Neural Network Models For The Subjective And Objective Assessment Of A Propeller Aircraft Interior Sound Quality

Lopes, Bernardo[1] ; Colangeli, Claudio; Janssens, Karl; Mroz, Agnieszka and Van der Auweraer, Herman
Siemens Industry Software NV
Interleuvenlaan, 68 – 3001 Leuven (Belgium)

ABSTRACT

This paper reports on the use of neural networks for modelling the relation between the salient objective and subjective psychoacoustic attributes of a propeller aircraft interior sound. The developed model grounds on a modular approach consisting in a series of two stages. The first stage is devoted to the data-driven estimation of sound quality features (loudness, sharpness, etc.) in time domain. In the second stage the estimated sound quality attributes are adopted to classify the input sounds in terms of passenger annoyance. This second module consists in an Artificial Neural Network model, trained on the basis of a subjective evaluation test. The paper describes the approach followed for the neural networks definition and for the collection of the subjective and objective propeller aircraft in-cabin psychoacoustic attributes. The adopted model has been compared with alternative machine learning instruments. We finally assessed the accuracy of the model in predicting the passenger response by validating it on experimental propeller aircraft in-cabin noise recordings whose annoyance was evaluated by a pool of jurors in a subjective evaluation test. Such a tool, integrated in a virtual prototyping framework, paves the way for the inclusion of the human perception in the aircraft design optimization process.

Keywords: Psychoacoustics, Neural Networks, Annoyance.
I-INCE Classification of Subject Number: 79

1. Introduction

The assessment of a propeller aircraft acoustic discomfort often occurs in the late stages of its development cycle. This makes it difficult to intervene to improve the resulting Sound Quality characteristics because many design parameters have already been fixed. Therefore, very often, the interior noise of an aircraft is only optimized regarding the Sound Pressure Level reduction the passenger acoustic discomfort, i.e., annoyance, is disregarded. The development of a prediction model able to, directly from pressure signal, quantify the human perception of sound enables the inclusion of subjective acoustic features in the early stages of the aircraft design process.

---

[1] bernardo.lopes@siemens.com

2. Sound Quality Prediction in a Propeller Aircraft

2.1 Overview

This paper analyses the prediction of Sound Quality characteristics in a propeller aircraft. Two flying conditions for this aircraft were held into account. The case when both propellers have a control system which tries to ensure coincident rotational frequencies is denominated by *synchronous* and the *asynchronous* case where no such control is active. Interior acoustic measurements have been carried out in the studied propeller aircraft during the cruising phase. Both for *synchronous* and *asynchronous* conditions, the noise was recorded in numbered positions of the aircraft, i.e. seats (1). In previous works, using these recordings, an algorithm was developed to, from a virtual model of the aircraft with changeable parameters, synthesize the sound sample corresponding to any position in the interior of the cabin, with the possibility of changing several design parameters of the virtual model. Therefore, it is possible to reproduce and study the interior noise in each typical propeller aircraft, without having to re-record sound samples in a flight. Consequently, the sound samples used for this paper were synthetized from a virtual model of a propeller aircraft. These samples were synthetized for both *synchronous* and *asynchronous* flying conditions, in 85 positions for each case, hence a total of 170 samples (2). For each sound sample, five psychoacoustic metrics were computed using Simcenter Testlab: Loudness, Sharpness, Tonality, Fluctuation Strength and Roughness. A spatial mapping of each one of these features can be seen in Figure 1. Also, these metrics were the subject of a *k-means* clustering analysis that allowed to, in a high dimensional group of data, find groups (i.e. clusters) with similar features (2). Four clusters of seats were obtained for each flying case (Figure 2).
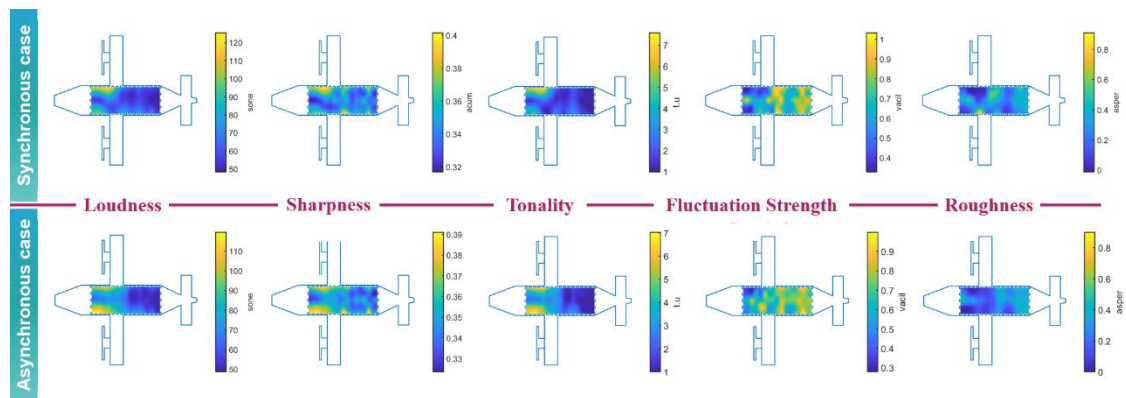

