ENGINE SOUND COMFORTABILITY: RELEVANT SOUND QUALITY PARAMETERS AND CLASSIFICATION

T. Coen [*,1] **N. Jans** [*] **P. Van de Ponseele** [**]
I. Goethals [*] **J. De Baerdemaeker** [***] **B. De Moor** [*]

[*] *Department of Electrotechnical Engineering, K.U.Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium*
[**] *LMS International, Interleuvenlaan 68, B-3001 Leuven, Belgium*
[***] *Department of Agro-Engineering and -Economics, K.U.Leuven, Kasteelpark Arenberg 30, B-3001 Leuven, Belgium*

Abstract: In order to be able to shorten the design cycle, automobile manufacturers are interested in modelling the human perception of engine sounds.
In the first part of the paper the relevant Sound Quality parameters for the prediction of engine sound comfortability are determined. The inputs are ordered with Automatic Relevance Determination and the obtained ranking is verified on the data. In the second part, models are presented to classify and compare cars on comfortability. Least Squares Support Vector Machines (LS-SVMs) is used for the classification. The influence of selecting the relevant inputs on the model performance is illustrated.

Keywords: Neural networks, Models, Automobile Industry, Classification, Inputs

1. INTRODUCTION

The relationship between automobile manufacturers and consumers has changed tremendously over the passed years. The design of a car has become more and more based upon the desires of the consumer. Since consumer desires are subject to change over time, the design specifications of a car change as well. This necessitates shorter design cycles in order to keep up with customer desires (Schöggl, 1998) (Keiper, 1997).

In the paper the focus lies upon determining which Sound Quality (SQ) parameters are the most relevant for modelling the comfortability (as perceived by a consumer) of engine sounds. In order to obtain the opinion of the consumer, jury tests have to be organized. In such a test, a person is asked to score each sound on a characteristic, for example comfortability.

The classic jury test practice is however incompatible with the current evolution of the automobile industry. Some of the drawbacks are:

Disturbances: A judge-specific bias to the scores is introduced due to variation of equipment, different interpretation of the questions, noise,...

Composition of the jury: A large and balanced (different background, age, ...) population is necessary for a jury test to be significant.

The above mentioned problems result in a considerable time span (about a month) that is needed to organize and process a jury test (Fastle, 1997)

[1] corresponding author: tom.coen@agr.kuleuven.ac.be

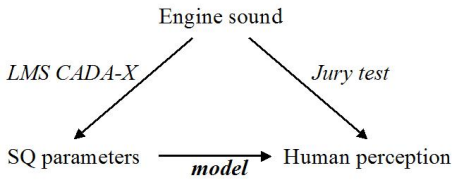

Fig. 1. Schematic overview

(Genuit, 1997). This is no longer deemed acceptable.

Objective The most relevant SQ parameters to predict comfortability of engine sounds will be selected from a group of 9 parameters. These 9 SQ parameters are recommended by experts from the automobile industry.

The performance of the model with all 9 SQ parameters is compared to the performance of models with reduced input dimension. An optimal input configuration is determined. This is illustrated with models to classify and compare cars on comfortability.

LS-SVM (Suykens *et al.*, 2002) is chosen as modelling technique. The most relevant inputs are determined with Automatic Relevance Determination (ARD) for LS-SVMs.

2. GLOBAL SET-UP

2.1 SQ parameters

The 9 SQ parameters suggested by the experts can be divided in 3 groups.

A first group of parameters, namely A-weighted Sound Pressure Level (SPLA), B-weighted Sound Pressure Level (SPLB) and Zwicker Loudness, is correlated with the Sound Pressure Level of the sound. SPLA and SPLB are Sound Pressure Levels with respective weighting functions A and B (Van der Auweraer and Wykaert, 1998). Zwicker Loudness is the human perception of sound, and is calculated from SPL levels by using a conversion Table (Zwicker, 1977).

The second group, namely Articulation Index (AI), Modified or Open Articulation Index (AIM), ANSI Speech Interference Level (ASIL) and Preferred Speech Interference Level (PSIL) (Van der Auweraer and Van de Ponseele, 1998), describes how comprehensible a conversation would be with the sound as background. AI and AIM are based on a special weighting of the SPL levels. Frequencies that are more important for the understanding of speech receive a higher weighting factor. The results are normalized. 100% means that a conversation is perfectly comprehensible. 70% or less means that conversation becomes difficult. ASIL and PSIL are the average of the sound pressure levels over the frequency bands that are important for speech. Thus, the lower the value of this parameter, the more comprehensible a conversation is.

