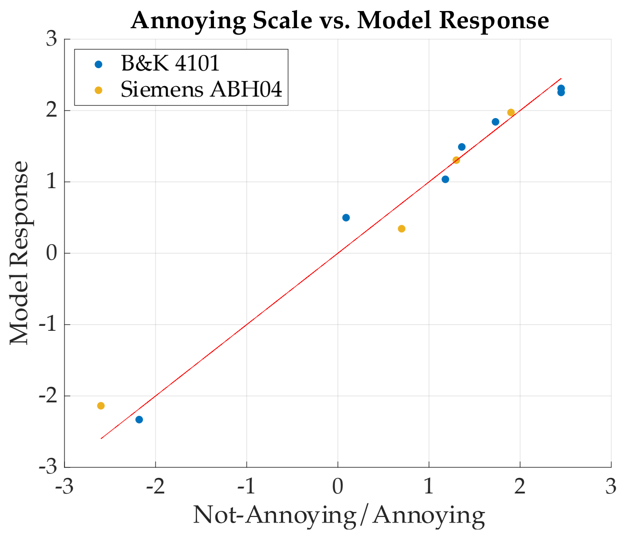

Fig 118: Comparison of subjective responses for the Not-Annoying/Annoying semantic scale and the prediction model responses by Eq. (14). The straight red line indicates the cases where model responses are equal to subjective responses.

The equation was validated using the technique adopted by [10]. This involved correlating the predicted rating of the model with the median ratings of 5 random subjects from the assessment. The process was repeated three times, resulting in correlation values of 0.98, 0.94, and 0.94 ($p<0.05$).

The **Quiet/Loud** semantic scale was predicted using LZeq and Loudness as independent variables, the **adjusted R-square** is 0.97, with a **RMSE** of 0.33. These parameters give a similar result compared to the Annoying scale, with a little difference in terms of root mean square error. **Figure 119** shows the subjective response plotted against the predicted values from the equation (15). Subjective responses align quite closely with the prediction model responses, as evidenced by their proximity to the red line in the plot. The equation was validated using the technique adopted by [10]. This involved correlating the predicted rating of the model with the median ratings of 5 random subjects from the assessment. The process was repeated three times, resulting in correlation values of 0.96, 0.94, and 0.94 ($p<0.05$).

$$Loud = -6.1938 + 0.0613 \times LZeq + 0.1084 \times Loudness \qquad (15)$$

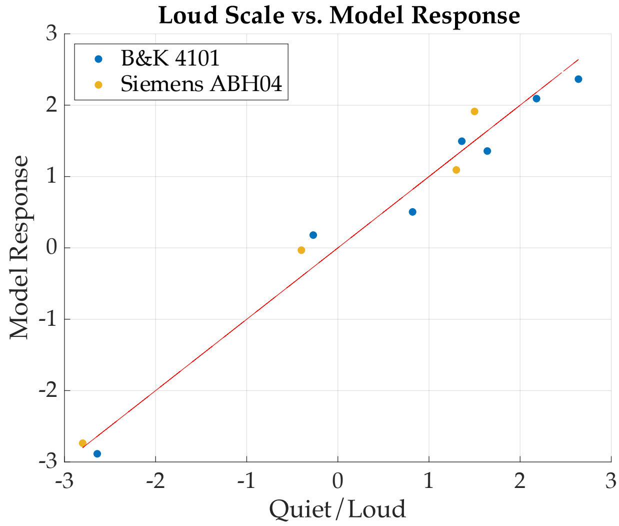

Fig 119: Comparison of subjective responses for the Quiet/Loud scale and the prediction of these responses by Eq. (15). The straight red line indicates the cases where model responses are equal to subjective responses.

The equation was validated using the technique adopted by [10]. This involved correlating the predicted rating of the model with the median ratings of 5 random subjects from the assessment. The process was repeated three times, resulting in correlation values of 0.97, 0.96, and 0.94 ($p<0.05$).
The **Rough/Smooth** semantic scale was predicted using LZeq and Roughness as independent variables, the **adjusted R-square** is 0.88, with a **RMSE** of 0.38. **Figure 120** shows the subjective response plotted against the predicted values from the equation (16). The RMSE of 0.38 is evident in the plot, as indicated by the larger distance of the points from the red line.