Figure 1: Spatial Distribution of the psychoacoustic metrics in the cabin of the aircraft.
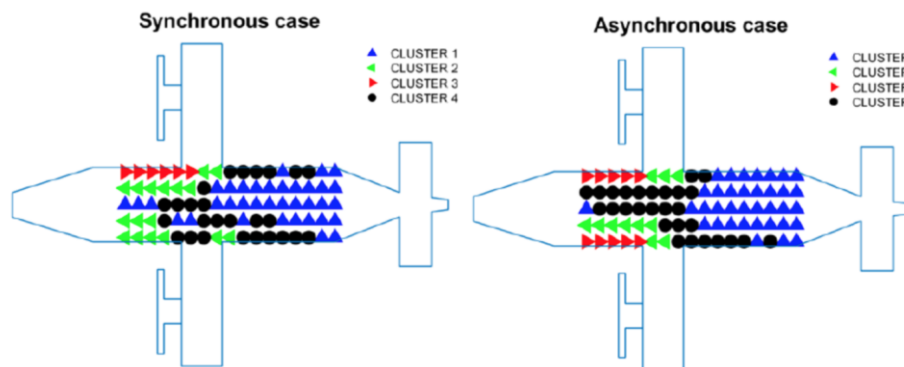

Figure 2: Distribution of the clusters in the cabin of the propeller aircraft (2).

2.2 A Data-driven Modular Approach for Sound Quality Prediction

In this paper, we developed a *Virtual Passenger* model with a modular approach for predicting Sound Quality in a propeller aircraft, where, in the first module, using *Convolutional Neural Networks* (CNNs), a prediction model was built by training the CNNs for predicting psychoacoustic metrics, directly from pressure signals. On the second module, the goal is to predict the passenger's acoustic discomfort, i.e., annoyance, in all the positions of the cabin in a propeller aircraft, where the input will be the psychoacoustic metrics predicted by the first module (Figure 3). For training this second module of the model, it was necessary to conduct a jury test campaign for collecting annoyance assessments from a pool of jurors. Pairing these evaluations of each sound sample with their corresponding psychoacoustic metrics, we used and compared different regression-based Machine Learning techniques for building the second block of the model.
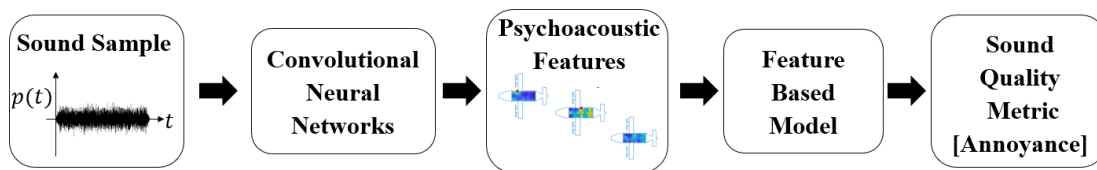
Figure 3: Block diagram describing the Virtual Passenger model.

2.3 Estimation of Psychoacoustic Features in Time Domain

The first module of the *Virtual Passenger* uses CNNs, which is a Machine Learning technique able to process data in its natural form. Therefore, the *manual* feature extraction step is bypassed, allowing to develop a *pipeline* where it is possible to directly predict acoustic discomfort from pressure signals. CNNs are able to recognize spatially or temporally invariant features from time-domain waveforms (3, 4, 5, 6). Also, the use of different psychoacoustic features guides the model in perceiving the different dimensions that will later be required for predicting Sound Quality. Such an end-to-end, data-driven, approach allows to overcome the need to compute semi-analytical algorithms for the extraction of the required psychoacoustic features (7, 8, 9, 10, 11). This simplifies the use of the predictive model in contexts such as multi-attribute optimization or control loops.