It is clear from the definition that there is a negative correlation between AI, AIM and ASIL, PSIL.

A third group of parameters consists of Sharpness and Roughness (Van der Auweraer and Van de Ponseele, 1998). Sharpness is based on the Loudness algorithm with higher weighting factors for the higher frequencies. Roughness is a measure of the degree of modulation weighted per third octave of the sound.

It is clear that not all these parameters are independent. Within the first and second group there is a strong correlation between the defined parameters. In a later stage of the modelling the most appropriate parameters will be selected.

2.2 Data acquisition

Run-ups of 30 significantly different cars were recorded during road tests (Coen *et al.*, 2004). A microphone was placed on the left and on the right of the head support of the driver. In this way the recorded sound is the actual sound heard in the car by the driver. This set-up implies that the engine sound as well as the effect of the isolation of the interior of the car is taken into account. After all the opinion of the driver is what is important for an automobile manufacturer.

The recorded sounds are then used in a jury test. The participants first fill out a form with some background information (age, driving habits,'car perception',...). The sound is played and the participants give a grade between 0 (not comfortable at all) and 10 (very comfortable). A judge grades each sound twice to check the consistency of the judge. If those 2 grades are too far apart on too many cars, the scores of this judge are removed from the dataset.

The jury test consists of 104 judges. The dataset used here is based on the average score given by the 79 judges that are consistent.

2.3 Scores processing

For the training of the model, one score for each car is needed. Simply averaging the scores over all judges is not a good idea. Each judge has a different mean and variation (over all cars) which is presumed not to be significant since the judges are no experts.

Therefore the scores are normalized to zero mean and unit standard deviation and subsequently averaged over all the judges. In this way a judge with a high variation no longer has a larger impact on the final score of a car.

3. SELECTING RELEVANT INPUTS

3.1 Automatic Relevance Determination

Automatic Relevance Determination (ARD) is in fact a special case of a classic LS-SVM with an RBF kernel (Suykens *et al.*, 2002). With LS-SVMs all inputs are accorded the same weight. If however not all inputs are equally relevant to the model, the model can be improved by removing the unimportant inputs.

An RBF kernel has the following form:

$$e^{\frac{-\|x-x_k\|_2^2}{\sigma^2}}. \tag{1}$$

If instead of the standard 2-norm a norm with a diagonal weighting matrix Σ^{-2} is used, an RBF kernel can be written as:

$$e^{-(x-x_k)^T\Sigma^{-2}(x-x_k)}. \tag{2}$$

ARD determines the elements on the diagonal of Σ. In this way each input i is accorded an own weighting factor σ_i.

Assuming the inputs are normalized (which is the case), a small σ_i indicates an important input. Even a small difference in this input dimension will have a large impact on the kernel function because of the small σ_i.

The performance of ARD will be illustrated by presenting the function estimated between the SQ parameters and perception of comfortability for input dimensions ranging from 1 to 9.

3.2 Relevant inputs for comfortability

In (Bisping, 1997) it is stated that not all SQ parameters are equally relevant to the perception of comfortability. The dependence between the proposed SQ parameters makes this an interesting test case for ARD.

ARD is now used to rank the available SQ parameters according to their importance to the prediction of the comfortability scores. The SQ parameters that don't provide a significant contribution to the prediction of the scores are marked in *italic*. The order given by ARD is:

(1) Zwicker Loudness
(2) ASIL
(3) AIM
(4) SPLB
(5) SPLA
(6) PSIL
(7) Sharpness
(8) *AI*
(9) *Roughness*

If certain SQ parameters are equivalent, they would be equally relevant for ARD. Since for example Zwicker Loudness, SPLA and SPLB are not all at the same position in the ranking, there is a clear difference in relevance between these parameters. Remarkable is also the difference in relevance between AI and AIM.

The most relevant parameter is Zwicker Loudness. This can be explained intuitively. Most people define comfortability as a lack of sound. Since Zwicker Loudness is an indication of the sound pressure level experienced by humans, this stands out as the ideal measure for comfortability. The 3 most important inputs are completely independent. The 4^{th} most important input however, SPLB, is linked to Zwicker Loudness.