$$Rough = -9.1139 + 0.1548 \times LZeq - 12.3291 \times Roughness \qquad (16)$$

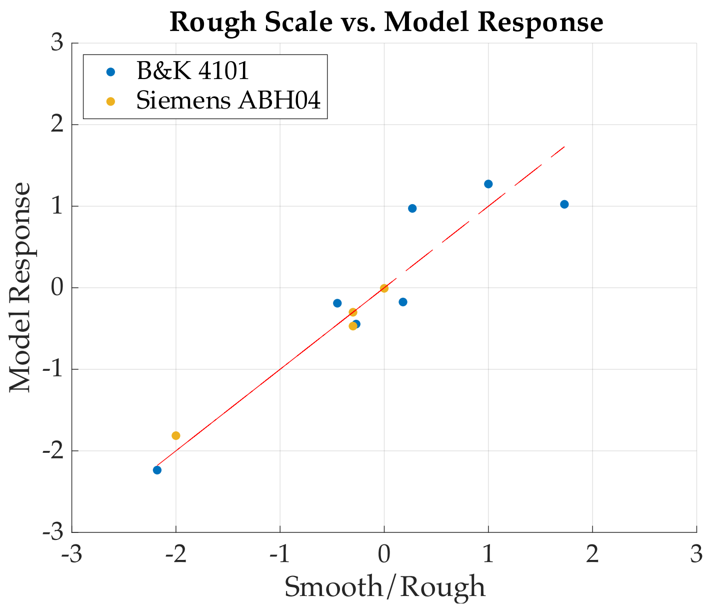

Fig 120: Comparison of subjective responses for the Rough/Smooth scale and the prediction of these responses by Eq. (16). The straight red line indicates the cases where model responses are equal to subjective responses.
The equation was validated using the technique adopted by [10]. This involved correlating the predicted rating of the model with the median ratings of 5 random subjects from the assessment. The process was repeated three times, resulting in correlation values of 0.88, 0.93, and 0.82 ($p<0.05$).

4. Discussion

This section analyses the ability of the Immersive Test and the Binaural Test to predict sound quality across various metrics. All R-squared values of multiple linear regressions were higher than 0.88, with a maximum of 0.97, indicating strong predictive capability across the evaluated perceptual dimensions. The results presented in **Table 18-21** suggest that the Immersive Test has slightly better predictive performance than the Binaural Test, especially in estimating Annoyance, Loudness, and Roughness. The Immersive Test demonstrates lower Root Mean Square Error (RMSE) values across multiple perceptual dimensions, including Annoyance (R-squared 0.97, RMSE 0.35), Loudness (R-squared 0.97, RMSE 0.25), and roughness (R-squared 0.88, RMSE 0.38). This suggests that the Immersive Test may provide a more reliable estimation of sound perception, particularly in environments where roughness plays a significant role. The Immersive Test has a notable advantage in accurately reproducing low frequencies, making it particularly relevant in scenarios involving low-frequency noise sources such as tractor engines. However, neither test was effective in predicting sharpness, which is an important parameter in the field of HVAC noise and highlights a current limitation in sound perception assessment methodologies. Further research is needed to develop more effective training and assessment methods for jury testing. Both tests achieved high R-squared values for annoyance prediction, but the Binaural Test resulted in higher annoyance and loudness ratings according to the Binaural Listening Test. This suggests potential biases or differences in perception between the two test methodologies.

The analysis of the multiple linear regression equations for each test indicates that linear sound pressure level (SPL) and loudness are the most effective parameters for predicting annoyance in this study. The use of A-weighted sound pressure level (SPL) with loudness resulted in lower prediction accuracy. Additionally, linear SPL is crucial in predicting roughness. Reducing these parameters will enhance the HVAC sound quality inside the cabin.

Our prediction models outperformed two out of three previous studies that used similar methodologies, particularly in predicting noise annoyance and roughness. This suggests that our approach with spatialized audio offers advancements in sound perception assessment. However, to further increase prediction reliability, future research should consider testing a higher number of HVAC operating modes and conducting in-field tests with tractors moving under different scenarios. The recording scenarios may have new sound characteristics, such as tonality, fluctuation strength, prominence ratio, which were not considered in this test due to their low measured values. These parameters could provide additional insights into sound annoyance.

Table 18: annoyance linear regression using the 1-10 rating scale, immersive listening test.

Rating Scale	Regression Equation	R-squared	RMSE
Annoyance	$\text{Annoyance} = -4.4705 + 0.1043 \times LZeq + 0.0844 \times Loudness$	0.97	0.35

Table 21: multiple linear regressions using semantic differential method scales, immersive listening test.