2.4 Bayesian Optimization

Adjusting the different available hyper-parameters represents one of the biggest challenges of using Machine Learning techniques, being this an often manual task that can require rules of thumb or even to conduct grid-searches (12). As shown in (13), Bayesian Optimization, by constructing a probabilistic model based on the Bayes theorem, outperforms other state of the art global optimization algorithms (14) because it is able to use the information available from previous evaluations of the objective function being used, and not simply rely on local gradient and Hessian approximations. Even though the evaluations of the objective function are expensive to perform, the ability to make better decisions justifies the extra computation cost (12).

2.5 Subjective Assessments of Sound Quality in a Propeller Aircraft

Acoustic discomfort, measured through annoyance, can be assessed by means of questionnaire items by appropriate response scales. We assessed the subjective perception of annoyance of the passengers in a propeller aircraft by asking a pool of jurors to perform a subjective evaluation test on a sample of the synthetized sound samples. After a one-week campaign, a total of 40 jurors evaluated the sample of sounds, in sessions of no more than 6 jurors at a time. For this campaign, the guidelines from (15) were followed.

For avoiding juror fatigue, a sample of 30 sound samples, with lengths of 6.5 seconds, were selected from the total of 170 pressure signals available. Considering the features of the sounds samples were clustered (as shown in 2.1), a sound sample was selected from each of the clusters. Also, the extreme cases for each psychoacoustic metric were selected, for ensuring the sample is representative. Finally, as suggested by Otto et al., a screening of the sound samples was done. The consistency of juror evaluations was checked by including 5 repeated sound samples. It should also be noted that all the sound samples were equalized to improve the sound quality during the test and fade in and fade out effects were included, for softening the start of the sound samples (15). The *Semantic Differential* method with a seven-point scale was used, where sound samples are presented one by one to the jurors, who has to rate them based on a pair of two opposing adjectives or expressions. However, it should be noted an *anchor sound* was used, i.e., before the juror listens to the sound sample he is assessing, he will first hear an *anchor sound*, which will be always the same sound. The *anchor sound* was selected by choosing a seat with average feature, thus seat number 38 (*synchronous*), located in the middle of the fuselage was the one chosen. This ensures the juror always has the same reference sound, increasing consistency.

2.6 Feature-based prediction with Machine Learning

Even though there is a high correlation between some psychoacoustic metrics and annoyance, the use of one single metric does not allow to model acoustic discomfort. A combination of features is necessary for building a model able to correlate psychoacoustic features with results from jury studies (16).

A comparative study between 4 different prediction techniques. A Multiple Linear Regression (MLR) allows to explain phenomena trough the combination of different variables (17), although being a linear technique which is a shortcoming when addressing strongly nonlinear regression (16). A more fitting approach is the use of a multilayer perceptron-based model, designated by Artificial Neural network is a common technique for non-linear modelling, where, the network training corresponds to the iterative process of computing the network parameters which minimize its error (4). Numerous training techniques can be used. The Levenberg-Marquardt (LM) algorithm (18) and the Bayesian Regularization (19) are two common ones and will be explored in this paper. Support Vector Machines (SVMs), a kernel-based method, corresponds to a statistical learning approach based on risk minimization principle and are also able to effectively model these phenomena (20, 21). Finally, the combination of several decision trees as a Random Forest (RF) will also be used for this type of nonlinear modelling (4). Similarly to the CNNs, the Machine Learning techniques above mentioned, have adjustable hyper-parameters. Bayesian optimization can also be a key tool when for having an automated way to perform this adjustment, enabling extensive robust performance assessment studies. However, ANNs parameters are a frequent target of manual hyper-parameter adjustments, where the number of hidden neurons or training algorithms can be carefully chosen and impact prediction performance (22).

3. Assessment of Prediction Performance

A systematic approach to assess prediction performance was used in this paper. Each time a prediction model is developed with p samples, an initial random division of the data into m training samples and n testing samples is done (being $p=m+n$). The model is trained only with the m training samples, being then the n testing independent variables inputted into the training model and the n response predictions compared with the n original responses from the training data (23). This comparison is done by computing

performance metrics. The Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and the Coefficient of Determination, denoted by R^2, where used, as defined in (17). Regarding this last coefficient, it ranges from 0 to 1 and the closer the value is to one, the better the fit, or relationship, between observed and predicted values (17). Also, the Pearson correlation coefficient was used in analyzing correlation between features and response (24). A consequence of randomly splitting a data set into training and testing samples, is that these samples may not be representative. For example, if the randomly selected *training data* happens to contain mostly samples with low annoyance values, then it would underperform when predicting on sounds of high annoyance. For ruling out these effects, the Monte Carlo simulation method was used, where this data division is randomly done many times (and for each time the model is trained and assessed). Afterwards, the average value of each performance metric can be computed and also its corresponding standard deviation during the simulation (25).