The results obtained by ARD are now verified by training LS-SVMs with different input dimensions and an RBF kernel. The input dimension is reduced sequentially from 9 to 1, retaining only the most relevant inputs (according to ARD). For each input dimension 10 independent models are trained with each time an other random partition of the available data in training- and testset. The performance over 10 runs for the different input dimensions is illustrated in Figure 2. The evolution of median, mean, minimum, maximum, first quartile and third quartile over the different input dimensions is shown.

Especially the median and the first and third quartile are interesting measures to evaluate the performance of a model, since they are not sensitive to outliers. The maximum and minimum percentage give an indication of the worst case performance.

For input dimension 4 the median over 10 runs is 100%, and mean and interquartile range reach a maximum respectively minimum. The inputs of this model are Zwicker Loudness, ASIL, AIM and SPLB. Two of these inputs are correlated, namely Zwicker Loudness and SPLB. In Figure 2 the model performance clearly improves by adding SPLB. At first sight it might appear strange that adding an input (SPLB) that is strongly correlated to an existing input (Zwicker Loudness) can improve the overall model performance. However, there can be given several reasons for this:

- The new input contains some extra information compared to the original input.

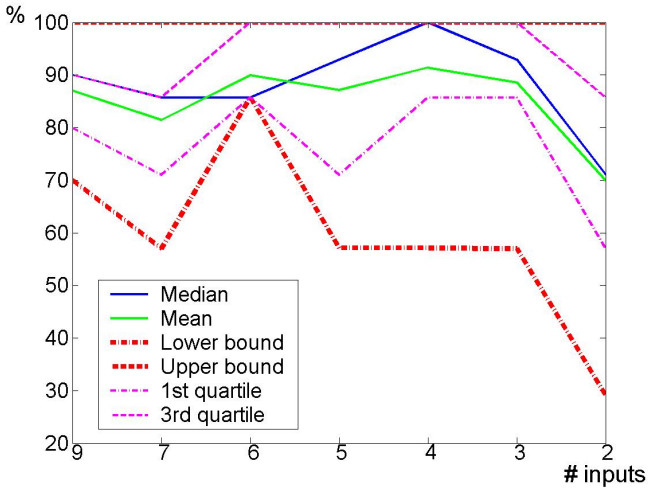

Fig. 2. Performance of models with different input dimensions. For each input dimension 10 runs are performed. The minimum and the first quartile give a lower bound on the model performance. The maximum and the third quartile give an upper bound. Median and mean illustrate the overall performance.
The model with input dimension 4 clearly obtains the best performance. The median, the first and third quartile and the mean reach a maximum for input dimension 4.

- Zwicker Loudness is more important than the other inputs of the original model. Adding a correlated input to the model adds extra weight to Zwicker Loudness.

If the second suggestion is correct, a model with ASIL, AIM and 2 times Zwicker Loudness as input should outperform the model with ASIL, AIM, Zwicker Loudness and SPLB. The results of both models are compared in Table 1.

Table 1. Model with RBF kernel and as input ASIL, AIM and 2 times Zwicker Loudness (A) versus model with an RBF kernel and as input ASIL, AIM, Zwicker Loudness and SPLB (B).

	Median	Mean	Std	Min.	Max.
A	100.0%	94.3%	7.4%	85.7%	100.0%
B	100.0%	91.4%	13.8%	57.1%	100.0%

It's clear that the model with 2 times Zwicker Loudness (A) performs better than the model with Zwicker Loudness and SPLB (B). The mean is higher and the standard deviation (Std) is smaller for model A.

Weighting inputs Increasing the importance of an input can be done by reducing the σ (as it is done in ARD), by rescaling an input after preprocessing or by applying an input a second time. These approaches are equivalent in LS-SVM with an RBF kernel. Rescaling an input or adjusting σ is of course the same.

As shown in equation (1), the norm of the difference between 2 datavectors is important for LS-SVM classifiers. Applying the same input twice, corresponds to rescaling the difference in that input with a factor $\sqrt{2}$. This effect is illustrated in Figure 3.

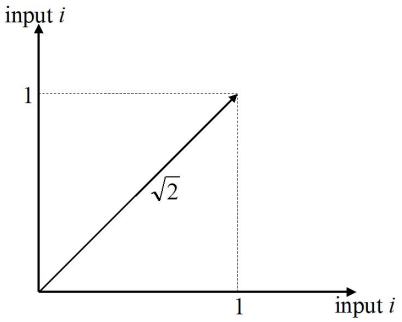

Fig. 3. The effect of applying an input twice

3.3 Correlation analysis

The results of ARD can be compared to the results of correlation analysis. For each SQ parameter the correlation is calculated as the cosine of the angle between the vector that contains the scores and the vector that contains the value of the SQ parameter for the different cars. This means the correlation is normalized between -1 and 1. The different SQ parameters, their correlation with the scores vector and their position in the ARD-ranking are shown in Table 2 for descending correlation with the comfortability scores.