SDM	Regression Equation	R-squared	RMSE
Annoying	$\text{Ann} = -5.9915 + 0.1243 \times LZeq + 0.3222 \times Loudness - 48.5507 \times Roughness$	0.95	0.36
Loud	$\text{Loud} = -6.0031 + 0.063 \times L\text{Zeq} + 0.0773 \times L\text{oudness}$	0.97	0.25
Rough	$\text{Rough} = -4.9950 + 3.5662 \times Roughness + 0.0529 \times LZeq$	0.88	0.27
Sharp	Not predictable	-	-

Table 20: annoyance linear regression using the 1-10 rating scale, binaural listening test.

Rating Scale	Regression Equation	R-squared	RMSE
Annoyance	$Annoyance = -3.9563 + 0.0956 \times LZeq + 0.1660 \times Loudness$	0.97	0.49

Table 21: multiple linear regressions using semantic differential method scales, binaural listening test.

SDM	Regression Equation	R-squared	RMSE
Annoying	$Ann = -6.32046 + 0.0758 \times LZeq + 0.0765 \times Loudness$	0.97	0.27
Loud	$Loud = -6.1938 + 0.0613 \times LZeq + 0.1084 \times Loudness$	0.97	0.33
Rough	$Rough = -9.1139 + 0.1548 \times LZeq - 12.3291 \times Roughness$	0.88	0.38
Sharp	Not predictable	-	-

5. Conclusions

An extensive review of the current literature regarding HVAC noise within tractor cabins was undertaken. This will serve as a baseline in order to identify the most significant attributes that characterize HVAC noise perception and the prevailing techniques in sound quality engineering used to evaluate this type of noise for tractors. The primary psychoacoustic parameters identified as significant contributors to HVAC noise annoyance and sound quality include Linear Sound Pressure Level (SPL), A-weighted Sound Pressure Level (A-SPL), Loudness, Roughness, and Sharpness. The thesis explores two subjective evaluation methods, including the 1-10 Rating Scale Method and the Semantic Differential Method. These were tested through binaural listening and a novel playback method: the immersive listening test with spatialized audio of 3rd order ambisonics. Through objective and subjective data, a predictive model was developed for Annoyance, Loudness, and Roughness using multiple linear regression. The study aimed to compare the effectiveness of the Immersive Test against the Binaural Test in forecasting sound perception across various metrics. Our findings suggest that the Immersive Test exhibits slightly superior predictive performance compared to the Binaural Test, particularly in estimating Annoyance, Loudness, and Roughness, while Sharpness was not predictable. The best performing models predicted Annoyance with a R-squared of 0.97 and a RMSE of 0.35, Loudness with a R-squared of 0.97 and RMSE of 0.25, Roughness with and R-square of 0.88 and a RMSE of 0.27.

6. Acknowledgements

I want to thank HEAD Acoustics for providing access to the ArtemiS suite software, which made analysing and testing my research data much easier. Special thanks to Professor Arianna Astolfi for her guidance and support, and to Louena Shtrepi for her contributions. And, of course, my family deserves a big thank you for their constant love and encouragement throughout this journey. Couldn't have done it without them!