3.1 Subjective Assessments of Sound Quality: Results

As shown in section 2.5, each juror classified a sound stimulus with respect to the anchor sound by choosing one of the adjectives presented in the software interface. Going from a discrete to a continuous annoyance scale, let 100 correspond to the extreme for maximum annoyance (Much More Annoying) and 0 be the extreme for minimum annoyance (Much Less Annoying). A total of 1200 subjective evaluations were collected, from 40 jurors, each one evaluating 30 different sound samples. The results were standardized according to the standard deviation of each juror and then, for each stimulus the average evaluation from all jurors was computed, thus, a vector of 30 annoyance evaluations being obtained. Finally, each stimulus evaluation, was re-scaled in a way that 100 corresponds to maximum annoyance value and 0 to its minimum value. This procedure was also used for the psychoacoustic metrics, being that the extreme values that correspond to each one of the features can be found in Table 1. Table 2 contains the rescaled values of the psychoacoustic metrics and corresponding annoyance, allowing to relate each stimulus with the corresponding seat on the aircraft.

Table 1: Maximum and minimum values for each psychoacoustic metric.

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

Table 2: Subjective evaluations of sound samples obtained through jury testing and their respective psychoacoustic metrics (s indicates synchronous)

Seat Nr.	Stimuli	Loudness	Fluc. Strength	Tonality	Sharpness	Roughness	Annoyance
1	1	97.02	34.69	77.49	93.09	1.72	97.96
5s	2	90.97	0.00	88.96	87.16	7.96	95.22
5	3	75.20	18.78	77.29	70.87	8.19	91.21
15s	4	92.04	0.06	100.00	90.15	1.93	94.99
20	5	87.01	23.96	91.97	85.86	1.93	94.03
20s	6	100.00	1.98	99.17	100.00	0.79	97.84
25s	7	99.96	6.50	99.92	99.88	0.00	97.46
30	8	64.93	24.13	63.29	63.46	2.94	81.79
1s	9	63.51	6.64	48.95	57.57	6.96	87.96

6	10	98.36	40.03	86.64	94.85	1.93	100.00
35s	11	52.85	39.47	40.12	49.98	5.47	79.12
40s	12	38.03	55.39	28.64	37.64	19.28	68.12
26s	13	17.77	66.50	3.24	17.13	100.00	54.07
37	14	34.66	33.14	29.95	32.06	9.68	71.01
50	15	24.01	84.17	9.59	33.93	62.95	63.80
71	16	24.93	100.00	2.08	37.48	48.43	59.46
22	17	60.45	72.31	54.82	58.90	6.24	92.13
48	18	7.94	82.11	6.76	13.61	34.72	22.07
59s	19	12.74	50.95	16.40	10.33	13.71	33.61
59	20	6.65	54.95	7.75	5.06	23.79	24.79
64	21	5.08	59.13	5.22	4.38	25.30	8.88
64s	22	8.93	42.41	9.39	7.26	17.05	20.24
68s	23	7.37	77.50	1.12	13.04	37.15	25.55
72s	24	11.11	79.32	2.46	14.20	33.60	21.99
73	25	3.38	57.79	1.26	3.87	37.01	6.92
79s	26	0.18	53.10	0.00	0.88	32.81	0.00
81s	27	13.92	71.95	0.00	19.69	36.87	38.87
82	28	0.00	58.44	0.00	0.00	40.70	1.33
84s	29	2.23	41.29	1.50	0.56	25.95	5.67
85s	30	12.90	68.73	0.82	17.25	31.71	42.95

Before starting to develop the prediction models, it is important to study the correlation between the psychoacoustic metrics and the annoyance obtained for each sound sample, with jury testing. Thus, the Pearson correlation coefficient was computed for each psychoacoustic metric and annoyance and also between the different psychoacoustic metrics. These results are presented in Table 3.

Table 3: Correlation matrix for the analyzed objective and subjective attributes.