The ranking obtained with ARD differs strongly from the one obtained with correlation analysis. There are several reasons for this:

Table 2. Relevance order by correlation analysis

SQ parameter	correlation	ARD ranking
SPLA	-0.95167	5
Zwicker Loudness	-0.94031	1
SPLB	-0.91952	4
PSIL	-0.81891	6
ASIL	-0.81080	2
AI	0.78781	8
AIM	0.78017	3
Roughness	-0.33459	9
Sharpness	-0.15915	7

- Correlation analysis is a linear technique. It cannot take non-linear relationships into account.
- Correlation analysis evaluates the relevance of each input separately. ARD tries to find the most relevant *group* of inputs.

4. MODELLING

In the paper several models are developed. Firstly a model that classifies cars on comfortability is discussed. Secondly a model to compare cars on comfortability is presented. For each of these models different input configurations are tested. In this way the results of ARD are evaluated. Because of the large variation of the scores over the different judges, predicting the scores is not useful.

All the models used here are classifiers. There are 2 possibilities to discretize the continuous scores:

- Divide the scores into classes, and encode the classes. Used codes are Minimum Output Code (MOC), One versus One (OO) encoding and One versus All (OA) encoding (Van Gestel *et al.*, 2002). Build a classifier (*classification*).
- Build a continuous prediction model (*function estimation*) (Suykens *et al.*, 2002), and then discretize into classes.

The best results for this dataset are obtained with function estimation (FE).

4.1 Qualitative judgement

4.1.1. Classification in 2 classes Two classes are defined, namely score smaller than 0 (class 1) and larger than 0 (class 2). The number of datapoints for each class is:

	class 1	class 2
Number	14	16

The medians for the performed experiments are shown in Table 3. The best results are obtained with an RBF kernel and FE.

Table 3. Median for different kernels and encodings, 2 classes comfortability

	MOC	MOC4	FE	FE4
Lin	71.4%	78.6%	85.7%	92.9%
RBF	85.7%	92.9%	100.0%	100.0%
Poly2	64.3%	78.6%	85.7%	92.9%
Poly3	57.1%	71.4%	71.4%	85.7%

Table 4. Median for different kernels and encodings, 3 classes comfortability

	MOC	MOC4	FE	FE4
Lin	42.9%	57.1%	71.4%	71.4%
RBF	42.9%	50.0%	71.4%	85.7%
Poly2	28.6%		42.9%	
Poly3	21.4%		57.1%	

Table 5. Median for different kernels and encodings, 4 classes comfortability

	MOC	OO	FE	FE4
Lin	35.7%	50.0%	78.6%	85.7%
RBF	28.6%	28.6%	85.7%	92.9%

Using the results obtained with ARD (see Section 3), this classifier can be derived based on the 4 most important inputs (which are Zwicker Loudness, ASIL, AIM and SPLB). This improves the model performance significantly. These results are also shown in Table 3 (*MOC4* and *FE4*).

4.1.2. Classification in 3 classes Three classes are defined as follows: scores clearly smaller than 0 (smaller than -0.25) (class 1), scores around 0 (between -0.25 and 0.25) (class 2) and scores clearly larger than 0 (larger than 0.25). The number of datapoints in each class is:

	class 1	class 2	class 3
Number	8	13	9

Experiments with different kernels and encodings are performed (see Table 4). Only FE with a linear kernel and an RBF kernel gives acceptable results. Reducing the input space to the 4 most relevant inputs (Zwicker Loudness, ASIL, AIM and SPLB) leads again to better results. The medians of these experiments are also shown in Table 4 (*MOC4* and *FE4*).

4.1.3. Classification in 4 classes Four classes are defined: scores smaller than 0.5 (class 1), scores between -0.5 and 0 (class 2), scores between 0 and 0.5 (class 3) and scores exceeding 0.5 (class 4) . The number of datavectors in each class is:

	class 1	class 2	class 3	class 4
Number	5	11	10	4

Experiments with a linear kernel and an RBF kernel are performed (see Table 5). FE clearly gives the best results. The performance improves by restricting the input space to the 4 most relevant inputs (*FE4*).