7. References

1. Zhipeng Wang, Yanyan Zuo, Liming Sun, The Impact of Sound Pressure Level, Loudness, Roughness, Sharpness, Articulation Index, Hand Vibration, and Seat Vibration on Subjective Comfort Perception of Tractor Drivers, Symmetry, (2023)
2. Liqiang Yanga, Pan Wanga, Jie Wang, Research on evaluation model for vehicle interior sound quality based on an optimized BiLSTM using genetic algorithm, Mechanical Systems and Signal Processing, (2023).
3. Wang Y.; Zhang S.; Meng D.; Zhang L. Nonlinear overall annoyance level modeling and interior sound quality prediction for pure electric vehicle with extreme gradient boosting algorithm, Applied Acoustics, (2022)
4. Chen P.; Xu L.; Tang Q.; Shang L.; Liu W. Research on prediction model of tractor sound quality based on genetic algorithm, Applied Acoustics, (2022)
5. Huang H.; Wu J.; Lim T.C.; Yang M.; Ding W. Pure electric vehicle nonstationary interior sound quality prediction based on deep CNNs with an adaptable learning rate tree, Mechanical Systems and Signal Processing, (2021)
6. Huang X.; Huang H.; Wu J.; Yang M.; Ding W. Sound quality prediction and improving of vehicle interior noise based on deep convolutional neural networks, Expert Systems with Applications, (2020)
7. Qian K.; Hou Z. Intelligent evaluation of the interior sound quality of electric vehicles, Applied Acoustics, (2021)
8. Huang H.B.; Wu J.H.; Huang X.R.; Yang M.L.; Ding W.P. The development of a deep neural network and its application to evaluating the interior sound quality of pure electric vehicles, Mechanical Systems and Signal Processing, (2019)
9. Ma C.; Chen C.; Liu Q.; Gao H.; Li Q.; Gao H.; Shen Y. Sound Quality Evaluation of the Interior Noise of Pure Electric Vehicle Based on Neural Network Model, IEEE Transactions on Industrial Electronics, (2017)
10. Li D.; Huang Y. The discomfort model of the micro commercial vehicles interior noise based on the sound quality analyses, Applied Acoustics, (2018)
11. Wang Y.S.; Shen G.Q.; Xing Y.F. A sound quality model for objective synthesis evaluation of vehicle interior noise based on artificial neural network, Mechanical Systems and Signal Processing, (2014)
12. Zhang E.; Hou L.; Shen C.; Shi Y.; Zhang Y. Sound quality prediction of vehicle interior noise and mathematical modeling using a back propagation neural network (BPNN) based on particle swarm optimization (PSO), Measurement Science and Technology, (2015)
13. Huang H.B.; Huang X.R.; Li R.X.; Lim T.C.; Ding W.P. Sound quality prediction of vehicle interior noise using deep belief networks, Applied Acoustics, (2016)
14. Gao Y.-H.; Qian K.; Liang J.; Liu Q.; Zhao J. 1915. Interior sound quality evaluation model of heavy commercial vehicles, Journal of Vibroengineering, (2016)
15. Jeong J.-E.; Yang I.-H.; Bin Abu A.; Cha K.-J.; Oh J.-E. Development of a new sound quality metric for evaluation of the interior noise in a passenger car using the logarithmic Mahalanobis distance, Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, (2013)
16. Chen S.; Wang D.; Liang J.; McDonnell M. Sound quality analysis and prediction of vehicle interior noise based on grey system theory, Fluctuation and Noise Letters, (2012)
17. Yoon J.-H.; Yang I.-H.; Jeong J.-E.; Park S.-G.; Oh J.-E. Reliability improvement of a sound quality index for a vehicle HVAC system using a regression and neural network model, Applied Acoustics, (2012)
18. Wang Y.; Shen G.; Guo H. Modelling for sound annoyance evaluation of vehicle noise based on neural network, Przeglad Elektrotechniczny, (2012)

19. Leite R.P.; Paul S.; Gerges S.N.Y. A sound quality-based investigation of the HVAC system noise of an automobile model, Applied Acoustics, (2009)
20. Back J.; Lee S.-K.; Min Lee S.; An K.; Kwon D.-H.; Park D.-C. Design and implementation of comfort-quality HVAC sound inside a vehicle cabin, Applied Acoustics, (2021)
21. PD ISO/TS 15666:2021
22. J1441_201609, Subjective Rating Scale for Vehicle Ride and Handling, SAE, (2016)
23. Norm O.; Scott A; Chris E.; Scott L. Guidelines for Jury Evaluation of Automotive Sounds, SAE, (1999)
24. Hugo Fastl; Eberhard Zwicker Psychoacoustics, Facts and Models, Springer, (2006)
25. Pedrielli F.; Carletti E.; Just noticeable differences of loudness and sharpness for earth moving machines. The Journal of the Acoustical Society of America, (2008)
26. Norm O.; Scott A.; Chris E.; Scott L.; Guidelines for Jury Evaluations of Automotive Sounds, SAE Technical Papers, (1999)
27. D. Dal Palù; E. Buiatti, G. E. Puglisi, O. Houix; P. Susini; C. De Giorgi; A. Astolfi; The use of semantic differential scales in listening tests: A comparison between context and laboratory test conditions for the rolling sounds of office chairs, Applied Acoustics, 2017
28. German patent DE 3709397C2, (1987)
29. Yu. H.; Dou L.; An empirical category-ratio scale for evaluating the subjective intensity of noise based on the comparison of estimated magnitudes and categories, Applied Acoustics, (2020)
30. Jun Z.; Kean C.; Hao Li; Xingshu C.; Ningjuan D.; The effects of rating scales and individual characteristics on perceived annoyance in laboratory listening tests, Applied Acoustics, (2023)