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

3.2 Predicting from Objective Metrics: Results

As described in section 2.6, four different types of feature-based prediction models were compared, where, for each one, a Monte Carlo simulation was done. The MLR has no relevant parameters to tune. Regarding the ANN, several combinations of its hyper-parameters are possible and play a role on its performance. Due to the small size of the data set and high correlation between predictor and response variables, only one hidden layer was used. The Monte Carlo simulation was performed, where, for two different training algorithms, the performance corresponding to different numbers of hidden neurons was computed. For each hidden neuron number, the data was divided 100 times and for, each one of this divisions, the model was trained and assessed, being computed the average RMSE and its standard deviation. These results are represented in Figure 4. Analyzing this figure, it is clear the best performance occurs for 2 hidden neurons with the LM training algorithm. Therefore, from hence on the ANN is used with 2 hidden neurons on the first hidden layer and trained with the LM training algorithm.
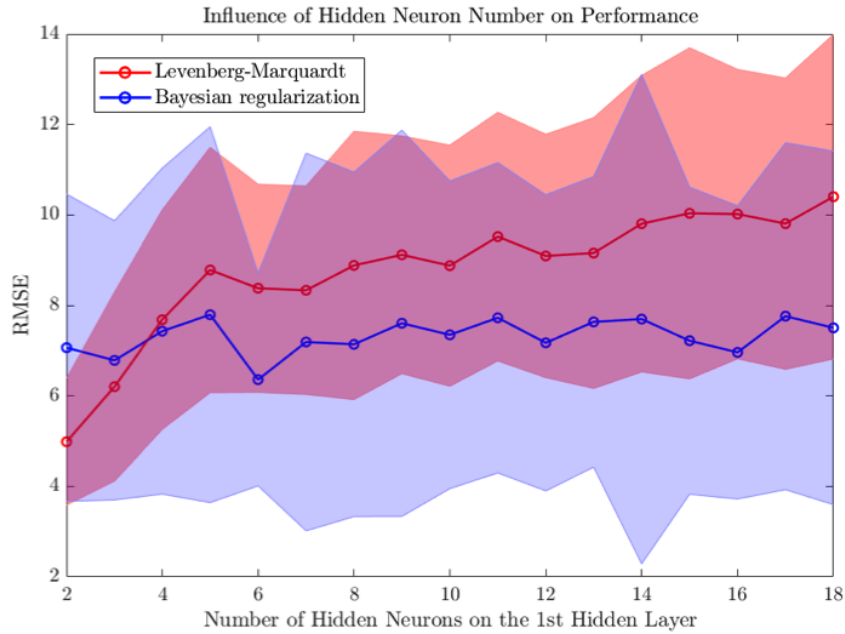
Figure 4: Study on the number of hidden neurons influence on performance, for two training types. Each point represents the mean RMSE of 100 random data divisions and the standard deviation of the RMSE in these 100 divisions. The shadowed colouring of each curve represents the standard deviation.
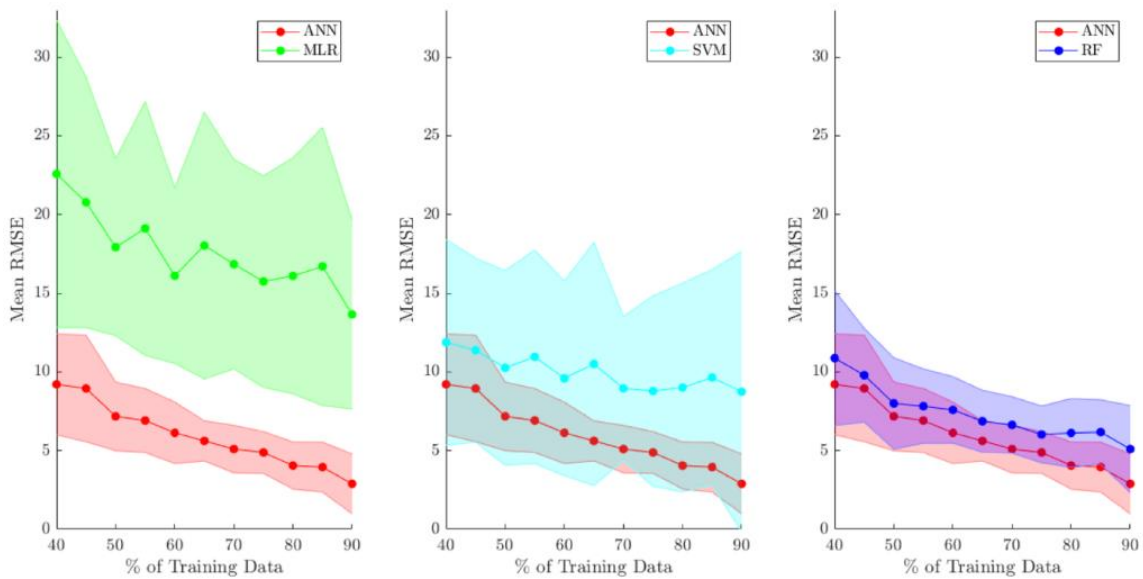

Figure 5: Comparison of the average RMSE and standard deviation of the RMSE (shaded colours) in 100 random divisions, for different percentages of training data.