Table 6. Median for comparing cars on comfortability

	MOC	FE	ΔMOC	ΔFE
Lin	82.1%	82.1%	85.7%	89.3%
RBF	82.1%	67.9%	85.7%	89.3%

4.2 Comparing 2 cars

The SQ vectors of 2 cars are used as input for the model. The output is a relative judgement of the comfortability of both cars. This kind of model can be used to establish a ranking of cars, and to fit a new car into an existing ranking.

The dataset was divided into 2 groups for every run: a trainingset of 23 cars and a testset of 7 cars. Within each group every car is compared to all other cars in order to compile the actual dataset. The trainingset thus contains $\frac{(22+1)22}{2} = 253$ datavectors, the testset contains $\frac{(6+1)6}{2} = 21$ datavectors. Models are trained with 2 different input configurations: either with 2 SQ vectors as input (dimension 18) or with the difference of both SQ vectors as input (dimension 9) (indicated by Δ).

Experiments are performed with an RBF and a linear kernel, combined with MOC encoding and FE (followed by discretization) (see Table 6).

The results improve significantly by applying the difference of both SQ vectors to the input. There are 2 reasons for this:

- The dimension of the input space is halved from 18 to 9.
- The structure of the input is more obvious. In the case with 18 inputs the corresponding SQ parameters of both the cars have to be mapped to one an other. This adds complexity to the modelling task.

5. CONCLUSIONS

In the paper ARD is used to determine the relevant SQ parameters for the modelling of the human perception of engine sound. The most important SQ parameters are Zwicker Loudness, ASIL, AIM and SPLB. This is confirmed for classifying and comparing cars on comfortability. Most models show a significant performance improvement when reducing the input space to the 4 most relevant SQ parameters.

ACKNOWLEDGEMENTS

The research work was done while the senior author was at K.U.Leuven, Department of Electrotechnical Engineering (ESAT). N. Jans was a student at K.U.Leuven. P. Van de Ponseele is with LMS. I. Goethals is a doctoral researcher with the FWO. J. De Baerdemaeker and B. De Moor are full professors with the K.U.Leuven. The research was conducted in cooperation with LMS International (http://www.lmsintl.com/). Our research is supported by Research Council KUL: GOA-Mefisto 666, GOA AMBioRICS, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects, G.0240.99, G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04 , G.0499.04; research communities (ICCoS, ANMMM, MLDM); AWI: Bil. Int. Collaboration Hungary/ Poland; IWT: PhD Grants,GBOU (McKnow) Belgian Federal Science Policy Office: IUAP P5/22; PODO-II; EU: FP5-Quprodis; ERNSI; Eureka 2063-IMPACT; Eureka 2419-FliTE; Contract Research/agreements: ISMC/IPCOS, Data4s, TML, Elia, LMS, Mastercard

REFERENCES

Bisping, R. (1997). Car interior sound quality: Experimental analysis by synthesis. *Acustica* **83**(5), 813–818.

Coen, T., N. Jans, I. Goethals, P. Van de Ponseele, A. Vecchio and B. De Moor (2004). Modelleren van het verband tussen menselijke waarneming en sound quality parameters gebruik makende van ls-svms. K.U. Leuven, ESAT, Leuven, Belgium.

Fastle, H. (1997). The psychoacoustics of sound-quality evaluation. *Acustica* **83**, 754–764.

Genuit, R. (1997). Background and practical examples of sound design. *Acustica* **83**(5), 805–812.

Keiper, W. (1997). Sound quality evaluation in the product cycle. *Acustica* **83**(5), 784–788.

Van der Auweraer, H. and K. Wykaert (1998). Sound quality: Perception, analysis and engineering. LMS International Internal Report.

Van der Auweraer, H. and P. Van de Ponseele (1998). Sound quality evaluation and measurement criteria. LMS International Internal Report.

Van Gestel, T., J.A.K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor and J. Vandewalle (2002). Multiclass ls-svms: moderated outputs and coding-decoding schemes. *Neural Processing Letters* **15**(1), 45–58.

Schöggl, P. (1998). Confort acoustique et vibratoire automobile et ferroviaire. Société des Ingénieurs Automobile, Courbevoie, 3-4 December.

Suykens, J.A.K., T. Van Gestel, B. De Moor and J. Vandewalle (2002). *Least Squares Support Vector Machines*. World Scientific Pub. Co., Singapore.

Zwicker, E. (1977). Procedure for calculating loudness of temporally variable sounds. *Journal of the Acoustical Society of America* **62**(3), 675–682.