A study was also done on the influence of percentage of *training data* used on the performance of the feature-based predictions models. For the SVMs and RFs, a Bayesian Optimization of their hyper-parameters was conducted for each time the models were trained. From early on it was notorious the ANN outperformed the other techniques. Therefore, in Figure 5, the performance of each one of the techniques is compared with the ANN, for different percentages of training data, where each point corresponds to the average RMSE (and corresponding standard deviation) of the performance of models with 100 different randomly selected samples of training and test data. The ANN

consistently has the smaller RMSE and also an inferior variability. Considering these results, an amount of 70% training data was for using in the further studies conducted, due to allowing to obtain a reasonable performance and ensuring an appropriate number of stimuli for testing (9). Also, it is notable the shortcomings of using MLR, being that it has a notably worse performance than the other methods, thus allowing to observe nonlinear modelling techniques are necessary for this phenomenon.

In Table 4, it is possible to find the performance of each technique, for 70% of training data, averaged over 100 random data divisions.

Table 4:Feature-based models averaged performance over 100 random data divisions (70% of data for training)

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

From the obtained results, the ANN was selected as the technique to use in complete prediction model. Selecting the best performing trained ANN from the ones previously trained, it is possible to input it with the data remaining from the 150 sound samples (i.e. their psychoacoustic metrics) and obtain a spatial mapping of the predicted annoyance values in all the seats in the aircraft, as shown in Figure 6. The obtained results allow to verify that annoyance has a great degree of variation in the cabin space and these variations occur in specific zones, being that the higher acoustic discomfort values are located in the seats near the propellers.
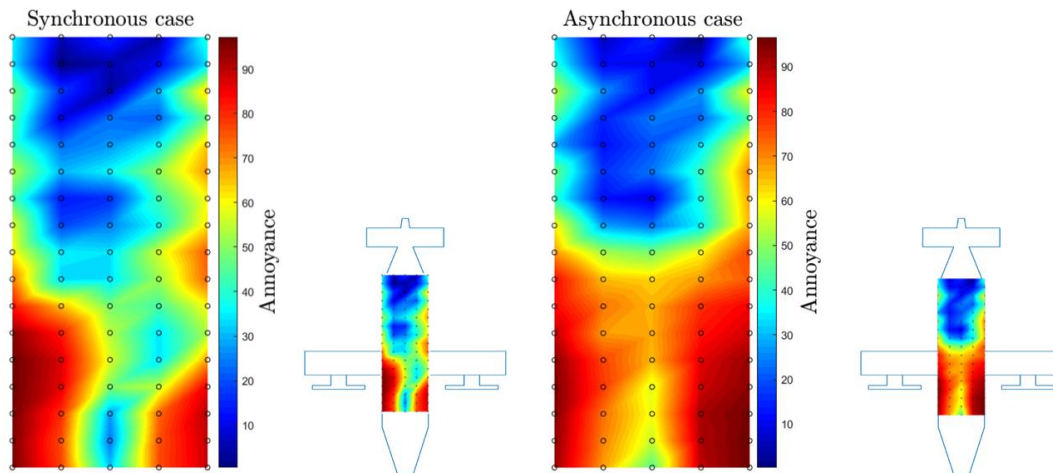
Figure 6: Annoyance prediction in the aircraft cabin using the trained ANN.

3.2 Predicting from Subjective Metrics: Results

As detailed in 2.5, CNNs are used for predicting features from time signals in the first block of the VP model. So, in this section, 5 CNNs are trained for predicting, from raw time signals the 5 psychoacoustic metrics previously used as inputs. The sounds samples used in this section were the ones that were not used for the jury testing. From the global set of 170 sound samples, 30 were used for jury testing, hence having 140 stimuli for training this 5 CNNs. These remaining 30 stimuli were used for assessing

prediction performance. Recalling from section 2.1 that their selection was done based on a cluster analysis, it is possible to consider these as a representative sample, thus being the Monte Carlo method not necessary. Due to the fact that five CNNs have to be trained, each one with a different data set, different architectures and hyper-parameters have to be selected in order to have a good performance. This was done using Bayesian Optimization, iterating between different layers until finding the stack that provides the best performance on the jury testing stimuli. In order to decrease the computing time necessary for training the models, the time signals were down-sampled from a sampling frequency of 44100 Hz to 8820 Hz. The selected architecture for each feature can be found on Table 5.

Table 5: Architectures used for predicting psychoacoustic metrics from time signals.

(a) Fluctuation Strength and Sharpness	(b) Loudness, Tonality and Roughness
Image Input Layer Convolutional Layer Batch Normalization Layer ReLU Layer Average Pooling Layer Convolutional Layer Batch Normalization Layer ReLU Layer Convolutional Layer Batch Normalization Layer ReLU Layer Dropout Layer Fully Connected Layer Regression Layer	Image Input Layer Convolutional Layer Batch Normalization Layer ReLU Layer Average Pooling Layer Convolutional Layer Batch Normalization Layer ReLU Layer Dropout Layer Fully Connected Layer Regression Layer

On Table 6 the obtained performance for each CNN based model is presented, scaled from 0 to 100 and also reconverted in each psychoacoustic metric original scale. Observing the analyzed results it is possible to conclude that for loudness, sharpness and tonality the performance is evidently superior than for fluctuation strength and roughness.

Table 6:Performance for the estimation of psychoacoustic metrics.

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

3.4 Feature Selection Study

Having now developed both blocks of the complete prediction model, it is possible to combine them having ready the Virtual Passenger model, where a sound sample can be inputted as a time signal, being the output the subjective sound evaluation (annoyance). However, as it was seen in the previous section, the 5 trained CNN that estimate features from the time signals have different performances. Therefore, feature selection allows to obtain a better overall performance of the Virtual Passenger model. Designating Loudness by *L*, Fluctuation Strength by *F*, Tonality by *T*, Sharpness by *S* and

Roughness by *R*, the effect of using different combinations of this features on the Virtual Passenger model performance was studied. Considering the results in Table 3, going from features with higher correlation with response to lower correlation ones, starting with all the features (*LFTSR*), these are sequentially removed one-by-one, until obtaining the combination *LS*. Also, *LT* was considered due to the high prediction performance obtained for tonality and because it represents an important psychoacoustic dimension. The 30 samples from the jury study were used for performance assessment and a Monte Carlo simulation was done. The 30 pressure signals are inputted into the CNNs and 30 feature estimations are obtained for the 30 samples. Then, for each feature combination, the predicted features are introduced into 100 differently trained ANNs. Finally, the performance in predicting annoyance is computed (comparing with the original juror response) and averaged. The results are shown on Table 7. It is possible to observe that *LS* (Loudness and Sharpness) is the best performing feature combination.

Table 7: Model average performances over 100 random data divisions for the second block, for each feature combination, subjectively assessed with the 30 sound samples.

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

When analyzing the developed model, it is necessary to recall that originally a juror would assess the sound sample with discrete classes. These classes would have a range of 17/100 in the continuous annoyance scale that was used. Therefore, firstly, accuracy can be defined as the number of times the model correctly predicts with an error inferior to the *width* of one of the original classes over the total number of predictions. Consequently, the developed predictive model has an accuracy of 80%, being that only 6 of the 30 stimuli have a prediction error superior to 17/100 (Figure 7).
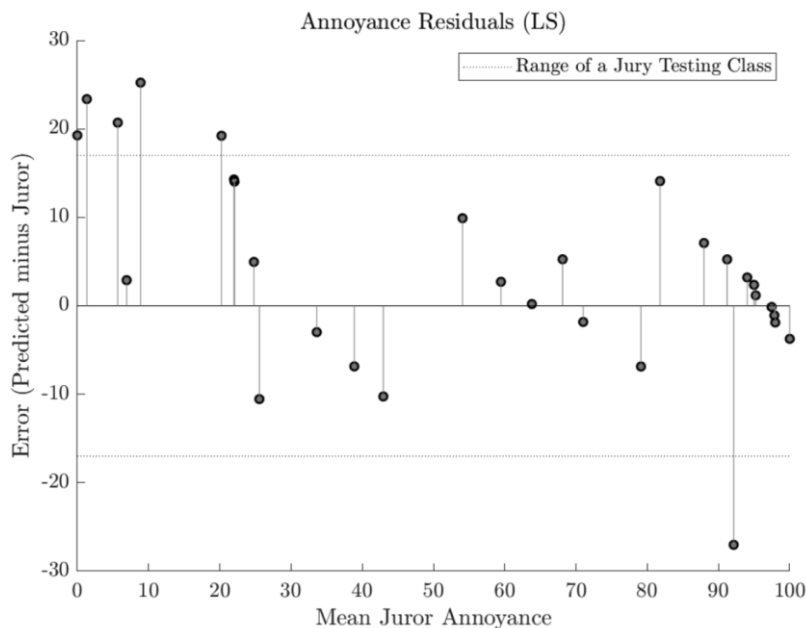

Figure 7: model predictions compared with the original mean juror annoyance.

4. CONCLUSIONS

This paper reports on the performance of a data-driven approach for the prediction of the acoustic discomfort of a passenger during a propeller aircraft flight. The proposed predictive model consists of two modules. The first one estimates, in time domain, psychoacoustic metrics using Convolutional Neural Networks tuned using Bayesian Optimization and the second one, based on jury testing data, is able to predict the passenger's acoustic discomfort from Sound Quality metrics. On this last block, four techniques were compared through Monte Carlo Simulation. It was concluded that Artificial Neural Networks are capable of outperforming Support Vector Machines, Random Forests and a Multiple Linear Regression with inferior average prediction error and variability. After having the whole model trained, a feature selection study allowed to conclude that the predictive model has maximum performance (overall accuracy: ~80%) when using as features Loudness and Sharpness.

In future steps, the developed predictive model will be integrated into a multi-attribute optimization workflow with the ambition of adopting the passenger annoyance as a parameter for the optimization of the design of the cabin of a virtual prototype of a regional propeller aircraft. In this perspective, ongoing studies are devoted to the assessment of the validity of the developed predictive model when applied to aircraft models of similar characteristics and in similar flight conditions.

6. REFERENCES

1. Janssens, K., Vecchio, A., and Van der Auweraer, H. "Synthesis and sound quality evaluation of exterior and interior aircraft noise." Aerospace Science and Technology 12.1 (2008): 114-124.

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

3. LeCun, Y., Bengio, Y., and Hinton, G. "Deep learning." nature 521.7553 (2015): 436.

4. Bishop, C. M. "Pattern recognition and machine learning." Springer, (2006).

5. Ba, J., Kiros, J. and Hinton, G.E. "Layer normalization." arXiv:1607.06450 (2016).

6. Srivastava, N., et al. "Dropout: a simple way to prevent neural networks from overfitting." The Journal of Machine Learning Research 15.1 (2014): 1929-1958.

7. Aytar, Y., Vondrick, C. and Torralba, A. "Soundnet: Learning sound representations from unlabeled video." Advances in neural information processing systems. (2016).

8. Dai, W. et al. "Very deep convolutional neural networks for raw waveforms." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (2017).

9. Sainath, T. N. et al. "Learning the speech front-end with raw waveform CLDNNs." 16th Annual Conference of the International Speech Communication Association. (2015).

10. Thickstun, J., Harchaoui, Z. and Kakade, S. "Learning features of music from scratch." arXiv:1611.09827 (2016).

11. Trigeorgis, G. et al. "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network." IEEE international conference on acoustics, speech and signal processing (ICASSP). (2016).

12. Snoek, J., Larochelle, H. and Adams, R. P. "Practical bayesian optimization of machine learning algorithms." Advances in neural information processing systems. (2012).

13. Jones, D. R. "A taxonomy of global optimization methods based on response surfaces." Journal of global optimization 21.4 (2001): 345-383.

14. Brochu, E., Cora, V. M. and De Freitas, N. "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning." arXiv:1012.2599 (2010).

15. Otto, N. et al. "Guidelines for jury evaluations of automotive sounds." SAE transactions (1999): 3015-3034.

16. Pietila, G., and Lim, T. C. "Intelligent systems approaches to product sound quality evaluations–A review." Applied Acoustics 73.10 (2012): 987-1002.

17. Rawlings, J. O., Pantula, S., and Dickey, D. A. "Applied regression analysis: a research tool." Springer Science & Business Media, (2001).

18. Zhang, H. F. et al. "Performance of the Levenberg–Marquardt neural network approach in nuclear mass prediction." Journal of Physics G: Nuclear and Particle Physics 44.4 (2017): 045110.

19. MacKay, D. J. C. "Bayesian interpolation." Neural computation 4.3 (1992): 415-447.

20. Liu, H. et al. "Sound quality prediction for engine-radiated noise." Mechanical Systems and Signal Processing 56 (2015): 277-287.

21. Vapnik, V. "The nature of statistical learning theory." Springer, (2013).

22. Kahn, M. "Sound quality evaluation of heavy-duty engines in free field conditions." Diss. Luleå tekniska universitet, (1998).

23. Russell, S., Norvig, P. and Davis, E. "Artificial Intelligence: A Modern Approach." Prentice Hall, (2010).

24. Fisher, R. "Statistical methods for research workers." Edinburgh Oliver & Boyd, (1925).

25. Kwak, Y. H. and Ingall, L. "Exploring Monte Carlo simulation applications for project management." Risk Management 9.1 (2007): 44-57.